Technische Universität München
TUM School of Engineering and Design

Towards Detecting Global Urban Poverty from Space

Thomas Daniel Stark

Vollständiger Abdruck der von der TUM School of Engineering and Design der Technischen Universität München zur Erlangung eines

Doktors der Ingenieurwissenschaften (Dr. Ing.)

genehmigten Dissertation.

Vorsitz: Prof. Dr.-Ing. Muhammad Shahzad

Prüfende der Dissertation:

1. Prof. Dr.-Ing. habil. Xiaoxiang Zhu
2. Prof. Dr. habil. Hannes Taubenböck
3. Assoc. Prof. Dr. Monika Kuffer

Die Dissertation wurde am 06.11.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Engineering and Design am 16.04.2025 angenommen.

Abstract

Extreme urban poverty in the Global South remains a persistent challenge, particularly in slum areas. Despite international initiatives like the Sustainable Development Goals (SDGs), progress in reducing extreme urban poverty has been slow. Slums continue to face significant grievances such as inadequate housing, lack of basic services, and vulnerability to environmental and social risks. While efforts have been made to address these issues and meaningful change is occurring, the scale of the problem and the complexity of urban poverty mean that it is not yet fully resolved.

This dissertation aims to help close the data gap by detecting slums using globally available remote sensing data and modern deep learning techniques. By applying advanced machine learning models to satellite imagery, the goal is to map slums on a large scale, providing policymakers and urban planners with the data they need to address extreme urban poverty. This work aims improve the understanding of slum locations and characteristics across cities in the Global South, where data is often scarce or outdated.

The methodology involves comparing different types of remote sensing data using semantic segmentation to determine which is best suited for slum mapping. In the case of Mumbai, three types of data are evaluated: very high-resolution QuickBird imagery, high-resolution Sentinel-2 optical data, and high-resolution TerraSAR-X radar data. A Fully Convolutional Network (FCN) is employed to assess the performance of these datasets in detecting slums. The next phase of the study extends this approach to PlanetScope data, which offers global coverage. Ten cities across the Global South, each with morphologically different slum structures, are mapped using a custom-designed fully convolutional Xception network (XFCN).

Capturing the highly variable morphological structures of slums in different regions is a major challenge. To do justice to this variability, Monte Carlo dropout and a specially developed convolutional neural network (STnet) are employed to estimate and minimize uncertainties in the slum mapping process. These uncertainty estimates are crucial for accurately distinguishing between slum and non-slum areas, particularly in regions with mixed urban features. Additionally, techniques such as Test Time Dropout and Test Time Augmentations are introduced to further enhance the model's performance, enabling large-scale slum mapping across 55 cities in the Global South. Achieving accurate and reliable results in such diverse urban environments poses a significant challenge due to the unique characteristics of each city.

Zusammenfassung

Trotz globaler Initiativen wie den Sustainable Development Goals (SDG) bleibt extreme urbane Armut in vielen Städten des Globalen Südens ein drängendes Problem. In den dicht besiedelten Slums leiden Millionen von Menschen unter schlechten Lebensbedingungen, fehlender Infrastruktur und mangelnden Chancen. Fortschritte zur Bekämpfung dieser Missstände sind trotz internationaler Bemühungen oft begrenzt, da viele Regionen nur unzureichend erfasst und analysiert werden. Es fehlt an Daten, die es ermöglichen, gezielte Maßnahmen zur Verbesserung der Situation zu ergreifen.

Diese Arbeit zielt darauf ab, eine Lücke in der Datenerfassung zu schließen, indem global verfügbare Fernerkundungsdaten und moderne Künstliche Intelligenz (KI) Methoden eingesetzt werden, um Slums zu identifizieren und zu kartieren. Durch den Einsatz von hochaufgelösten Satellitendaten und tiefen neuronalen Netzwerken wird es möglich, urbane Armut in großem Maßstab zu analysieren und die unterschiedlichen morphologischen Slumstrukturen zu erkennen. Diese Methodik soll auf Städte des Globalen Südens angewendet werden.

Zunächst werden verschiedene Fernerkundungsdaten mittels semantischer Segmentierung gegenübergestellt. In Mumbai werden Satellitendaten von QuickBird, Sentinel-2 und TerraSAR-X verglichen, um zu bestimmen, welche Daten am besten zur Slum-Detektion geeignet sind. Hierfür wird ein Fully Convolutional Network (FCN) verwendet. Im nächsten Schritt wird getestet, ob sich auch in anderen Regionen des Globalen Südens Slums mit hochaufgelösten PlanetScope-Daten kartieren lassen. In zehn Städten mit morphologisch unterschiedlichen Slums kommt ein eigens entwickeltes Fully Convolutional Xception Network (XFCN) zum Einsatz.

Die Erfassung der stark variierenden morphologischen Strukturen von Slums in unterschiedlichen Regionen stellt eine große Herausforderung dar. Um dieser Variabilität gerecht zu werden, werden in der nächsten Phase Monte-Carlo-Dropout und ein speziell entwickeltes Convolutional Neural Network (STnet) eingesetzt. Diese Methoden helfen, Unsicherheiten im Prozess der Slum-Kartierung zu erkennen und zu reduzieren. Das ist besonders wichtig, um genau zwischen Slum- und Nicht-Slumgebieten zu unterscheiden, vor allem in städtischen Regionen mit gemischten Strukturen. Zusätzlich werden Verfahren wie Test Time Dropout und Test Time Augmentations angewendet, um die Genauigkeit des Modells weiter zu steigern und eine großflächige Kartierung von Slums in 55 Städten des Globalen Südens zu ermöglichen. Die Herausforderung besteht darin, in so unterschiedlichen städtischen Umgebungen zuverlässige Ergebnisse zu erzielen, da jede Stadt ihre eigenen Besonderheiten aufweist.

Acknowledgements

Who knew that six years could feel like both a lifetime and a heartbeat? This dissertation journey has been an extraordinary adventure filled with thrilling highs and also challenging times. Navigating the uncharted waters of a global pandemic, adapting to working from home, and finally reuniting with coworkers have all shaped this work in unforgettable ways. Raising a family alongside pursuing this research has added profound meaning to this chapter of my life, making it a journey I will cherish forever.

This work would not have been possible without the kind support of many people in my life. First and foremost, my deepest gratitude goes to my supervisor, Prof. Xiaoxiang Zhu, for providing exceptional supervision, unwavering motivation, and fostering a great work environment. Equally, I extend heartfelt thanks to Prof. Hannes Taubenböck for his encouraging and consistently positive outlook, and for generously sharing his expertise. Their open-door approach, readiness whenever needed, comprehensive knowledge, and broad perspectives enabled the ambitious, meaningful, and engaging conceptual ideas behind this thesis.

I would also like to express my sincere gratitude to Prof. Monika Kuffer. Her unparalleled work on detecting urban poverty using remote sensing has been a constant inspiration for many aspects of this dissertation. Meeting her at conferences and exchanging ideas have always been enriching experiences that greatly influenced my work.

Equally, my sincere gratitude goes to my mentor, Dr. Michael Wurm, for his unparalleled daily supervision and his personable, knowledgeable, and passionate mentorship style. From day one, his liberal and flexible open-door policy made him consistently available, and I feel truly fortunate—I couldn't have asked for stronger support. He has enormously helped me in critical thinking, problem-solving, experiment design, paper writing, presentations, conflict resolution, establishing collaborations, mentoring students, and managing time and energy. More than just a mentor, he is a friend. With mutual trust, we've always had open, efficient, and effective communication. He knows exactly how to give me the praise I need and understands the struggles I haven't voiced. I am deeply thankful.

My heartfelt thanks also go to my wonderful colleagues and friends who have made this journey both enjoyable and enriching. I cherished the relaxing, warm, and welcoming environment filled with research seminars, coffee and beer times, and casual conversations. Whether from TUM or DLR, collaborating and working with all of you has been a true pleasure. The great times we shared at summer schools, Eisstockschießen, and Oktoberfest were not only fun but also incredibly beneficial—many of the cool research ideas that found their way into this dissertation were sparked during these events.

I especially want to express my heartfelt gratitude to my amazing office colleagues, Dorothee Stiller and Jeron Staab. Over the past six years in our beloved "ST-Büro," I couldn't have wished for better office mates. The deep friendships we've established are among the most cherished outcomes of this journey. Working alongside them has been so rewarding that it makes me wish I could remain in this environment indefinitely. Cheers to the incredible memories and the bonds we've built!

I want to express my deepest gratitude to my parents and my sister for their unwavering support and endless encouragement throughout my life. Their belief in me has been a constant source of strength, inspiring me to reach goals I once thought unattainable. Without their guidance and love, I wouldn't be where I am today.

Most importantly, to my family—the true heart of my journey—I am eternally grateful. Angi, the love of my life, has been my greatest cheerleader and confidante, always encouraging me to strive for greatness. Together, we have the incredible joy of raising our two wonderful children, Raffael and Isabella, who are the brightest lights in my life. Their smiles and laughter have been a constant source of inspiration, giving me strength during the most challenging times.

Your love and support have made this extraordinary journey both possible and meaningful. Thank you for being by my side every step of the way.

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Network
DSM	Digital Surface Model
FCN	Fully Convolutional Network
GLCM	Grey Level Co-occurrence Matrix
HR	High Resolution
IoU	Intersection over Union
LCZ	Local Climate Zone
MM	Mathematical Morphology
OBIA	Object Based Image Analysis
SDG	Sustainable Development Goal
STnet	Slum Transfer Network
VHR	Very High Resolution
XFCN	Fully Convolutional Xception Network

Contents

1. Introduction

“Spare no effort to free our fellow men, women and children from the abject and dehumanizing conditions of extreme poverty”.

These powerful words marked the opening of the 2015 Millennium Development Goals Report, reflecting the unanimous commitment of 189 countries at the Millennium Summit in September 2000 to eradicate extreme poverty and uplift humanity (United Nations, 2000).

From the Millennium Development Goals Report evolved the 17 Sustainable Development Goals (SDGs) (United Nations, 2023b). The first SDG aims to "End poverty in all its forms everywhere". With multiple ambitious targets, this goal focuses on eradicating extreme poverty, halving the number of people living in poverty, and establishing robust social protection systems for all by 2030.

Eradicating poverty is one of the greatest global challenges today, especially for developing countries. It is a multifaceted issue with roots in both national and international contexts. No single solution fits all; instead, country-specific programs, international support, and a conducive global environment are vital. The complexity of poverty threatens social cohesion, economic development, and political stability. Despite some progress, inequalities persist, achievements remain uneven or even regress (United Nations, 2023b).

Poverty has always existed in human history and is present on all continents. It manifests in various forms and types. Since poverty is present worldwide, its physical manifestations can be observed everywhere. However, there is a noticeable disparity between the Global South and the Global North (Walker, 2023).

The global landscape is undergoing a massive urban transformation driven by diverse societal phenomena. These include a rural-urban migration (Tacoli et al., 2015), housing shortages (Wang et al., 2017), inequality (Young, 2013), deportation and displacement (Lipton, 1980). Additionally climate change (Sedova and Kalkuhl, 2020), natural disasters (Hunter, 2005) and wars (Melander and Ãberg, 2007) further contribute to this transformation.

These factors have significant impacts on cities worldwide, leading to the emergence of informal, spontaneous occupied settlements, slums and other forms of urban poverty (UN-Habitat, 2020). The drivers of urbanization are as varied as the manifestations of poverty in the housing sector. In particular, these urban areas associated with poverty are considered highly dynamic compared to other parts of urban society (Kuffer et al., 2016; Taubenböck et al., 2015).

1.1. Motivation and Objectives

From the Millennium Summit in 2000 (United Nations, 2000) to today's Global SDGs Report in 2023 (United Nations, 2023b), progress is being tracked closely. Specifically, SDG1, which aims to end poverty, shows only moderate progress by 2023. Especially, the goal to eradicate extreme poverty has only seen limited advancements, with progress stagnating or even reversing between 2020 and 2023 (United Nations, 2023b).

In the United Nations call for better data, the use of geospatial, big data analysis tools and tapping new data sources like satellite imagery are specifically mentioned in order to better track the SDGs progress (United Nations, 2024). The challenge to process the big surge and demand for data has also helped to open up the gap in our understanding of the world. The 2030 Agenda motivates countries around the world to begin or to improve monitoring data from air and water quality to the prevalence of discrimination and access to electricity, sanitation and clean water, as well as census, population, and income data (United Nations, 2023c).

Another valuable source of data for tracking poverty comes from the research domain. Research studies, particularly case studies using survey or socio economic data, often provide accurate and detailed insights into poverty (Tarozzi and Deaton, 2009). These studies are typically conducted on a small scale, allowing researchers to delve deeply into specific communities or regions. However, while case studies offer a precise and nuanced understanding of poverty in the areas they cover, they are geographically limited. This means their findings might not be easily generalizable to larger populations or different regions.

How can large-scale data on the current situation of urban poverty be obtained? This is where geospatial and remote sensing data become invaluable. These technologies enable us to gather detailed information about urban areas, identifying regions affected by poverty (Kuffer et al., 2016). By utilizing satellite imagery and other remote sensing tools, the physical characteristics and infrastructure of cities can be mapped out, highlighting areas where poverty is most concentrated (Taubenböck et al., 2015). This approach helps to close the information gap, providing meaningful and reliable data about the location and extent of urban poverty. With this information, policymakers and researchers can develop more targeted and effective strategies to address the challenges of urban poverty and improve living conditions for those affected.

As remote sensing data becomes more accessible and the means of processing this data become more advanced (Shahtahmassebi et al., 2021; Weng and Quattrochi, 2006; Yin et al., 2021), this technology can be utilized to gain a comprehensive, bird's-eye view of the physical characteristics of urban poverty (Mahabir et al., 2018). This allows us to analyze urban landscapes with neutral, coherent, and unobstructed precision. Remote sensing data is the first pillar needed for large-scale analysis to detect the physical presence of urban poverty settlements. Equally important is the second pillar: the means to process this data and derive meaningful information from it. Recent advancements in machine learning, particularly the increased usage of deep learning methods in remote sensing applications, play a crucial role (Ma et al., 2019; Zhu et al., 2017). One of the key advantages of deep learning is its ability to generalize well even in complex settings (Lunga et al., 2021; Maxwell et al., 2022). This capability is essential for detecting the diverse physical characteristics of urban poverty across different geographical regions.

In this dissertation, the aim is to apply deep learning methods to detect the physical presence of urban poverty from remote sensing data and answer multiple questions: Can deep learning methods effectively map urban poverty, and which types of remote sensing data are best suited for this task? What challenges arise in varied and complex slum environments, and how can they be overcome?

The objective is to develop advanced deep learning methods capable of handling diverse conditions and geographical settings, ensuring a comprehensive and scalable approach to mapping urban poverty. This dissertation seeks to address these questions and aims to contribute to the effective identification and analysis of slum areas on a global scale.

1.2. Outline of this Thesis

This cumulative dissertation comprises four journal papers, included in Appendix A.1, A.2, A.3, A.4 with each paper addressing one of the aforementioned research questions while progressively building on the findings of the preceding studies.

In chapter 1, the current state of urban poverty, its drivers, and the impact of the SDGs are presented. The motivations and objectives highlights the need for better, more reliable data and how deep learning methods can help for large scale poverty detection applications.

Chapter 2 presents an overview of the use of remote sensing data for mapping poverty. It explores the key terminology related to urban poverty manifestations and analyzes the physical characteristics of slums that can be detected through remote sensing. Additionally, it provides a comprehensive review of studies that have utilized remote sensing datasets to map and identify slum settlements, while also addressing the challenges associated with accurately detecting slums using remote sensing techniques.

Chapter 3 provides a summary of the four contributed journal publications. The first discusses the transfer learning capabilities of FCNs for slum mapping across various satellite images. The second introduces a transfer-learned XFCN model, capable of distinguishing between formal built-up areas and different categories of slums in high resolution (HR) satellite data, using a large sample of slums from globally distributed cities. The third publication presents efficient methods for slum detection by applying transfer learning with minimal sample sizes and estimating prediction probabilities. Lastly, the fourth explores advanced machine learning techniques and uncertainty-aware approaches for mapping slum areas across 55 heterogeneous cities, contributing to a deeper understanding of global slum morphologies.

Chapter 4 critically examines the limitations of the contributions, whether due to methodological constraints or the nature of the data sources. It also discusses the broader implications for the SDGs, non-governmental organizations (NGOs), and the scientific community. Additionally, the chapter addresses the ethical considerations, not only regarding the use of deep learning itself but particularly in the context of slum mapping, highlighting the potential risk of stigmatizing underrepresented social groups.

Chapter 5 presents a conclusion and outlook on the topic of mapping global urban poverty using deep learning methods.

2. Mapping Poverty from Space

2.1. Indicators and the Concept of Extreme Poverty

Extreme poverty is a complex concept with varying definitions depending on the perspective or context in which it is considered. The United Nations define it as severe deprivation of basic human needs, including food, safe water, sanitation, health, shelter, education, and information, not solely based on income but also on access to essential services (United Nations, 2023a). The World Bank sets the threshold for extreme poverty at living on less than $1.90 per day, adjusted for purchasing power parity, which serves as a global benchmark (World Bank, 2023).

The United Nations Development Programme (UNDP) broadens this definition to include not just low income but also lack of access to basic services, social exclusion, and vulnerability to shocks like disease and natural disasters. This perspective is reflected in the Multidimensional Poverty Index (MPI), which also measures factors like health, education, and living standards (UNDP, 2023). Similarly, the Oxford Poverty and Human Development Initiative (OPHI) defines extreme poverty as being deprived in one-third or more of the indicators used in the MPI (OPHI, 2023).

The World Health Organization (WHO) focuses on the health aspect, defining extreme poverty in terms of lack of access to essential health services (WHO, 2023). In the European Union, extreme poverty is sometimes understood in relative terms, such as living on less than 40% of the median income, though this relates more to relative poverty (European Union, 2023). These definitions highlight that extreme poverty is not just about income but also about access to essential services and overall well-being.

While the goal is not to revisit the various established definitions of extreme poverty, inequality, and social segregation or to examine poverty according to a precise definition, it is important to understand their differences and limitations, as extreme poverty can manifest in various forms across different countries around the globe.

2.2. The Morphological Slum

All the aforementioned indicators predominantly focus on economic or social metrics, such as income levels, employment rates, educational attainment, and social integration. The built environment, however, is mentioned only in a limited number of parameters and often receives less emphasis in analytical frameworks. In this context, the term "built environment" refers to the human-made surroundings that provide the setting for human activity, including buildings, infrastructure, and urban layouts. Despite its minimal direct consideration, the built environment is indirectly utilized as a means to locate and identify individuals residing in more impoverished or disadvantaged areas. By analyzing aspects of the built environment, such as housing quality, infrastructure availability, and spatial organization, it becomes possible to infer the socioeconomic conditions of a population. Consequently, the physical characteristics of the built environment serve as valuable proxies for detecting and studying communities affected by poverty. This indirect utilization underscores the importance of incorporating spatial and environmental factors into assessments of social and economic well-being.

It should be emphasized that the aim of this approach is to systematize the morphological characteristics of housing types within slum settlements rather than focus on their statistical aspects of extreme urban poverty from the previous section. The methodological concept, is based on morphological features of slum settlements.

2.2.1. The Physical Appearance of Slum Settlements

Slum settlements and formal built-up areas are characterized by distinctly different morphological features, as extensively documented in various studies (Baud et al., 2010; Kuffer et al., 2014, 2016; Taubenböck et al., 2018). In slum areas, buildings are typically small, substandard, and densely packed, lacking the spatial organization and infrastructure found in formal settlements. These areas often consists of very high roof coverage densities, with minimal or no public or green spaces, contrasting sharply with the lower density and well-planned layouts of formal neighborhoods that include provisions for parks and recreational areas (Debray et al., 2023).

The layout of slum settlements is often chaotic, with organic, irregular patterns that lack orderly road arrangements and fail to comply with setback standards. This disordered structure is a far cry from the regular, systematic layouts of formal areas, where roads are well-organized and setback rules are strictly followed (Taubenböck et al., 2020). Furthermore, slums are frequently located in hazardous environments, such as flood-prone zones, near industrial sites, or on steep slopes (Kühnl et al., 2023; Müller et al., 2020). These locations, while sometimes offering proximity to infrastructure and economic opportunities, pose significant risks to the residents' safety and well-being. In contrast, formal settlements are typically established on land that is suitable for construction, equipped with basic infrastructure, and free from significant environmental hazards.

The physical appearance of slum settlements is presented in Table 1 in a comparative analysis of the morphological features between slum settlements and formal settlements (Baud et al., 2010; Kuffer et al., 2014, 2016; Taubenböck et al., 2018). While this characterization applies to most cases, there are exceptions where slum settlements may not exhibit all these features (Debray et al., 2023). Slum settlements are characterized by smaller, substandard building sizes, high building density, and a lack of public or green spaces. Their organic layout results in irregular road patterns that do not conform to standard setback regulations. Furthermore, slums are often located in hazardous areas such as flood-prone zones or steep slopes, though they are in close proximity to infrastructure and livelihood opportunities. In contrast, formal settlements exhibit larger building sizes, lower to moderate density, and provision of public or green spaces. The pattern layout in formal areas follows a planned, regular structure with compliance to setback rules, and the land is suitable for construction with access to basic infrastructure. This table highlights the distinct morphological differences between these two settlement types.

Table 1 Comparison of morphological features between slum settlements and formal settlements (adapted from Baud et al. (2010); Kuffer et al. (2014, 2016); Taubenböck et al. (2018))

Features	**Slum Settlements**	**Formal Settlements**
Building Size	Small, substandard building sizes	Generally larger building sizes
Building Density	Very high roof coverage, with a lack of public or green spaces within or near slum areas	Low to moderate density with provision of public or green spaces within or near planned areas
Pattern Layout	Organic layout with irregular road arrangements and non-compliance with setback standards	Regular layout with planned roads and adherence to setback rules
Site Characteristics	Often located in hazardous areas (e.g., flood-prone zones, near industrial sites, or on steep slopes) with proximity to infrastructure lines and livelihood opportunities	Land is suitable for construction, with basic infrastructure provided

2.2.2. Slum Categories

Based on the detailed analysis presented in Table 1, slum settlements can be categorized into several distinct groups, each defined by specific morphological characteristics that influence the spatial and structural dynamics of these areas (Taubenböck et al., 2018). These categories reveal the varying degrees of organization, density, and development found within slum settlements, reflecting the diverse ways in which these communities are established and evolve across different regions.

Category 1 (C1) Morphologic Slums: This category includes neighborhoods that align most closely with the extreme characteristics of slum morphologies. These areas are characterized by small, makeshift shelters that are densely packed and arranged in complex, often chaotic patterns. The disorganized nature of these settlements is a defining feature, with little to no formal planning. The dense and often haphazard arrangement of buildings results in very high population densities, with limited access to basic services or infrastructure. Examples of areas that fall into this category include some of the most notorious slums in the world, which are often depicted as epitomes of urban poverty and informality. C1 is characterized by exhibiting all four features outlined in Table 1.

Category 2 (C2) Mixed Structured Slums: These neighborhoods exhibit significant deviations from the typical slum morphology in more than one physical aspect but still retain a closer resemblance to slum characteristics than to formalized urban areas. These mixed forms might include areas that have undergone some degree of development or improvement. Alternatively, they may consist of older, deteriorating urban blocks that have begun to resemble slum conditions due to neglect or economic decline. The structures in these areas are often a blend of informal and formal elements, resulting in a patchwork of building types and layouts. Despite the deviations from the typical slum morphology, the overall appearance and functionality of these neighborhoods still align more closely with slum-like conditions than with those of structured, planned neighborhoods. C2 displays many of the features, specifically three out of four, as described in Table 1.

Category 3 (C3) Untypical Slums: This category encompasses neighborhoods that range from areas exhibiting a mix of structured and unstructured forms to fully structured, formalized urban areas. These neighborhoods often represent transitional spaces where slum-like conditions are gradually integrated into more formal urban planning schemes. In some cases, these areas have undergone significant redevelopment, leading to the incorporation of formal, planned elements such as geometric street layouts and lower building densities. However, they may still retain some characteristics of informal structured slums. C3 encompasses only a subset of the features, meeting two out of the four criteria listed in Table 1.

Overall, these categories underscore the complexity and diversity of slum settlements and highlight the varying degrees of formality, development, and planning that can exist within them. Each category represents a different stage or type of urban evolution and cultural background, reflecting the dynamic nature of urban poverty and the ongoing challenges in addressing it through urban planning and policy (Taubenböck and Kraff, 2014). These distinctions are essential for understanding the spatial and structural characteristics of slum areas and for developing targeted methods capable of effectively distinguishing the sometimes blurred feature space between slums and formal settlements.

2.3. Methods Using Remote Sensing Data to Map Slums

Slum mapping has evolved from local and labor-intensive surveys (Baud et al., 2009; Weeks et al., 2007) to more advanced and large scale analysis using remote sensing data (Kuffer et al., 2016). Remote sensing is a powerful tool for detecting and analyzing the physical features of urban areas, including those associated with poverty, as outlined in Table 1 (Esch et al., 2010a). By providing a bird's-eye view of the Earth's surface, remote sensing allows for monitoring local scale to large-scale views of the urban environments. This technology utilizes various sensing methods, including active and passive systems, which capture data through different means such as reflected sunlight or emitted energy. Additionally, remote sensing can be configured with varying geometric and radiometric resolutions, enabling detailed analysis of spatial patterns and material properties. A wide variety of sensors, from optical systems to radar systems at different geometric resolutions, can be employed to identify and map indicators of urban poverty, such as building density, layout patterns, and infrastructure conditions (Kuffer et al., 2016). These capabilities make remote sensing an invaluable asset in urban studies, particularly in the context of understanding and addressing poverty in urban areas.

2.3.1. Object Based Methods

In object-based image analysis (OBIA), images are treated as compositions of distinct objects characterized by attributes such as size, shape, texture, and relationships with neighboring objects (Giada et al., 2003). One of the earliest studies applying OBIA to slum detection was conducted by Hofmann et al. (2001), where multi-resolution analysis was employed to segment Ikonos satellite imagery at various spatial scales in Cape Town, South Africa. Objects identified at different scales were linked using a class hierarchy, with larger super objects such as informal settlements, encompassing smaller descriptors like physical characteristics (e.g., dwelling size) and contextual elements (e.g., texture). The classification process utilized a set of fuzzy logic rules to describe these object-based characteristics.

Subsequent studies following the approach introduced by Hofmann et al. (2001) have applied OBIA for slum identification, albeit with some modifications in segmentation parameters to account for the varying physical characteristics of slums (Escalante, 2012; Kit et al., 2012; Rhinane et al., 2011). The primary difference across these studies lies in the parameterization of the segmentation process, which reflects the physical diversity of slum environments. While many studies have employed OBIA for slum detection, only a few, such as Novack and Kux (2010), have attempted to automate the selection of segmentation parameters.

Despite its advantages, OBIA, like other techniques used to extract information on slums, presents several challenges. One common issue is the presence of vegetation and shadows, which can obscure dwellings and reduce the accuracy of slum detection (Novack and Kux, 2010). Another challenge stems from the materials used in slum construction, which often generate high spectral noise. For instance, unpaved roads may exhibit similar spectral reflectance as slum rooftops, complicating the differentiation of individual dwellings. Moreover, the rules developed to extract slum features tend to be image-specific, limiting their transferability to other geographic regions (Duque et al., 2015; Owen and Wong, 2013). As a result, automated OBIA methods often yield suboptimal results in dense urban areas due to the significant intra- and inter-slum variability (Kit et al., 2012).

2.3.2. Mutli Scale Approaches

Multi-scale approaches analyze slums by examining features that emerge across different spatial scales, with common methods including fractal and lacunarity measures. Fractal geometry assesses the geometrical complexity of slum shapes, while lacunarity evaluates the internal heterogeneity by analyzing the distribution of open spaces between features (Filho and Sobreira, 2005; Kohli et al., 2013). Studies in Quezon City, Philippines, and in Recife, Brazil, have applied these methods to distinguish between different types of settlements (Barros Filho and Sobreira, 2008; Galeon, 2011).

Most studies on multi-scale approaches for slum detection have focused on large cities, where slums can be more easily distinguished from other settlement types. However, this distinction may not be as apparent in smaller cities in developing countries, where urbanization challenges are expected to be particularly severe (Leao and Leao, 2011). Given these limitations, further research is needed to assess the effectiveness of multi-fractal approaches for slum analysis, especially since not all features in an image exhibit self-similarity (Barros Filho and Sobreira, 2008).

In terms of monitoring slum growth, multi-scale methods are particularly effective for tracking slum development during the consolidation and maturity phases. However, these approaches tend to lose their effectiveness when feature density decreases at the local level, limiting their ability to capture early-stage slum formation (Thomas et al., 2008). Therefore, while multi-scale techniques offer some advantages, they may require refinement and further investigation to ensure their broader applicability, particularly in smaller urban contexts and at different stages of slum development.

2.3.3. Image Texture Analysis

Image texture analysis, which examines the repeated variations in intensity and color within images, is a key method for inferring structural and spatial patterns in slum areas. This process, however, is inherently complex due to the scale-dependent nature of image features (Su and Hu, 2004; Valous et al., 2010). One approach, known as mathematical morphology (MM), is often employed to refine outputs from binary images. In the context of slum detection, MM is commonly used to eliminate unwanted artifacts such as trees, fences, or other features that could interfere with the analysis (Rhinane et al., 2011; Sulik and Edwards, 2010).

In contrast to structural methods, statistical approaches analyze the spatial relationships and pixel intensities to detect patterns within groups of pixels. One widely used technique is based on the Grey Level Co-occurrence Matrix (GLCM), which evaluates texture by measuring statistical relationships between pixel values (Haralick et al., 1973). For instance, Kohli et al. (2013) employed GLCM texture measures, such as entropy, contrast, variance, and mean, to extract slums at the settlement level, as these metrics are well-suited to identifying the high density of dwellings typical of slum areas. Similarly, studies by Kuffer et al. (2015); Wurm et al. (2017) found that slums in Mumbai, India, exhibited significantly lower variance in texture values compared to formal settlements. Alternatively, Stasolla and Gamba (2008) proposed the use of autocorrelation texture measures to distinguish between slums and formal settlements using radar data, highlighting a different statistical approach to texture analysis.

Despite its utility, image texture analysis faces several limitations when applied to slum detection and mapping. One challenge is the variability of texture measures across different slum areas, influenced by factors such as building materials and spatial resolution, which can complicate the interpretation of results (Schmitt et al., 2018). Structural approaches, such as those using MM operations like dilation, may struggle to differentiate between individual dwellings in densely packed slum areas, where buildings are located close together. Additionally, these structural methods often rely on scene-specific rules to extract certain features, which limits the generalizability of the techniques across different geographic regions or datasets (Soille and Pesaresi, 2002).

In summary, while both structural and statistical texture analysis methods provide valuable tools for slum detection, they are constrained by scale dependencies, local conditions, and the need for scene-specific adaptations. Further research is needed to improve their accuracy and applicability in diverse urban contexts.

2.3.4. Landscape Analysis

Landscape analysis uses quantitative metrics to describe the spatial patterns of land cover, focusing on the composition and configuration of patches, adjacent pixel regions with uniform land cover. This approach has been applied to study slums by creating indices like the Unplanned Settlement Index (USI), which evaluates metrics such as patch density, contagion, and aggregation (Kuffer et al., 2014). Studies in New Delhi, India, and Dar es Salaam, Tanzania, have shown that different landscape metrics are more effective in different contexts (Kuffer et al., 2014). However, the use of landscape metrics also presents challenges, including dependency on initial spectral characterization and potential correlations between metrics, which complicate their selection and application (Huang et al., 2007). Furthermore, the accuracy of landscape metrics depends on the homogeneity of the input data, and their effectiveness may vary with changes in scale and spatial resolution (Liu et al., 2006).

2.3.5. Single Building Detection

Building feature extraction, often utilizing digital surface models (DSMs), focuses on identifying individual dwellings in slum areas by analyzing height data. This approach assumes that the vertical dimension can help distinguish slum dwellings from other features, resulting in 3D models of slums (Temba et al., 2015). Techniques like stereo image matching and active contour models (snakes) have been used to derive DSMs and extract building features (Rüther et al., 2002).

However, this approach may face challenges in areas where slum settlements are multiple stories high, such as in Medellín (Kühnl et al., 2023), or in high-rise poverty settlements as seen in Shenzhen (Pan and Du, 2021). In such cases, the vertical complexity of the structures can make it difficult to distinguish between slum and formal housing using DSM data alone.

2.3.6. Deep Learning Techniques

Since the release of AlexNet in 2012 from Krizhevsky et al. (2012), the field of machine learning has experienced a profound shift from traditional methods to the widespread adoption of neural networks, particularly CNNs. AlexNet's revolutionary performance in the ImageNet competition demonstrated the immense power of deep learning, particularly in complex tasks such as image classification. Its ability to automatically learn hierarchical features directly from raw data, combined with its scalability, allowed it to far surpass traditional methods, reshaping the landscape of machine learning. The field of deep learning is constantly evolving, with new architectures, incremental improvements in layers, and novel learning approaches emerging regularly. Since AlexNet, significant advancements have been made in the field of urban remote sensing task, such as scene classification tasks (Aravena Pelizari et al., 2023; Mou et al., 2017; Qiu et al., 2019), population regression estimation (Doda et al., 2024), semantic segmentation of land-use and land-cover classes (Chen et al., 2020; Mou et al., 2020; Wurm et al., 2021), object detection methods (Li et al., 2020), and instance segmentation of single buildings (Schuegraf et al., 2024). Even small modifications in architecture or training methods can lead to substantial improvements in performance, further demonstrating the dynamic nature of this field.

With the rise of deep learning across various scientific disciplines, its application in slum detection has also gained momentum. Different approaches using various deep learning methods enhance the accuracy and efficiency of identifying informal settlements.

In Jean et al. (2016), survey data combined with satellite imagery from five African countries, Nigeria, Tanzania, Uganda, Malawi, and Rwanda was used to demonstrate how a CNN can be trained to identify image features that explain up to 75% of the variation in local economic outcomes. The study addresses data scarcity through a multi-step transfer learning approach, where a readily available but noisy proxy for poverty, nighttime lights, is leveraged to train a deep learning model. This trained model is then applied to estimate either average household expenditures or average household wealth at the neighborhood level, providing valuable insights into economic conditions in data-limited regions.

Very high resolution (VHR) remote sensing data has demonstrated significant success in classification tasks, as illustrated by Mboga et al. (2017), where a patch-based CNN approach achieved high accuracy using QuickBird imagery from Dar es Salaam, Tanzania. Similarly, Verma et al. (2019) applied transfer learning with CNNs to both VHR and HR satellite imagery, yielding overall higher accuracies for VHR data than when using HR data, when validated against manually delineated slum boundaries in Mumbai. Furthermore, Prabhu et al. (2021) proposed a dilated kernel-based deep CNN (DK-DCNN) for slum detection in South India, incorporating post-processing through morphological spatial pattern analysis to enhance accuracy. Additionally, Persello and Kuffer (2020) explored CNNs for identifying socio-economic variability within poor neighborhoods in Bangalore, India. Their model, pretrained on a slum classification dataset, was fine-tuned to predict a continuous socio-economic index, effectively capturing multiple levels of deprivation.

These studies highlight the growing potential of deep learning techniques in leveraging satellite imagery to address urban challenges, particularly in resource-constrained environments. Similarly Ajami et al. (2019) addresses the challenge of identifying slum variations in Bangalore, India, with an integrated approach that combines VHR satellite images with socio-economic data. By applying multiple correspondence analysis (MCA) and a data-driven index of multiple deprivation (DIMD), the study predicts slum deprivation levels. A two-step transfer learning approach was used and best results were achieved using an ensemble model.

Semantic segmentation techniques are increasingly employed to map and analyze informal settlements, with many methods utilizing variants of the U-Net architecture (Ronneberger et al., 2015). For instance, Dufitimana and Niyonzima (2023) applied a MobileNetV2 U-Net for mapping informal settlements in Kigali, Rwanda, using VHR satellite data. By incorporating dilated convolutional operations, the model effectively distinguished informal settlements from other urban areas, enhancing spatial accuracy. Similarly, Dabra and Kumar (2023) explored the detection of green and open spaces within informal settlements in Mumbai. Their approach involved training three modified CNNs, VGG16 U-Net, MobileNetV2 U-Net, and DeepLabV3+, on HR imagery, achieving high precision in urban slum mapping.

Gram-Hansen et al. (2019) demonstrated the successful use of HR data and DeepLabV3+ to detect informal settlements across diverse geographic contexts, including Kenya, South Africa, Nigeria, Sudan, Colombia, and Mumbai. Likewise, Fisher et al. (2022) utilized HR Sentinel-2 multispectral data to map slums in Mumbai using a U-Net architecture combined with Monte Carlo Dropout for uncertainty estimation. This method not only provided accurate slum delineations but also incorporated uncertainty estimates to improve the robustness of the results.

The question of optimal architecture and backbone combinations for slum mapping was investigated by Lumban-Gaol et al. (2023), who found that a Feature Pyramid Network (FPN) with a VGG16 backbone yielded the best performance for this task. Despite promising results, the variation in slum characteristics across different regions continues to hinder generalization of these methods in diverse urban areas.

Additionally, Persello and Stein (2017) introduced a custom FCN for detecting informal settlements in Dar es Salaam, Tanzania, using VHR imagery. This approach outperformed standard patch-based architectures, such as the one used by Mboga et al. (2017), both in processing efficiency and accuracy. These studies illustrate the growing potential of deep learning models, particularly those based on semantic segmentation, to enhance the mapping and analysis of informal settlements in urban environments, though challenges remain in achieving consistent performance across varied slum characteristics and geographic contexts.

2.4. Remote Sensing Data and Reference Data

2.4.1. Relevant Remote Sensing Data for This Work

QuickBird is a VHR Earth observation satellite that offers a pansharpened geometric resolution of 0.5 meters. This high level of detail makes QuickBird particularly useful for urban mapping, allowing for the identification of small-scale features such as individual buildings, road networks, and even vegetation within urban environments. Due to its fine spatial resolution, QuickBird is widely used in applications requiring detailed imagery, such as land cover classification (Lu et al., 2010), forest tree species (Abdollahnejad et al., 2017), and disaster management (Pradhan et al., 2016). The VHR imagery helps in detecting subtle changes in urban morphology, which is critical for analyzing urban poverty and other socio-economic conditions.

PlanetScope is a constellation of small satellites operated by Planet Labs, providing daily imagery with a geometric resolution of 4.77 meters. While not as detailed as QuickBird, PlanetScope offers frequent coverage, making it ideal for monitoring dynamic urban areas where changes occur rapidly (Gosteva et al., 2019; Wang et al., 2022a). Its HR imagery is suitable for identifying broader urban patterns, such as land use changes, and overall urban sprawl (Frazier and Hemingway, 2021).

Sentinel-2 is part of the European Space Agency's Copernicus program, offering multispectral imagery with a geometric resolution ranging from 10 to 60 meters, depending on the spectral band. Sentinel-2's wide coverage and multi-spectral capabilities make it useful for a variety of applications, including vegetation monitoring (Rußwurm and Körner, 2018), land cover classification (Zhu et al., 2019), and large-scale urban analysis(Qiu et al., 2020). While its resolution is lower than QuickBird and PlanetScope, Sentinel-2's ability to capture data across multiple spectral bands allows for the detection of broader environmental and urban characteristics.

TerraSAR-X is an active remote sensing satellite that uses synthetic aperture radar (SAR) technology to capture HR imagery with a geometric resolution of 6 meters. It operates with multiple polarization modes, which enhance its ability to detect and analyze surface features under various conditions, including during the night and through cloud cover. TerraSAR-X is particularly effective for mapping urban areas, identifying structural details, and monitoring changes in built environments (Esch et al., 2010b). Its ability to penetrate through weather and lighting conditions makes it a reliable tool for consistent urban monitoring, especially in assessing infrastructure and detecting signs of urban poverty, such as the expansion of informal settlements (Wurm et al., 2017).

2.4.2. Reference Data Related to Slum Mapping

Ground truth labels and masks are critical for deep learning methods, serving as reference data for training and evaluating models. Several studies have utilized manual annotation of satellite images, involving human experts to identify and delineate urban villages, slum regions, and areas of deprivation (Friesen et al., 2024; Huang et al., 2023; Owusu et al., 2024; Wahbi et al., 2023). This manual process ensures high accuracy and reliability of reference data, which is essential for robust model development.

In addition, multiple studies have integrated various data sources, including remote sensing imagery, street view data, and social sensing information, to create comprehensive labels (Ajami et al., 2019; Chen et al., 2022; Najmi et al., 2022). Combining these diverse datasets enhances the richness and precision of the labels, providing more detailed insights into urban environments. Expert knowledge plays a crucial role in this process, offering valuable perspectives on the distinct characteristics of urban areas, thus improving the reliability and accuracy of the generated labels.

The quality of reference data is crucial for deep learning model performance because it directly impacts the accuracy of training, validation, and testing phases, guiding the model to learn relevant patterns. Poor or noisy reference data can introduce biases or errors, reducing the model's ability to generalize and produce reliable predictions (Mishkin et al., 2017; Zhang et al., 2021). When developing deep learning approaches, especially for urban applications, it is essential to account for the potential noise or coarseness in the resolution of the data. Addressing these challenges ensures more reliable model outcomes, particularly when dealing with complex and heterogeneous urban environments (Zhu et al., 2017).

2.5. Challenges in Slum Mapping

With the continuous advancements in slum mapping, particularly with the advent of deep learning techniques, significant improvements in accuracy and scalability have been achieved. The integration of deep learning has enhanced the generalizability of urban remote sensing tasks, making large-scale applications more feasible (Ma et al., 2019; Qiu et al., 2019; Zhu et al., 2017). Moreover, the increasing availability and accessibility of remote sensing data have further supported these developments. However, the application of remote sensing data for slum mapping still presents numerous challenges. These challenges stem from the complexity and heterogeneity of slum environments, the variability in data quality and resolution, and the need for more sophisticated algorithms to effectively capture the unique characteristics of slum areas (Kuffer et al., 2014; Taubenböck et al., 2018). In the following, some of the major challenges associated with slum mapping using remote sensing data are addressed.

Slum morphology is typically characterized by high population density and the prevalence of small, often irregularly shaped buildings (Friesen et al., 2018; Kraff et al., 2020). These physical features shown in Table 1 present unique challenges for detection and mapping, necessitating the use of remote sensing data capable of accurately capturing such fine details. VHR satellite data is generally preferred for this purpose, as it can detect small and densely packed structures effectively (Verma et al., 2019). However, HR data can also be utilized, especially for mapping larger slum areas where the spatial patterns are more discernible even at a slightly lower resolution. The choice between VHR and HR data largely depends on the size and complexity of the slum being studied; while VHR data is ideal for small, intricate slum environments, HR data may suffice for larger slums where the general patterns are more apparent (Kuffer et al., 2016).

The reliance on remote sensing data introduces another significant challenge: the availability and quality of labeled data for training deep learning models in slum mapping. Although deep learning techniques have substantially improved accuracy compared to traditional machine learning methods, they are inherently data-intensive. To achieve robust generalization across diverse slum environments, large amounts of high-quality, accurately labeled data are required (Stark et al., 2023). However, acquiring such data is challenging due to the heterogeneous nature of slums and the variability in their geographic locations. Effective training datasets must encompass a wide range of geographical contexts to ensure that models do not overfit to specific regions but instead generalize well across different urban landscapes (Zhu et al., 2017).

One of the primary difficulties in compiling these datasets is the small physical footprint of many slums, which, despite housing large populations, occupy relatively small land areas (Friesen et al., 2018). Notable exceptions exist, such as Kibera in Nairobi or the slums of Mumbai, which are expansive enough to be well-documented (Kraff et al., 2019; Taubenböck and Kraff, 2014). However, for many smaller, less prominent slums, data is exceedingly scarce and difficult to obtain. Additionally, slums are dynamic entities; they frequently undergo significant changes in structure and appearance over time, or they may be subject to eviction and redevelopment (Liu et al., 2019). Therefore, the temporal alignment between the remote sensing data and the labeled ground truth data is crucial to ensure accuracy in mapping and monitoring these areas. Figure 1 illustrates a comparative analysis of settlement structures in Delhi using QuickBird VHR data, pansharpened to 0.5 m, and PlanetScope HR data at 4.77 m. Three areas of interest are highlighted with corresponding Google Street View imagery. While VHR data enables discernment of these settlement structures, distinguishing them remains challenging due to their compact and irregular layout. The difficulty significantly increases with HR data, where the spatial resolution is insufficient to accurately separate individual structures. This comparison underscores the limitations of lower-resolution imagery for detailed urban structure analysis and highlights the need for VHR data when addressing complex settlement differentiation tasks in dense urban environments.

Acquiring high-quality ground truth data in smaller or more unknown urban areas, poses yet another challenge. While prominent slums in cities like Mumbai, Nairobi, Lagos, and Caracas are relatively well-documented, many smaller slums in less prominent cities remain underrepresented in available datasets (Kuffer et al., 2016). These overlooked areas could account for a significant portion of global slum populations, as the lack of comprehensive statistics from smaller cities conceals their true scale and impact. The scarcity of accurate and up-to-date ground truth data could hamper the ability to effectively apply deep learning models in smaller urban centers (Whang et al., 2023).

A further challenge lies in the diversity of slum categories and their spatial relationship with formal settlements. Slums are not homogeneous; their morphology can vary significantly not only between different cities (inter-urban variability) but also within a single city (intra-urban variability) (Taubenböck et al., 2018). While many urban forms can be clearly classified into well-defined categories, such as those outlined in the Local Climate Zone (LCZ) classification scheme (Stewart and Oke, 2012), slums are highly dynamic and often defy easy categorization (Liu et al., 2019). The LCZ scheme divides urban areas into ten distinct settlement types, considering factors such as building height (high, mid, and low-rise buildings), building density, and the presence of green cover within these settlement types. However, slums frequently fall outside these clear-cut classifications due to their fluid and evolving nature, making them difficult to categorize within this framework. Within the LCZ classification schema, only class 7, characterized by a dense mix of single story buildings constructed from lightweight materials, can be considered comparable to some slum morphologies.

Two specific issues arise in this context: First, formal settlements can sometimes exhibit physical characteristics similar to slums, such as high density and irregular building patterns (e.g. LCZ class 7), as seen in Figure 1 in Delhi in the area of interest highlighted in (2) and (3). Here This similarity can lead to misclassification, as the algorithms may struggle to distinguish between densely packed formal settlements and slum areas. Conversely, some slum areas may adopt characteristics of formal settlements, as observed in

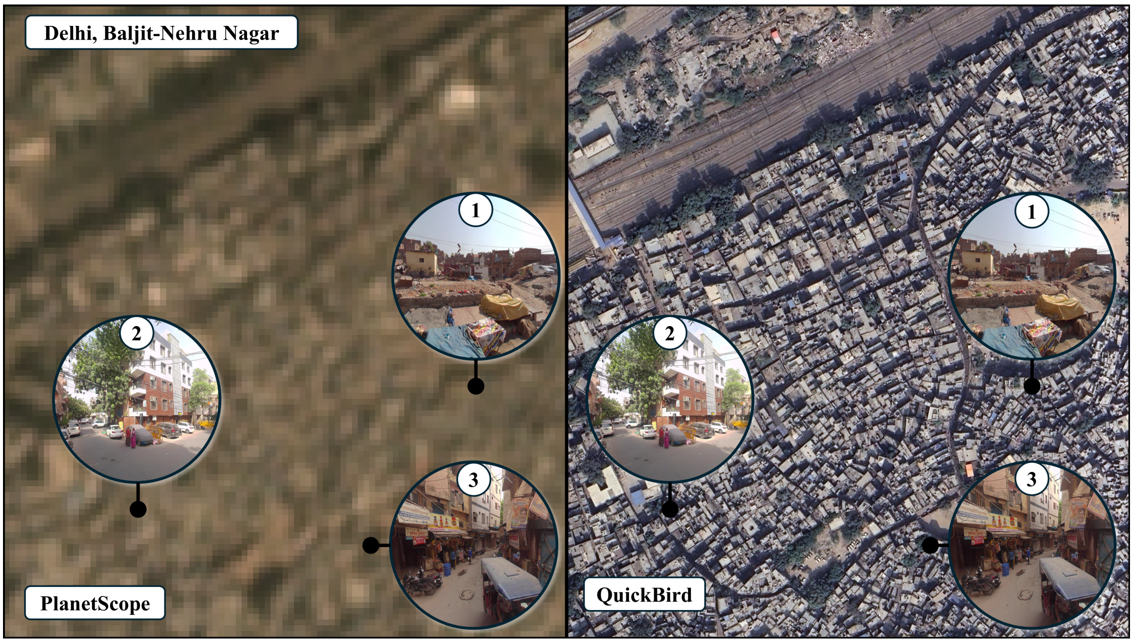

Figure 1 Comparison of three areas of interest in Delhi using VHR QuickBird and HR PlanetScope scenes. All areas exhibit dense and irregular layouts; however, only the first area qualifies as a slum, while the other two display more formal settlement structures.

certain South American favelas, where improved housing and infrastructure blur the lines between informal and formal settlement types (Kühnl et al., 2023; Wurm and Taubenböck, 2018).

Moreover, the boundaries between slums and formal settlements are rarely distinct, often blending gradually from one to the other. This gradual transition presents a significant challenge for accurate classification and mapping (Dovey and Kamalipour, 2017). The intermingling of slum and formal settlement characteristics can result in ambiguous zones that are difficult to categorize using standard classification methods. The spatial and morphological overlap between these areas complicates the development of automated detection and classification algorithms, requiring more sophisticated techniques capable of handling such nuanced and dynamic urban landscapes (Chen et al., 2021).

3. Summary of the Contributions

This chapter highlights the key contributions of this dissertation, which are based on four journal articles, including three first authorships and one second authorship.

Section 3.1 focuses on the transfer learning capabilities of FCNs for slum mapping across different satellite sensors. A model initially trained on QuickBird optical satellite imagery was adapted to lower-resolution Sentinel-2 and TerraSAR-X data. While TerraSAR-X did not show significant improvement due to its unique image characteristics, the transfer from VHR to high resolution achieved high accuracy.

In Section 3.2 a XFCN for distinguishing between formal built-up areas and various slum categories using HR PlanetScope data is presented. The XFCN, trained on a diverse global sample of slums, effectively managed to distinguish between heterogeneous morphological features of slums and formal settlements.

Section 3.3 covers learning with minimal samples and estimating probabilities for slum prediction. By incorporating Monte Carlo dropout, the study improved classification performance in noisy datasets while also assessing prediction uncertainty. The custom CNN STnet model developed here matched the performance of well-known models like ResNet50 and Xception, offering greater efficiency in training and inference, particularly with limited data.

Finally, Section 3.4 discusses advanced machine learning techniques and uncertainty-aware methodologies to map slum areas across 55 diverse cities. This approach effectively managed the challenges posed by limited labeled data, producing robust slum probability maps that offer a nuanced understanding of slum distributions. The study also incorporated test-time dropout and augmentation to estimate uncertainty in slum predictions.

3.1. Slum Mapping in Mumbai Using Transfer Learning Within Different Remote Sensing Datasets

3.1.1. The Feasibility of Slum Mapping Using Remote Sensing Data

The primary objective of this study is to explore the feasibility of accurately mapping slums using semantic segmentation techniques under ideal conditions, such as those found in the slums of Mumbai. Mumbai's slums present the typical morphological structure characterized by extremely dense settlements, small non-concrete houses made from lightweight materials, heterogeneous building alignments, and irregular road networks that are either inaccessible by vehicles or nonexistent (Dabra and Kumar, 2023; Kuffer et al., 2015; Taubenböck and Kraff, 2014). These unique features make Mumbai an ideal case study for testing the capabilities of semantic segmentation in detecting and mapping slum areas.

The secondary research question aims to determine the most suitable remote sensing data for slum mapping, with a particular emphasis on efficiency and the dataset size required for training deep learning models from scratch. Given the scarcity of labeled data and the challenges in acquiring such data for slum areas, transfer learning becomes a crucial strategy. Transfer learning allows the adaptation of pre-trained models on similar tasks, which can significantly reduce the amount of data and computational resources required for effective training (Gopalakrishnan et al., 2017; Zhu et al., 2017).

In this context, three different remote sensing data sources are available for Mumbai: VHR QuickBird imagery, HR Sentinel-2 data, and HR TerraSAR-X data. Each of these data sources offers distinct advantages and limitations. QuickBird imagery, with its very high spatial resolution, can capture fine details of slum morphology (Mboga et al., 2017). Sentinel-2 offers the potential for global coverage and easily accessible data acquisition, while TerraSAR-X radar data can penetrate through cloud cover and provide complementary information on surface roughness (Verma et al., 2019; Wurm et al., 2017). The study will assess the performance of semantic segmentation models using these data sources individually and in combination transfer learning between image sensors, aiming to determine the optimal approach for efficient and accurate slum mapping in urban environments like Mumbai.

3.1.2. Transfer Learning Between Different Remote Sensing Datasets

For the experiments, three different sensors, QuickBird, Sentinel-2, and TerraSAR-X, are used, with their distinct specifications presented in Table 2 and example imagery in Figure 2. The primary dataset is derived from QuickBird, providing pansharpened images at a 0.5-meter resolution. A false color composite of green, red, and near-infrared bands is utilized. Sentinel-2 offers high-resolution optical imagery across 12 spectral and thermal bands; here, a false color composite of green, red, and near-infrared at 10-meter resolution is employed. TerraSAR-X images have a ground sampling distance of 6 meters in its commonly used stripmap mode, offering dual and cross-polarized images. To compose the

Table 2 Specifications of satellite sensors used in the experiments

	GSD	Source Size	Bands \ Polarization	Date	Incidence Angle	Image Tiles
QuickBird	0.5m	103km^2	blue, green, red, nir	17.11.2008	16.6	7487
Sentinel-2	10m	781km^2	blue, green, red, nir	19.11.2017	4.8	219
TerraSAR-X	6m	242km^2	HH\VV	29.09.2013	33.7	2113
	6m	242km^2	VV\VH	11.12.2013	33.7	
	6m	308km^2	HH\VV	10.10.2013	34.7	
	6m	308km^2	VV\VH	04.12.2013	34.7	

available polarizations into three channels, a principal component analysis (PCA) method is applied, as shown in Table 2.

The satellite images were divided into 224 × 224 pixel tiles with a 28-pixel overlap to increase data volume and address segmentation issues near edges. The images were classified into four semantic classes: urban, vegetation, water, and slums. Fully labeled images were created for training and evaluation using a multi-step process involving hierarchical, knowledge-based, and object-based classification, combined with machine learning and visual interpretation. Initial land cover classes formal built-up, water and vegetation were classified using a random forest classifier. Slum areas were identified through visual interpretation. Figure 2 shows the remote sensing data and the corresponding reference data used for the experiments.

FCNs, introduced by Long et al. (2015), enable end-to-end and pixel-to-pixel training for semantic segmentation tasks, predicting dense outputs from its input images. FCNs perform learning and inference on entire images using dense feedforward computations and backpropagation. A key feature of FCNs is the use of upsampling layers, which facilitate pixelwise predictions and enable learning from subsampled pooling layers. The backbone CNN for the FCN is based on the VGG19 classification architecture (Simonyan and Zisserman, 2015). VGG19 is characterized by small receptive fields of 3 × 3 pixels, where convolutions are applied at every pixel. To adapt VGG19 for an FCN , several modifications are made. The final classification layer is replaced with a 1 × 1 convolutional layer, matching the number of output channels to the number of target classes. Additionally, deconvolutional layers are introduced for bilinear upsampling, transforming coarse outputs into dense predictions. This process involves transpose convolutions, where the

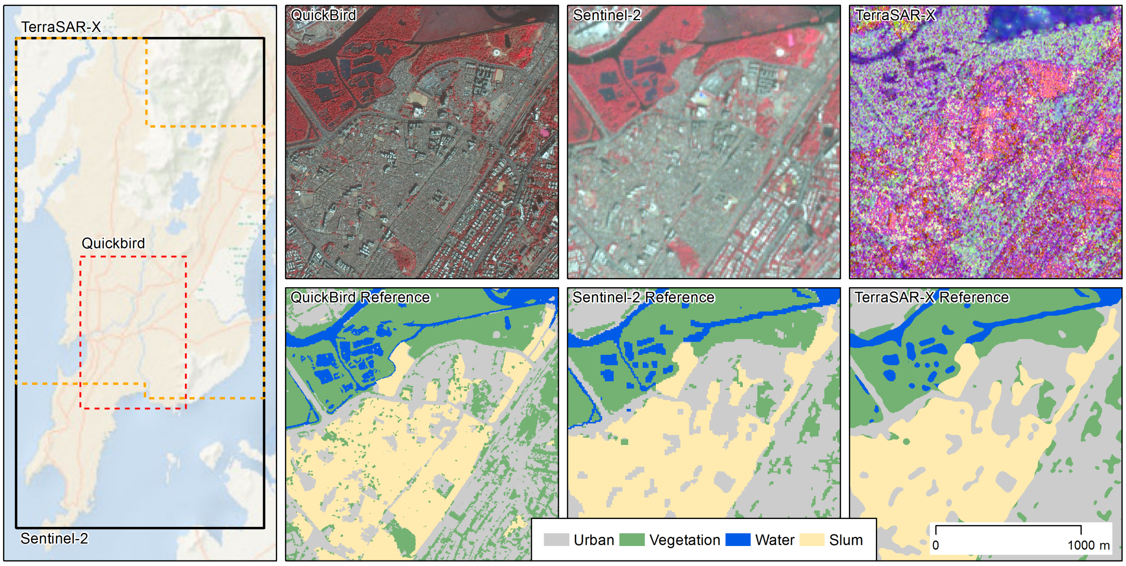

Figure 2 Composites and reference labels for all datasets: QuickBird and Sentinel-2 in false color and TerraSAR-X as PCA composite for a subset of central Mumbai.

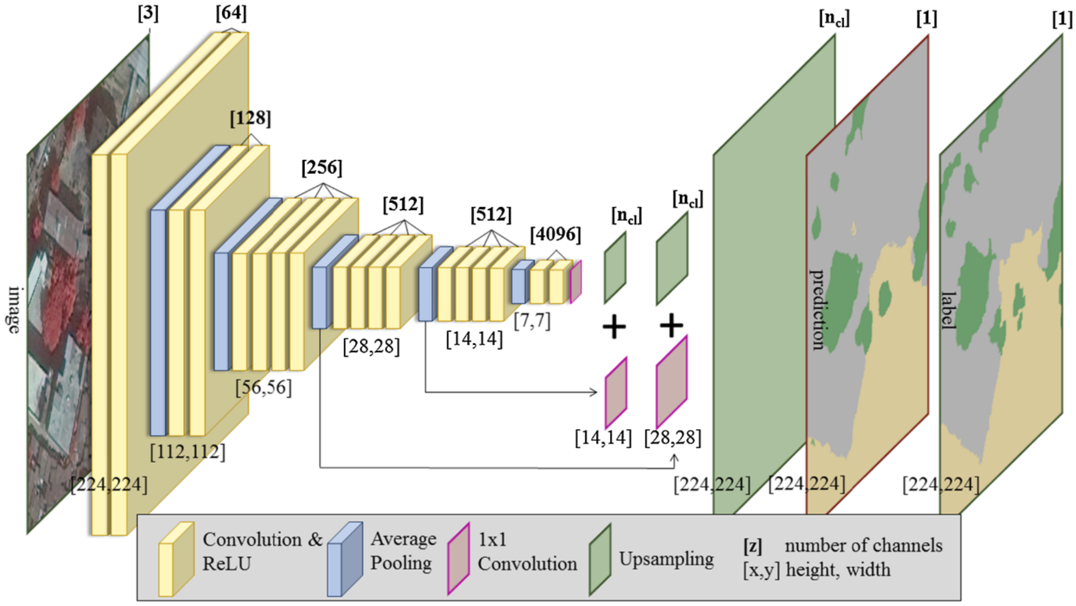

Figure 3 Architecture of the FCN VGG19 adapted from (Long et al., 2015). Prediction is performed using upsampling layers with four channels for the all classes $[n_{cl}]$ in the reference data. Upsampling layers are fused with 1 × 1 convolutions of the third and fourth pooling layers with the same channel dimension [x,y,n cl].

convolution's forward and backward passes are reversed.

The FCN seen in Figure 3 also implements skip connections, which combine predictions from lower-level layers with higher-level outputs. This fusion of fine and coarse layers helps the model generate local predictions while preserving global structural information. Specifically, the model fuses the upsampled output from VGG19 with predictions from the third and fourth pooling layers, enhancing its overall performance in semantic segmentation tasks.

As shown in Table 2, each data source covers a different area and produces a varying number of image tiles due to differences in geometric resolution. The Sentinel-2 dataset, despite covering the largest area with its 10-meter resolution, contains the smallest number of image tiles, with only 219. In contrast, the TerraSAR-X dataset contains 2,113 image tiles, and the QuickBird dataset, despite covering the smallest area, has the largest number of image tiles, totaling 7,487.

The FCNs were trained using the Adam optimizer, with a batch size of two image tiles, a fixed learning rate of 10^{-5}, and a dropout value of 15%. Two sets of experiments were conducted. The first experiment involved training the model for 100 epochs on each of the three remote sensing datasets (FCN-QB, FCN-S2, FCN-TX). The second experiment included transfer learning model trained on the QuickBird dataset two times. First the FCN-QB was transfer learned towards the to Sentinel-2 data (FCN-TL-S2) and secondly to the TerraSAR-X dataset (FCN-TL-TX).

Performance was evaluated using 4-fold cross-validation, where each scene was split into four strips, with three used for training and one for testing. This process was repeated four times to cover the entire scene. Accuracy was assessed using the kappa value, overall accuracy (OA), and Intersection over Union (IoU). These metrics provided insights into the general and class-specific performance of the segmentation results.

3.1.3. Evaluation of Results and Conclusions

Quantitative results in terms of overall performance for the semantic segmentation are presented in Table 3 for all five experiments. With regards to overall measures, all five experiments obtained considerable accuracies with Kappa values between 0.72 and 0.85. The best performing set-up is reported for the QuickBird model (FCN-QB). The Kappa value (0.85) and the Overall Accuracy (90.62%) show a very high agreement. This is followed by the transfer learned Sentinel-2 experiment (FCN-TL-S2) with the same Kappa value (0.85) and marginally lower OA (89.64%). Interestingly the highest IoU (87.43%) is reported for the Sentinel-2 model (FCN–TL-S2) which can be considered as being mostly related to the substantially larger area of interest for Sentinel-2 seen in Figure 2.

The baseline results presented in Table 3 show that the FCN trained on QuickBird data achieves an overall accuracy of 90.92%. In comparison, the accuracy drops to 86.71% when trained on Sentinel-2 data and further decreases to 80.68% for TerraSAR-X data. When applying transfer learning, where the FCN initially trained on QuickBird data is adapted to Sentinel-2 and TerraSAR-X data, notable changes in overall accuracy are observed. Specifically, the transfer from QuickBird to Sentinel-2 results in an overall accuracy of 89.64%, representing a slight decrease from the baseline QuickBird performance but an improvement over the Sentinel-2 baseline. For TerraSAR-X, transfer learning yields an overall accuracy of 80.03%, a marginal decline from the baseline accuracy for this dataset, indicating challenges in transferring knowledge from optical to radar data.

Table 3 Comparison of different approaches based on Kappa, Overall Accuracy (OA), and IoU.

Approach	Kappa	OA (%)	IoU (%)
FCN-QB	0.85	90.62	84.12
FCN-S2	0.81	86.71	83.94
FCN-TL-S2	0.85	89.64	87.43
FCN-TX	0.73	80.68	73.96
FCN-TL-TX	0.72	80.03	73.02

Table 4 Performance evaluation of the FCN for all classes. IoU: Intersection over Union; PPV: Positive Prediction Value; Sens: Sensitivity; A: area (percentage of scene coverage).

	Urban				Vegetation				Water				Slum			
	Sens (%)	PPV (%)	IoU (%)	A (%)	Sens (%)	PPV (%)	IoU (%)	A (%)	Sens (%)	PPV (%)	IoU (%)	A (%)	Sens (%)	PPV (%)	IoU (%)	A (%)
FCN-QB	91.37	90.34	83.24	47.57	92.90	95.35	88.88	36.96	90.78	90.97	83.28	4.98	85.70	88.39	77.02	10.48
FCN-S2	87.47	75.87	68.43	36.36	96.42	98.44	94.97	35.20	85.35	89.72	77.75	26.12	38.21	78.82	35.51	2.32
FCN-TL-S2	87.62	82.00	73.49	33.70	97.47	98.57	96.12	36.81	90.14	90.61	82.44	26.37	55.47	85.25	51.23	3.12
FCN-TX	84.29	83.13	71.99	36.14	93.86	94.03	88.59	39.84	78.46	75.65	62.63	19.54	51.64	72.50	46.27	4.47
FCN-TL-TX	85.78	80.21	70.80	34.93	93.49	93.58	87.85	42.02	75.82	75.64	60.94	19.56	43.64	78.43	38.42	3.49

While overall performance measures provide a general assessment, class-specific evaluations offer deeper insights into the segmentation results. Table 4 presents these class-specific performance measures. The highest accuracies for the urban and slum classes are achieved by QuickBird (FCN-QB). For the vegetation class, Sentinel-2 (FCN-TL-S2) performs best, likely due to its broader aggregation of information. Water class accuracies are similar between QuickBird (FCN-QB) and Sentinel-2 (FCN-TL-S2), with only minor differences.

QuickBird (FCN-QB) significantly outperforms Sentinel-2 (FCN-TL-S2) in the urban class. The most striking result is for the slum class, where the high geometric resolution of QuickBird allows for accurate segmentation of small-scale buildings. High accuracies in positive prediction value (88.4%) and sensitivity (85.7%) are achieved, with an IoU of 77%, indicating reliable detection of slum areas and minimal false positives.

Using transfer learning, Sentinel-2 shows a significant accuracy improvement in all classes, particularly in slums, highlighting the value of transfer learning. However, no positive effect is observed for TerraSAR-X data, where most classes are better represented by training directly of the source dataset (FCN-TX), except for an increase in PPV of slums.

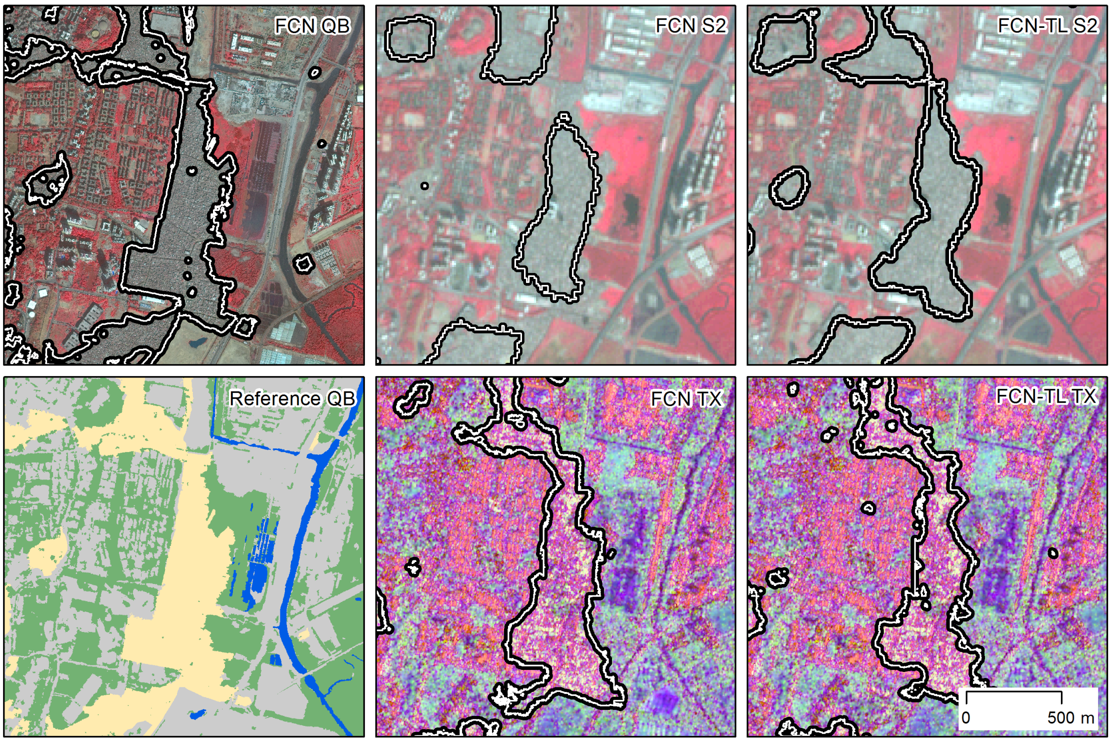

Figure 4 Comparative alignment of a slum patch showing differences in segmentation results. Slum patches in the reference map are depicted in yellow.

Figure 4 presents the visual outcomes for all conducted experiments, illustrating the performance of semantic segmentation across the entire study area. The visual analysis highlights the effectiveness of the FCN when applied to QuickBird data, where the segmentation results closely resemble the fine-grained structure observed in the reference dataset. For Sentinel-2 data, the impact of transfer learning is distinctly noticeable. While the original FCN-S2 results depict broader patches, the transfer learning approach (FCN-TL-S2) significantly enhances granularity, allowing for the detection of small vegetation patches and slum areas even at a 10-meter resolution. This improvement underscores the value of transfer learning in refining segmentation results, especially in complex urban landscapes. In contrast, the results for TerraSAR-X (FCN-TX) show minimal changes with the application of transfer learning.

Previous studies emphasize the considerable variability in slum patch sizes within cities (Friesen et al., 2018; Wurm et al., 2017). Most slums cover areas smaller than 5 hectares (ha), with only a few exceeding 25 ha. Based on these observations, a patch size-based accuracy assessment was conducted specifically for the slum class to evaluate the impact of slum patch size on classification performance.

A quantitative sensitivity assessment was performed in Table 5. For small slum patches ($<$ 5 ha), QuickBird imagery demonstrated excellent mapping capabilities (FCN-QB: 78.57%). Sentinel-2 showed a significant sensitivity increase between pretrained and transfer-learned models (9.32% vs. 24.67%). However, for TerraSAR-X, sensitivity decreased in the smallest patches (31.26% vs. 20.78%).

For medium-sized patches, a similar trend was observed: QuickBird achieved the highest sensitivity (FCN-QB: 83.63%), and transfer learning improved Sentinel-2 performance (28.19% vs. 50.64%). TerraSAR-X again showed a performance drop (47.36% vs. 37.98%).

Finally, for large slum patches, QuickBird detected 88.39% of reference slum pixels, while transfer learning significantly enhanced Sentinel-2 mapping (47.18% vs. 62.46%). A slight performance decline was noted for TerraSAR-X (48.36% vs. 55.34%). These findings underscore the strong influence of slum patch size on detection rates across all datasets.

Table 5 Sensitivity measurement as a function of varying slum patch size.

	Small	Medium	Large
	<5ha	5-25ha	>25ha
FCN-QB	78.57%	83.63%	88.39%
FCN-S2	09.32%	28.19%	47.18%
FCN-TF-S2	24.67%	50.64%	62.46%
FCN-TX	31.26%	47.36%	55.34%
FCN-TF-TX	20.78%	37.98%	48.36%

In summary, the study reveals that the FCN trained on QuickBird excels in classifying complex urban environments, demonstrating high accuracy in heterogeneous settings. Additionally, transfer learning significantly enhances the performance of Sentinel-2, particularly in detecting slum areas, where finer details are captured more effectively. However, the TerraSAR-X data shows lower overall performance compared to optical data like QuickBird and Sentinel-2. Moreover, transfer learning does not improve the overall performance of TerraSAR-X, although it does offer marginal benefits for slum classification. These findings highlight the strengths and limitations of different data sources and learning approaches in semantic segmentation tasks.

3.2. Large Scale Slum Mapping in 10 Cities Using Transfer Learning

3.2.1. Categorizing Slums and Their Morphological Variability

In Section 3.1 FCNs have been demonstrated as an effective method for slum mapping using a variety of remote sensing datasets within the city of Mumbai. However, large-scale slum mapping presents several challenges, including the fuzzy boundaries between formal and informal settlements, a significant imbalance between slum and non-slum areas, and the diverse morphological characteristics of slums from different geographical regions. These challenges complicate the scalability of slum mapping methods when applied to larger scale.

In this study these challenges are addressed by using semantic segmentation across 10 cities in the Global South: Cape Town (South Africa), Caracas (Venezuela), Delhi and Mumbai (India), Lagos (Nigeria), Medellin (Colombia), Nairobi (Kenya), Rio de Janeiro and São Paulo (Brazil), and Shenzhen (China). This research emphasizes the complexity of defining slums due to their highly variable morphological features, which differ significantly not only between cities (interurban variability) but also within the same city (intraurban variability). This variability highlights the need for adaptable and robust methods in large-scale slum mapping efforts.

The study addresses the challenges of categorizing slums due to their varied and inconsistent characteristics. As seen in Figure 5 slums differ significantly in physical appearance across and within cities, such as dense, low-rise shacks in Mumbai 5(a) and three-story buildings in Medellin 5(d), or even within Lagos, where some slums have floating shacks 5(b) while others have regular road networks 5(c). To tackle this, the study categorizes slums into three groups based on its morphologic type as presented in (Taubenböck et al., 2018). Category C1 includes extreme cases like those in Mumbai and Nairobi, marked by high density and irregular building orientation. Category C2 features slums with slight deviations, such as those in Delhi and Lagos. Category C3 covers slums in cities like Cape Town and Rio de Janeiro, where slum morphology can resemble formal settlements. The study uses transfer learning a custom FCN to improve slum mapping. Tested on multichannel HR remote sensing data from PlanetScope, this approach aims to enhance global slum mapping, especially in areas where traditional methods fall short. Incorporating auxiliary data, like road layouts from OpenStreetMap.

3.2.2. Transfer Learning Using Remote Sensing and Auxiliary Data

CNNs pretrained on natural images, e.g. ImageNet dataset, oftentimes limit input image depth to three channels, thereby neglecting the rich multispectral data available in remote sensing imagery. To fully exploit the spectral depth of optical satellite sensors, CNNs can be trained from scratch on multispectral data with any number of input channels. However, this approach is computationally expensive (Li and Liu, 2019; Senecal et al., 2019). Specific architectures can balance depth and efficiency, making the model lighter

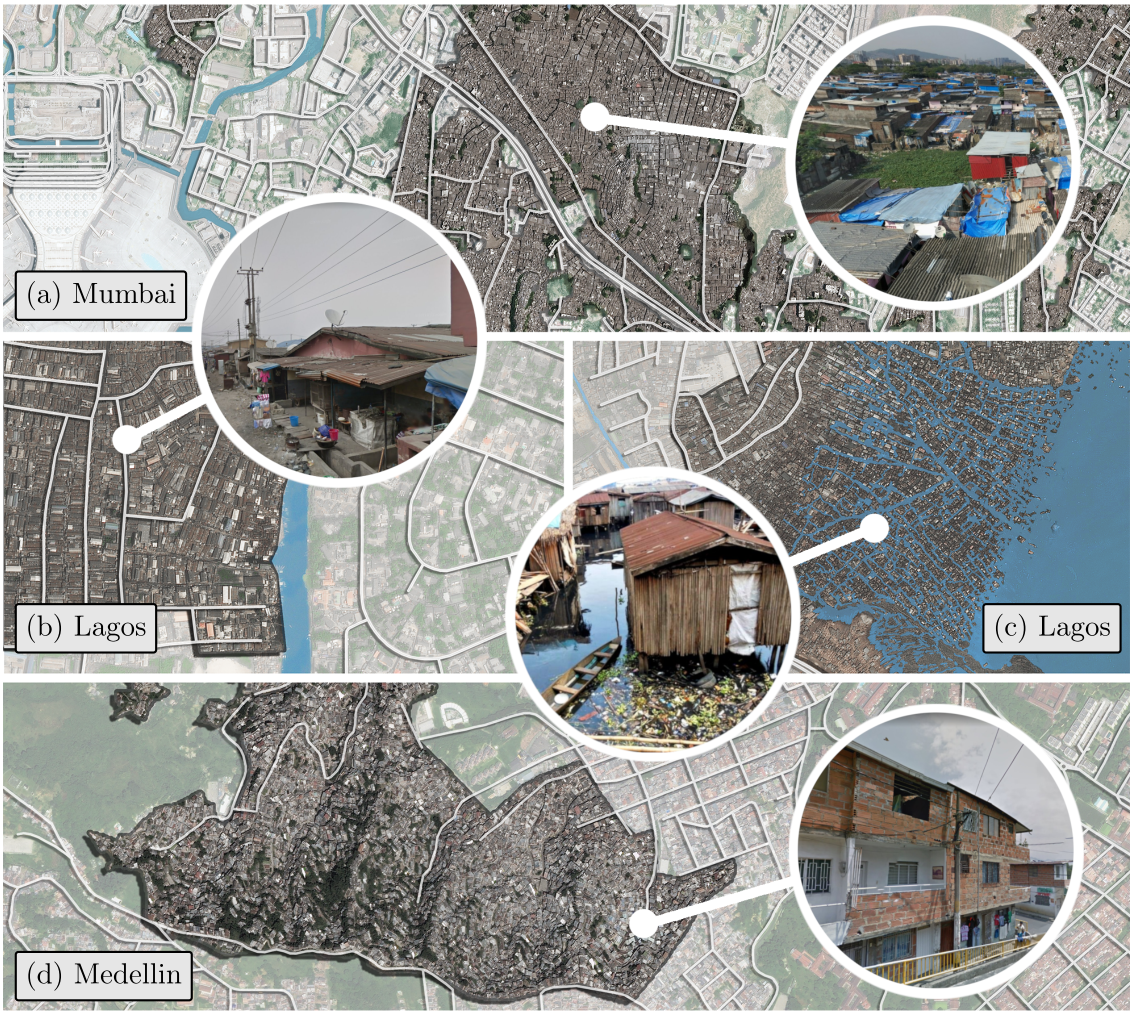

Figure 5 A comparison of the inter- and intra-urban variability of slums. Image (a) shows a typical slum in Mumbai, India, consisting of very densely built shacks. The images (b, c) in the middle show two very different slums in Lagos, Nigeria: poverty areas in the city's periphery as well as the downtown floating slum of Makoko in the Lagoon of Lagos. Last, image (d) depicts a slum in Medellin, Colombia, with three-story buildings made of concrete. Images from Google Street View provide additional close-up information on the local built-up structure.

and easier to train. The Xception network, an evolution of the Inception models, exemplifies this balance (Chollet, 2017; Szegedy et al., 2015). It employs depthwise separable convolutions to reduce the model's parameter size while maintaining performance. Thus a modified Xception network is used as the backbone architecture for a FCN called a fully convolutional xception network XFCN.

The Xception network's architecture is composed of modules designed to decouple cross-channel and spatial correlations, thus shrinking the model's parameter size (Chollet, 2017). These modules, integral to the Xception network, perform depthwise 3×3 convolutions followed by pointwise 1×1 convolutions. This design reduces the number of connections required, making the model lighter and more efficient. The architecture consists of an entry flow, a middle flow, and an exit flow, with residual connections throughout to maintain information flow. The final XFCN model includes 41 convolutional layers, combining batch normalization, ReLU activation fucntions, and dropout layers.

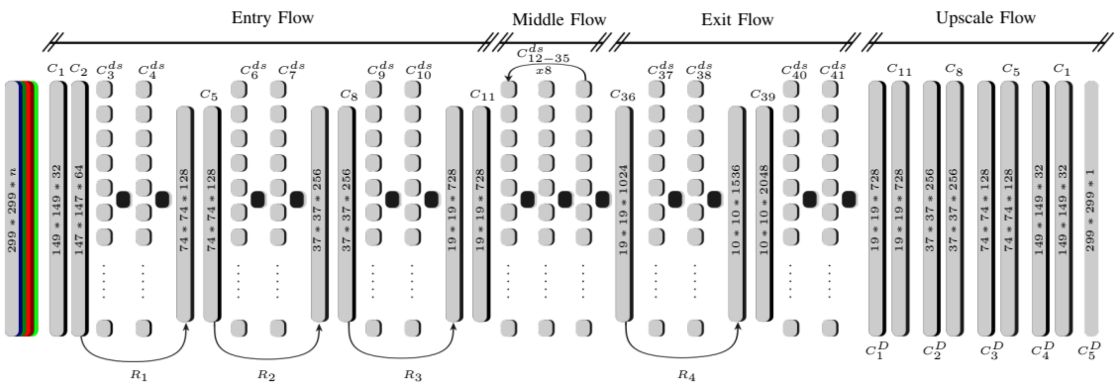

Figure 6 The architecture of the XFCN. The Xception backbone is slightly changed to allow for a multi-dimensional input and more rigorous regularization. After the exit flow a fully convolutional flow follows. All convolutional blocks are a combination of standard 2D convolutions C_n or depthwise separable convolutions C_n^{ds} in combination with batch normalization, dropout, and ReLU activation functions. The XFCN features residual skip connection throughout the whole network (R_n), and during the upscale flow the dilated convolutions (C_n^D) are fused with the long distance skip connections from the entry flow.

For the decoder, the XFCN uses an upsampling approach similar to the original FCN (Long et al., 2015), with five dilated convolutions to upscale the output back to the original input dimensions. A softmax classifier then produces a single prediction tensor. The decoder also incorporates long-distance skip connections to preserve low-level features, ensuring fine-grained upsampling performance.

The XFCN was specifically designed to map slums using HR PlanetScope data. PlanetScope imagery, with a geometric resolution resampled to 3m and four spectral bands (blue, green, red, and near-infrared), is an ideal data source for large-scale poverty mapping. To enhance this data, the normalized difference vegetation index (NDVI) was added as a fifth feature, improving the ability to detect subtle landscape variations. All datasets used in this study were 16-bit surface reflectance data, normalized to a range of 0–1 to be compatible with the deep learning framework. Additionally, the OpenStreetMap (OSM) road network was incorporated as an auxiliary layer. Only paved roads accessible by automobiles were selected to ensure consistency across cities. A binary logarithmic proximity to each road was computed, providing information on road distance and settlement structures. This auxiliary road network data complements the spectral data by offering further insights into urban morphology and slum patterns, strengthening the XFCN's capacity for accurate slum detection (Ibrahim et al., 2019; Kuffer et al., 2017).

The reference data for the 10 cities were composed of manually mapped polygons for each PlanetScope scene in order to maintain consistent delineation of slum boundaries across the cities and to ensured coherent transfer learning between datasets. Bing aerial imagery and Google Street View were used to construct the reference datasets. Only slums larger than 1 hectares were included. Table 6 provides an overview of the datasets utilized for training the XFCN model. The accompanying table presents detailed information on the dataset from each city, including the total area covered by the satellite imagery, the

	Caracas	Mumbai	Nairobi	Delhi	Lagos	Medellin	Shenzhen	Cape Town	Rio	São Paulo
	CA	MU	NA	DE	LA	ME	SH	CT	RI	SP
Category	C_1	C_1	C_1	C_2	C_2	C_2	C_2	C_3	C_3	C_3
Area [km^2]	357	1379	211	852	230	59	1471	356	2086	3764
Number of slums	104	452	47	232	51	49	1872	70	404	905
Slum area [km^2]	30.4	41.3	8.2	5.3	15.0	4.1	46.2	6.1	26.4	51.3
Mean size[ha]	29.2	9.1	17.5	2.3	29.4	8.4	2.5	8.3	6.5	4.4
Training data										
Training samples	10,902	19,109	2,300	2,162	3,090	1,565	23,909	2,117	10,722	18,822
Slum proportion [%]	38.2	26.4	22.7	7.3	46.1	24.1	22.7	19.8	19.4	16.9

Table 6 Overview on the datasets used for training the XFCN. The table shows information on each city's dataset with the total area of the satellite scene, the number of slums within the city, the total area of slums and the mean size of slums for each site. The number of image patches used for training the XFCN and the slum sample percentage gives more insight about the available datasets.

number of slum regions identified within the city, the total area these slums occupy, and the average size of the slums at each site. Additionally, the number of image patches employed in training the XFCN and the percentage of slum samples within these datasets offer further insights into the data availability with a total number of 94,698 samples including 22,798 slum samples.

In order to prevent overfitting, batch normalization, dropout layers, and early stopping are used. The models are trained using a softmax cross-entropy loss function and the Adam optimizer, with an exponential decaying learning rate. Unseen image patches from the test dataset, with dimensions of 299 × 299 pixels, are predicted with an overlap of 199 pixels in both the x and y directions. This method allowed for multiple predictions over the same area, ensuring robust handling of uncertainties and minimizing edge prediction issues. The results are evaluated using the F1-score and IoU.

Three sets of experiments are conducted. In the first experiment named $XFCN^{city}$, individual datasets for each city are used for both training and testing, with geographical separation between the training and testing areas within the same city. This experiment establishes a baseline for performance in scenarios with limited data availability. The second experiment involves a leave-one-out cross-validation approach, where datasets from all cities except one are combined for training. The experiment is named $XFCN_{LSP}$ for its Large Scale Poverty dataset. The trained model is then tested on the remaining city, with the test dataset for each city being identical to the first experiment. This setup assesses the model's ability to generalize across different geographic regions. In the final experiment $XFCN_{LSP}^{TF}$, the leave-one-out cross-validation method is extended with a transfer learning approach. Here, models trained on the leave-one-out cross-validation dataset are further transfer-learned using the same city-specific datasets from the first experiment. This experiment evaluates the effectiveness of transfer learning in improving model performance when applied to geographically distinct areas.

Each of these experiments is conducted twice: first using a five-dimensional dataset (B, G, R, NIR, NDVI), which consists solely of remote sensing data from PlanetScope. The second set of experiments utilizes a six-dimensional dataset (B, G, R, NIR, NDVI, OSM), which incorporates auxiliary road network data from OpenStreetMap (OSM). This additional road network data complements the spectral information by providing insights into urban morphology and slum patterns.

3.2.3. Evaluation of Results and the Implications of Slum Categories

Results for all experiments using the IoU and the F1-score accuracy measures for each city can be seen in Table 7. In the first experiment $XFCN^{city}$, models were trained on individual city datasets and tested on spatially separated areas within the same city. For the five-dimensional input data (B, G, R, NIR, NDVI), the mean IoU across all cities was 62.93%, with a mean F1-score of 66.86%. When the road network was included as an additional input layer, the six-dimensional data achieved a higher mean IoU of 67.98% and a mean F1-score of 73.35%. Notably, the inclusion of the road network improved the IoU by 5.05% and the F1-score by 6.49%. The best results were observed in Mumbai, Nairobi, and São Paulo.

The $XFCN_{LSP}$ experiment trained models on a large scale poverty (LSP) dataset, where nine cities are combined into one training dataset and tested the model on the remaining city. For the five-dimensional input data, the mean IoU was 57.81% and the mean F1-score was 63.87%. With the six-dimensional input, these scores increased significantly to 71.64% and 75.30%, respectively. The inclusion of the road network was particularly beneficial, improving the mean IoU by 13.82% and the F1-score by 11.41%. Cities like Caracas, Lagos, and Shenzhen exhibited the highest IoU accuracies with the six-dimensional input data.

In the $XFCN_{LSP}^{TF}$ experiment, models were first trained on the LSP dataset and then transfer learned on individual city datasets. The mean IoU for the five-dimensional data was 72.60%, with a mean F1-score of 77.69%. For the six-dimensional data, these scores increased slightly to 74.53% and 78.10%, respectively. The highest accuracies were observed in Mumbai and Shenzhen, where IoU scores exceeded 80%.

Dataset		CA	MU	NA	DE	LA	ME	SH	CT	RI	SP	Mean
Slum category		C_1	C_1	C_1	C_2	C_2	C_2	C_2	C_3	C_3	C_3	
	n_{dim}						IoU					
$XFCN_{city}$		59.13	71.16	**73.13**	56.23	61.04	51.47	63.24	66.56	58.89	68.42	62.93
$XFCN_{LSP}$	5	58.16	50.58	49.05	57.97	**67.48**	61.27	63.01	57.48	56.74	56.43	57.82
$XFCN_{LSP}^{TF}$		80.70	**80.86**	75.63	58.10	70.44	68.35	70.11	77.98	73.37	70.49	72.60
$XFCN_{city}$		76.73	78.49	78.09	60.22	71.91	48.95	**85.98**	72.28	54.48	61.19	68.83
$XFCN_{LSP}$	6	78.14	66.32	64.54	67.18	80.77	70.51	**80.99**	78.63	65.25	64.08	71.64
$XFCN_{LSP}^{TF}$		81.62	81.80	79.73	64.65	74.72	69.83	**86.29**	81.54	60.95	64.20	74.53
							F1-score (F1)					
$XFCN_{city}$		63.66	76.29	56.63	61.22	51.76	**77.23**	71.19	71.66	64.12	74.84	66.86
$XFCN_{LSP}$	5	63.87	54.81	56.15	63.66	70.78	66.30	**77.31**	59.12	63.50	63.37	63.89
$XFCN_{LSP}^{TF}$		85.93	**86.98**	79.76	59.20	74.56	75.73	73.62	83.89	80.03	77.25	77.70
$XFCN_{city}$		81.48	83.98	82.67	68.47	71.99	60.15	**89.49**	73.81	57.27	64.16	73.35
$XFCN_{LSP}$	6	82.68	70.33	66.94	71.81	76.78	76.58	**83.59**	82.61	71.52	70.20	75.30
$XFCN_{LSP}^{TF}$		86.17	86.63	83.24	67.44	79.52	72.72	**89.48**	83.76	64.46	67.53	78.10

Table 7 Results for all experiments using the IoU and the F1-score accuracy measures. The top part of the table shows the experiments for the five-dimensional remote sensing data, while the bottom part includes the proximity to the road network as an additional sixth input dimension. The highest accuracies for each experiment are presented in bold; the highest overall accuracy for each accuracy score is highlighted in gray.

Intra- and inter-urban variability present significant challenges for deep learning models when mapping slum morphologies. These morphologies can vary not only between cities (inter-urban variability) but also within a single city (intra-urban variability), as exemplified by the case of Lagos (Figure 5). The segmentation results of the XFCN models in Figure 7 highlight the complexity of variability across three different cities using the three proposed approaches.

In slum settlements where most or all typical slum morphologies are present, the models tend to perform with high accuracy. This slum category C1 as seen in Caracas, Mumbai, and Nairobi, consistently achieves high accuracies across all experiments with both the five-dimensional and six-dimensional datasets. In the city of Mumbai the results are presented in Figure 7 and demonstrate robust performance for all experiments in the segmentation tasks. The slum areas exhibit the characteristic morphological features typical of informal settlements, including irregular building patterns and dense spatial layouts, which are distinctly different from the organized structure of formal settlements.

For Shenzhen, seen in Figure 7 and the other cities within the second category C2, Delhi, Lagos, and Medellin, the accuracies achieved using both the five-dimensional and six-dimensional datasets are somewhat lower compared to the cities of the first slum category C1. The urban villages in Shenzhen can still be effectively classified into slum category

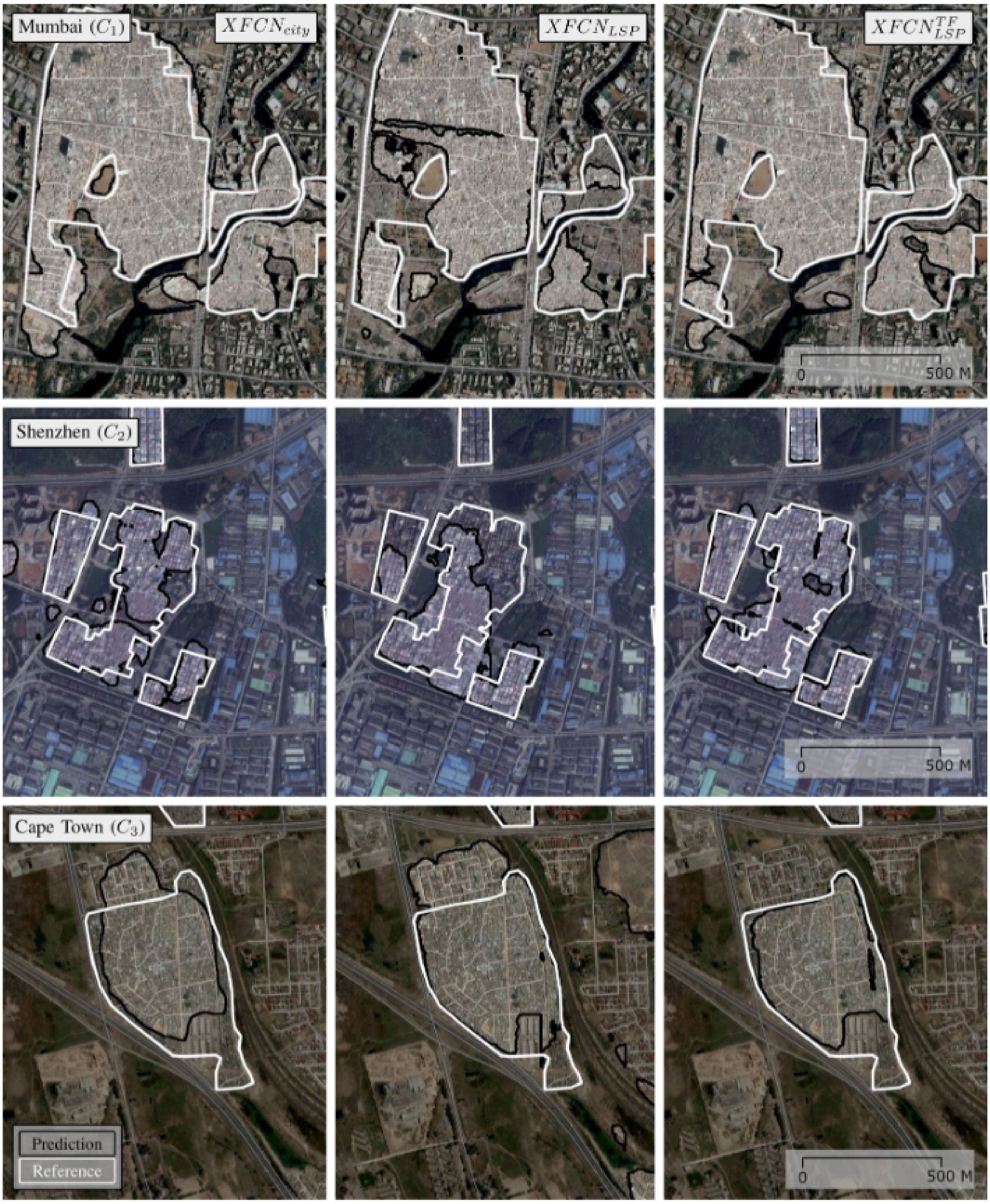

Figure 7 Comparative alignment for three cities of each slum category (C1-3). All results were trained on the six dimensional input data. The left column shows the results for the $XFCN_{city}$, the middle column shows results from the $XFCN_{LSP}$ model, and finally, the right column shows the transfer learned $XFCN_{LSP}^{TF}$ results.

C2. As shown in Figure 7, the segmentation results remain reliable despite the challenges posed by these areas, which feature a mix of dense, irregular layouts and multi-story or even high-rise buildings. Although not fully representative of typical slum morphologies, the models successfully capture the unique spatial characteristics of Shenzhen's urban villages, enabling accurate mapping and classification.

In the case of Cape Town, the accuracies are the lowest among the experiments, as the townships fall into slum category C3 with their untypical slum morphologies. These areas often have a regular road layout and lower building density, making them less characteristic of conventional slums. Nevertheless, the segmentation results for Cape Town, Rio de Janeiro, and Sao Paulo, still demonstrate the model's capability to successfully map slum areas, even in these more structured environments, as shown in Figure 7. This underscores the robustness of the methodology, which adapts well to diverse urban forms.

Despite this, it remains critical to focus on these atypical slum settlements, as they are often understudied and underrepresented in research, and frequently overlooked by NGOs and governments (Kuffer et al., 2024, 2016; United Nations, 2024). Overall, the inclusion of the road network as an additional input layer consistently improved the model's performance, particularly in complex urban environments. The transfer learning approach was especially effective, demonstrating that models trained on a diverse set of slum morphologies could be successfully adapted to local conditions, yielding significant improvements in accuracy. This approach was most beneficial in cities with challenging datasets, where training from scratch was insufficient.

3.3. Slum Detection Using Uncertainty Quantification With Transfer Learning on Limited Data

3.3.1. Addressing Uncertainty in Slum Detection with Fuzzy Boundaries and Limited Data

In section 3.2, it was observed that while semantic segmentation was highly successful in detecting slums in HR imagery, one of the challenges encountered was the presence of multiple categories of slums. For instance, segmenting slums with distinct and typical slum morphologies proved to be relatively straightforward. However, it was significantly more difficult to delineate slum boundaries when the slum settlements lacked typical morphological features. This challenge was particularly evident in cases where slum morphologies gradually transitioned into formal settlements, resulting in an indistinct separation between slum and formal areas. Due to these complexities, a new approach was adopted for this study. In this study, a scene classification approach is presented, combined with an approximation of uncertainty in slum predictions, to better capture and understand the gradual transitions between slum settlements and formal settlements.

The challenge illustrated in Figure 8 draws upon the findings of Zhu et al. (2019), where predictions for LCZ class 7, characterized by dense, low-rise buildings, highlights two regions within Nairobi, Kenya. Although both areas exhibit similar structural traits, closer inspection using Google Street View imagery reveals that only a portion of one area can be identified as a slum. This underscores the limitation of classifying slums based solely on morphological features like density and building height. Dense, low-rise structures do not automatically indicate slums, nor does the absence of such density negate the possibility of slum classification. A more nuanced assessment involving multiple morphological characteristics is essential. Given the noisy nature of the dataset and the challenges in acquiring accurate ground truth data, this study focuses on assessing the typical morphological features associated with slum settlements to better understand the confidence levels of slum predictions.

The study addresses two primary challenges: Limited data availability and noisy datasets in the context of slum mapping using HR PlanetScope data. The primary objective is to develop an efficient method for detecting slums with a limited number of training samples and to estimate the uncertainty within these predictions. To achieve this, a transfer learning approach is employed, leveraging a large, imbalanced dataset to effectively transfer-learn towards a smaller, balanced dataset. This approach ensured that only a few samples were required for successful slum detection. To address the issue of noisy datasets, Monte Carlo dropout is utilized, enabling the approximation of uncertainty associated with slum settlement predictions, thereby providing a more robust and reliable analysis. Additionally, a custom CNN, named the slum transfer network (STnet), is introduced. STnet, as seen in Figure 9, is specifically designed for HR remote sensing data and engineered to enhance training efficiency with a limited number of samples, while also significantly

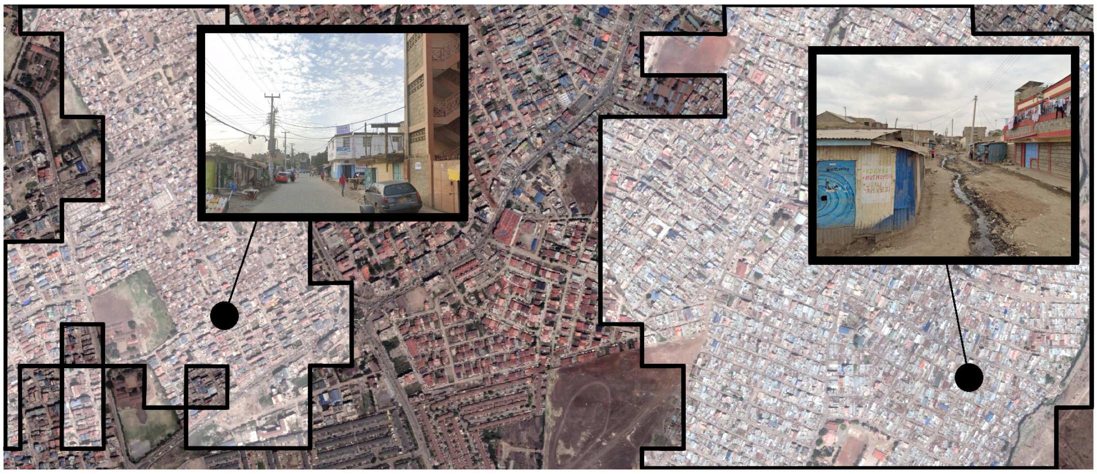

Figure 8 Dense and low-rise areas shown with a black outline for the city of Nairobi (Zhu et al., 2019). Google Street View imagery is used to show that only some parts of the dense areas can also be considered a slum settlement highlighting the challenge of slum mapping.

improving processing time compared to standard CNN models. The research aims to demonstrate the effectiveness of STnet in accurately detecting slums in diverse urban environments, contributing to advancements in both urban studies and remote sensing.

3.3.2. Transfer Learning STnet Using Efficient Learning Strategies

The custom STnet is optimized for processing HR remote sensing imagery. STnet is based on a modified Xception network, with a simplified structure seen in Figure 9. Its entry flow includes five convolutional layers with residual skip connections, using large 9x9 kernels in the first two layers to capture more extensive areas. Feature pyramid pooling is applied in the middle flow for multi-scale feature extraction. The classification flow consists of two linear functions, while batch normalization and dropout are used throughout. STnet contains 22 layers and 3.3 million trainable parameters.

The learning strategy in this study is divided into two phases using two datasets, which are created using a leave one out cross validation approach. Initially, the STnet model is pretrained on a class-imbalanced dataset, denoted as D_{base}. To address the class imbalance, a weighted loss function is employed, giving greater importance to under represented classes. This ensures that the model learns effectively from all available data, despite the imbalance. After pretraining, the STnet undergoes transfer learning using another dataset, D_{loocv}, which is class-balanced through under sampling. This balanced dataset ensures that each class is equally represented, mitigating any bias introduced during the pretraining phase on the imbalanced D_{base}.

To estimate uncertainty in model predictions, Monte Carlo dropout was employed. This technique involves averaging multiple predictions, each made with a different dropout configuration, to model the distribution of the predictive posterior (Gal and Ghahramani, 2016; Seoh, 2020). By using a dropout probability of 0.3, the model achieves a balance between maintaining accuracy and capturing uncertainty. The final prediction is obtained

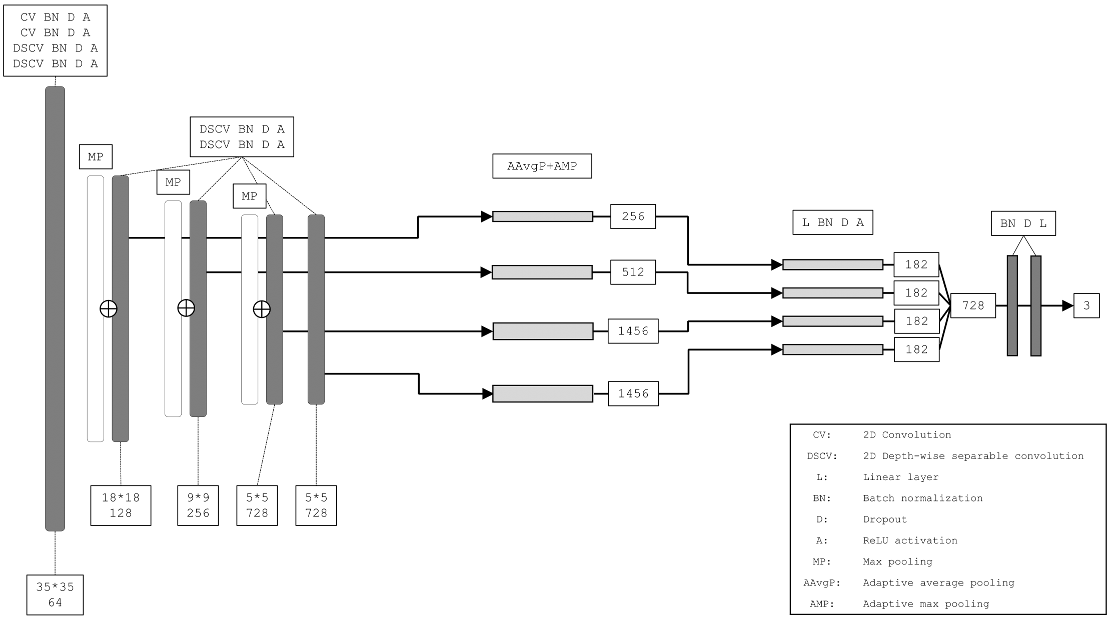

Figure 9 Simplified schematic of the STnet architecture, comprising five convolutional variants in the entry-flow, succeeded by feature pyramid pooling layers and a classification-flow in the end. This light-weight architecture encompasses 3.3 million trainable parameters.

by averaging the results of multiple forward passes, providing a more robust and reliable output.

The remote sensing data used in this study is collected from PlanetScope satellites in 2021, covering eight cities in the Global South; Cape Twon (South Africa), Caracas (Venezuela), Lagos (Nigeria), Medellin (Colombia), Mumbai (India), Nairobi (Kenia), and Rio de Janeiro and Sao Paulo (Brazil). The data is divided into 88×88 pixel patches, which are used to train and test the deep learning models. The dataset includes three classes: background, formal built-up areas, and slums. The class distribution of each city can be seen in Figure 10. In total 64,746 samples are available, of which 6.5% are slum samples. Slum areas were manually mapped by experts using up-to-date aerial imagery to create accurate reference data (Stark et al., 2020). A patch was labeled as a slum if it contained at least 25% slum pixels; otherwise, it was discarded or classified based on the highest pixel count.

The experimental setup included the use of an Adam optimizer, weighted soft cross-entropy loss, and Monte Carlo dropout for uncertainty estimation. The models were evaluated primarily on their ability to identify slum areas, using metrics such as F1-score, precision, and recall. Additionally, the study assessed the impact of 25 Monte Carlo iterations on model stability, training-, and inference time, ensuring a comprehensive evaluation of both performance and computational efficiency. This rigorous approach provides a clear understanding of the model's general performance across varying datasets, helping to optimize the overall quality of the results.

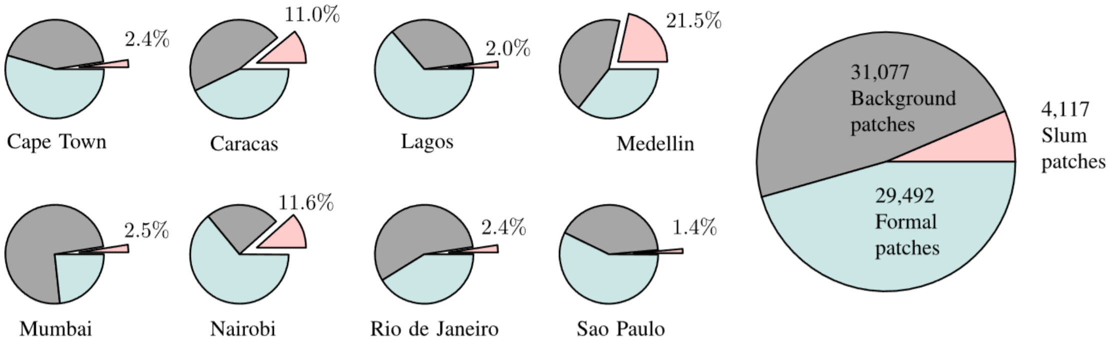

Figure 10 Class distribution in eight cities and combined distribution. The figure displays nine pie charts depicting the class distribution in eight cities, with the slum sample proportion highlighted for each city. The final pie chart showcases the combined distribution, illustrating the overall class proportions across all cities.

3.3.3. Evaluation of Results and Conclusion of Using Limited Data

The transfer learning results for the STnet model demonstrate a clear relationship between the number of samples per class used during training and the resulting F1-scores. As seen in Table 8, increasing the number of samples generally leads to improved F1-scores, although there is a notable plateau effect after 50 samples, where F1-scores stabilize around the high 80% range. Interestingly, even with just one sample per class, the model achieves a respectable F1-score of 73.24%, underscoring STnet's potential to perform well with limited training data. The highest F1-score, 86.24%, is observed when using 100 samples per class, highlighting the benefits of using more data in the transfer learning process.

In a comparative analysis of three different CNN architectures, STnet, Xception, and ResNet-50, STnet exhibits competitive performance despite its lower parameter count of 3.3 million. This efficiency is particularly notable in scenarios with limited samples for transfer learning, where STnet outperforms the more complex Xception and ResNet-50 models. As shown in Table 8, F1-scores for all models were averaged over five independently seeded runs, with standard deviations provided to illustrate performance variability. Although all three CNNs deliver similar overall F1-scores, STnet's streamlined architecture results in significantly faster training times, as detailed in Table 9. However, STnet requires more epochs to achieve optimal performance compared to the other models, suggesting a trade-off between speed and the number of training iterations.

The influence of varying Monte Carlo dropout rates on STnet's performance was also examined, focusing on inference time, F1-scores, and entropy values (a measure of prediction

Table 8 Comparison of F1-scores for STnet, Xception, and ResNet50, averaged over five differently seeded runs shown with their standard deviations. All models employed 25 Monte Carlo iterations.

	Inference	1	5	10	25	50	100
STnet	**0.7201** ± .11	**0.7324** ± .09	**0.7806** ± .05	**0.8082** ± .05	0.8358 ± .04	0.8432 ± .05	0.8624 ± .05
Xception	0.6724 ± .16	0.7309 ± .08	0.7804 ± .05	0.8031 ± .06	0.8380 ± .06	0.8502 ± .04	0.8775 ± .05
ResNet50	0.7150 ± .09	0.7010 ± .12	0.7811 ± .05	0.7941 ± .04	**0.8553** ± .04	**0.8709** ± .04	**0.8957** ± .02

Table 9 Comparing different CNN architectures based on their size and training time.

	Parameters	Training time	
		per step	total time
STnet	3.3m	33.25it/sec.	56:34 36 epochs
Xception	20.8m	21.28it/sec.	1:22:45 24 epochs
ResNet50	23.5m	17.85it/sec.	1:15:37 20 epochs

Table 10 Comparison of STnet's inference time, F1-score, and Entropy across different Monte Carlo Dropout iterations, trained on 100 samples per class.

Monte Carlo Iterations	Inference time	F1-score	Entropy
1	31.59 it/sec. 2:11 min	0.8423	–
5	20.43 it/sec. 3:20 min	0.8515	0.7845
25	5.68 it/sec. 12:02 min	0.8624	0.7832
50	3.10 it/sec. 22:17 min	0.8679	0.7810

uncertainty) as presented in Table 10. As the number of Monte Carlo dropout iterations increased from 5 to 50, slight improvements in F1-scores and reductions in entropy values were observed, indicating more confident predictions. However, this came at the cost of significantly longer inference times, with a 275% increase from 5 to 25 iterations, and an additional 84% increase from 25 to 50 iterations. Despite these trade-offs, 25 iterations were selected as the optimal configuration, balancing accuracy and computational efficiency.

Monte Carlo dropout provided valuable insights into the uncertainty associated with slum classification as seen in Figure 11. The STnet model showed high confidence in predicting typical slum settlements, particularly in cities with well-defined slum morphologies such as Caracas, Medellin, and Mumbai. However, challenges arose in cities like Lagos, where underclassification was prevalent, and in Nairobi, Rio de Janeiro, and Sao Paulo, where overclassification occurred. These difficulties stem from the lack of distinct slum features and the similarity of some formal settlements to slums in terms of density and building patterns. The findings underscore the complexity of slum classification and highlight the importance of considering local morphologic characteristics and the surrounding built-up areas when applying the model to different urban contexts.

In Figure 12, results from the STnet model over the same area of interest seen in Figure 8 are shown. The slum reference polygons are outlined in black, and slum probability results using different amount of training samples are displayed using the same red colorbar. These results highlight the model's performance in identifying slum settlements. All images (12a–12f) are displayed at a scale of 1:10,000. Figure 12a shows VHR Google satellite imagery of the point of interest, while Figures 12b–12f demonstrate the outcomes of applying the STnet with different number of available samples.

Figure 12b shows results without transfer learning, while Figures 12c–12f depict the impact of transfer learning with 1 to 50 samples, showing the improvement in performance. Interestingly, increasing the sample count from 50 to 100 produces negligible changes in accuracy and visual outcomes, so the figure concludes at 50 samples. Transfer learning

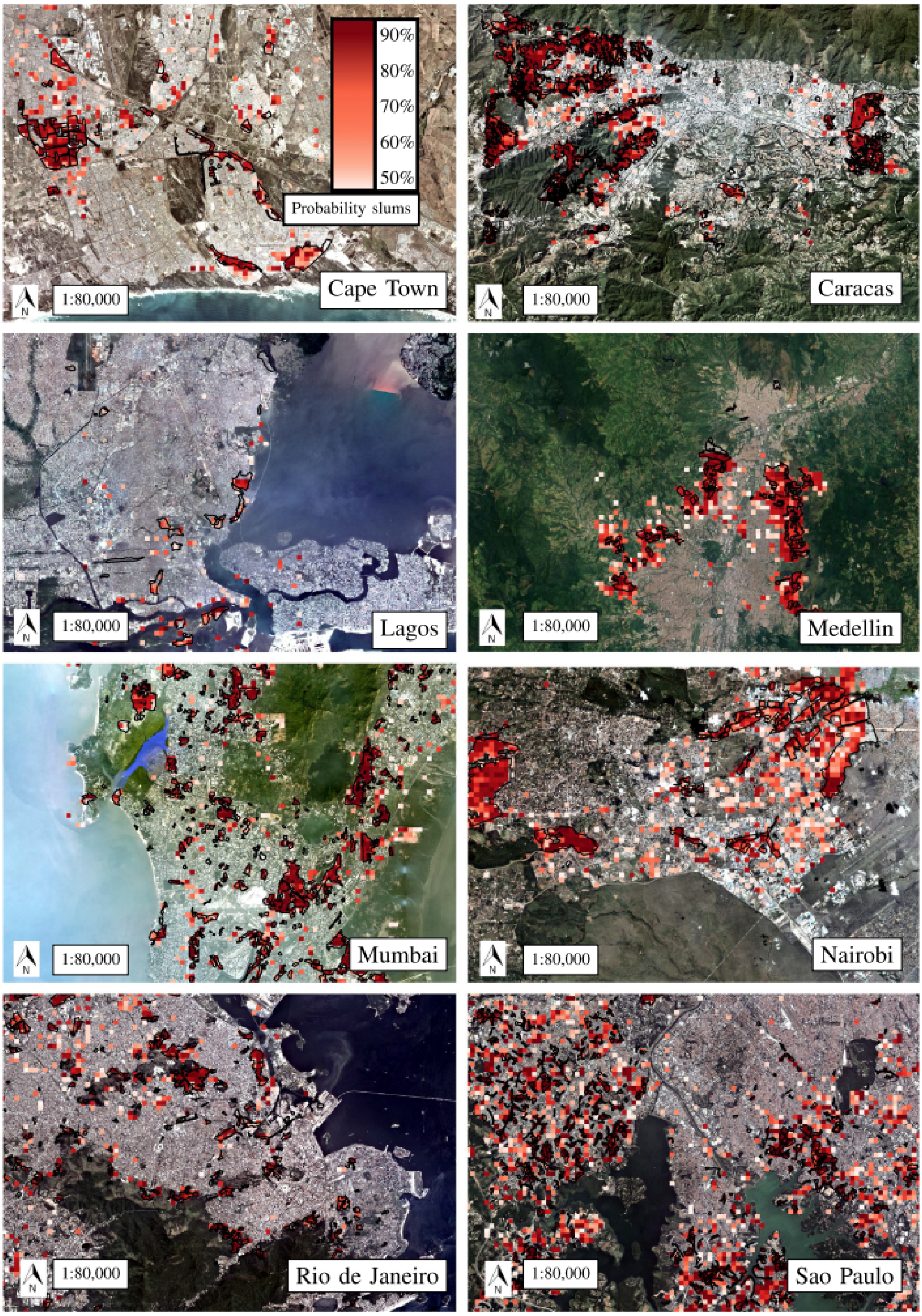

Figure 11 Results for all eight cities using the transfer-learned STnet trained on 100 samples per class. All results are in the same scale of 1:80,000 and use the same color-bar for the probability value of the slum class. Black outlines are used for the reference slum polygons.

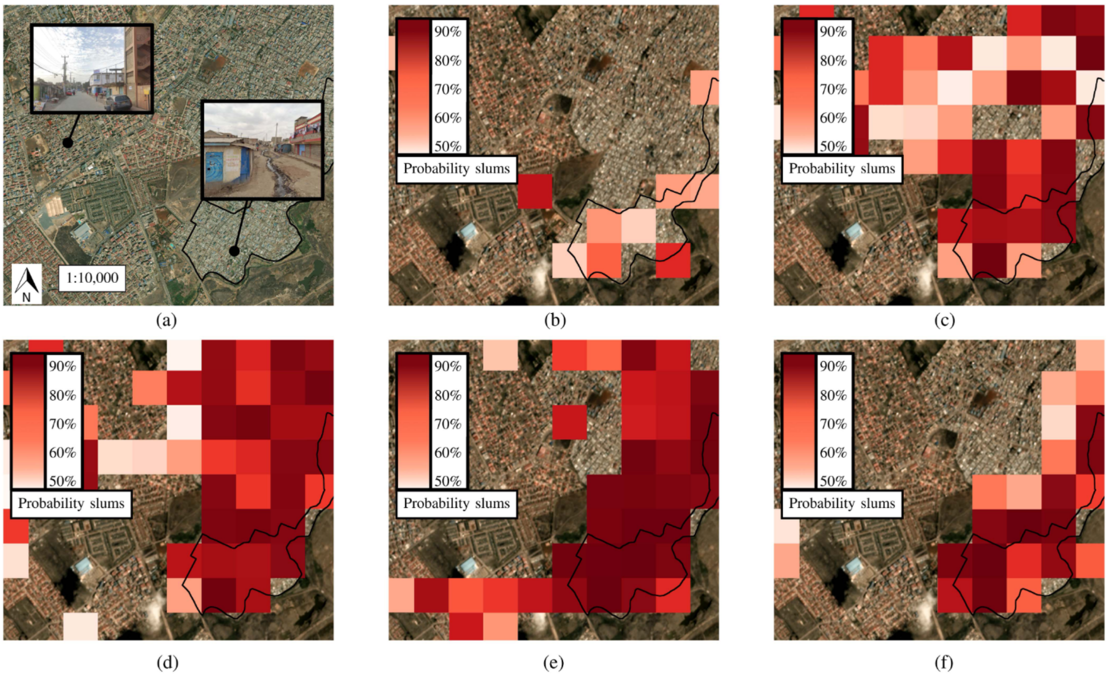

Figure 12 Results for the STnet in a comparable area of interest, as depicted in Figure 8. All images (a-f) are presented in a consistent scale of 1:10,000. Image (a) showcases a VHR Google satellite imagery of the identical point of interest shown in Figure 8. Images (b-f) exhibit the outcomes obtained using the STnet , with variations from no transfer learning (b) to transfer learning from 1 to 50 samples per class (c-f).

significantly improves the model's ability to recognize Nairobi's urban features.

Figure 11 provides a detailed overview of the STnet 's performance across Nairobi. Without transfer learning, the model achieved an F1-score of 49.06%, struggling to map slums accurately. While Figure 12b initially shows promising results, the predictions come with low confidence values. However, with the introduction of transfer learning using 1 sample per class, the F1-score rises to 66.78%, showing the effectiveness of even a small number of labeled samples in improving the model's understanding of Nairobi's urban morphology. As shown in Figures 12c to 12e, the F1-score improves progressively, but over-classification and overconfidence become issues. By Figure 6f, when enough samples are used, both the F1-score and visual outcomes improve significantly, with only minor instances of over- and under-classification.

These results highlight the capability of the proposed transfer learning approach to distinguish between areas with mixed formal and informal settlements. The presence of slums that either gradually transition into formal settlements or exhibit atypical slum morphologies presents challenges for traditional classification methods. This underscores the robustness of the transfer learning approach in handling such complexities. The findings have broader implications for cities like Lagos, Rio de Janeiro, and São Paulo, where similar challenges in distinguishing between formal and informal dense settlements arise. This highlights the generalizability of the transfer learning approach to other urban environments with complex morphological characteristics.

3.4. Slum Detection Using Uncertainty Quantification in 55 Globally Distributed Cities

3.4.1. Mapping Urban Poverty on a Large Scale

In Stark et al. (2024a), significant advantages were demonstrated through the use of a scene classification approach combined with uncertainty-aware prediction utilizing Monte Carlo dropout. The methodology was further refined in this study, where uncertainty estimation was enhanced by integrating both test-time augmentation and test-time dropout, resulting in higher-quality uncertainty estimations. This advancement enabled the application of the approach on a large global scale for the first time.

Leveraging HR data from 55 cities across the Global South, this study marks a substantial step toward the detection of global urban poverty. The identification and assessment of slums across diverse urban areas, from small to large, present unique complexities (Friesen et al., 2018; Kraff et al., 2019). Unlike well-known mega-cities, smaller and mid-sized cities often lack visibility and recognition in academic and policy circles, resulting in limited resources and attention towards their urban dynamics, including slum prevalence and characteristics. Recognizing the gradual change from informal to formal settlements and the inherent uncertainties in this spectrum is essential for advancing the field. Uncertainty aware methods provides valuable insights into the diverse morphologies of slum settlements and aids in developing more robust methodologies.

The cities used in this study were chosen based on their significant presence of densely built-up areas, specifically those classified under LCZ classes 3, 6, and 7, as identified by Zhu et al. (2019), and based on the classification scheme of (Stewart and Oke, 2012). These LCZ classes were selected because they typically exhibit dense settlement patterns, which are often associated with potential slum areas, posing a challenge in distinguishing between formal and informal settlements. Figure 13 illustrates the geographical dispersion of the selected cities, ranging from smaller urban centers like Ilorin, with 842,000 inhabitants, to megacities like Delhi, with over 17 million inhabitants.

These cities were analyzed using PlanetScope data from 2022, which provides Red, Green, and Blue channels at a 8bit radiometric resolution and a geometric resolution of 4.77 meters. Though most scenes featured minimal cloud coverage, some areas were occasionally obstructed. A consistent comparison of city scales was ensured by cropping the PlanetScope data using a bounding box around the morphological urban areas, as defined by (Taubenböck et al., 2019). This approach ensured an adequate representation of vegetation and water bodies within the classification schema, offering a broader environmental context.

The reference dataset was generated by integrating our own slum dataset from Stark et al. (2024a, 2020) with the LCZ dataset. The slum dataset was created through manual

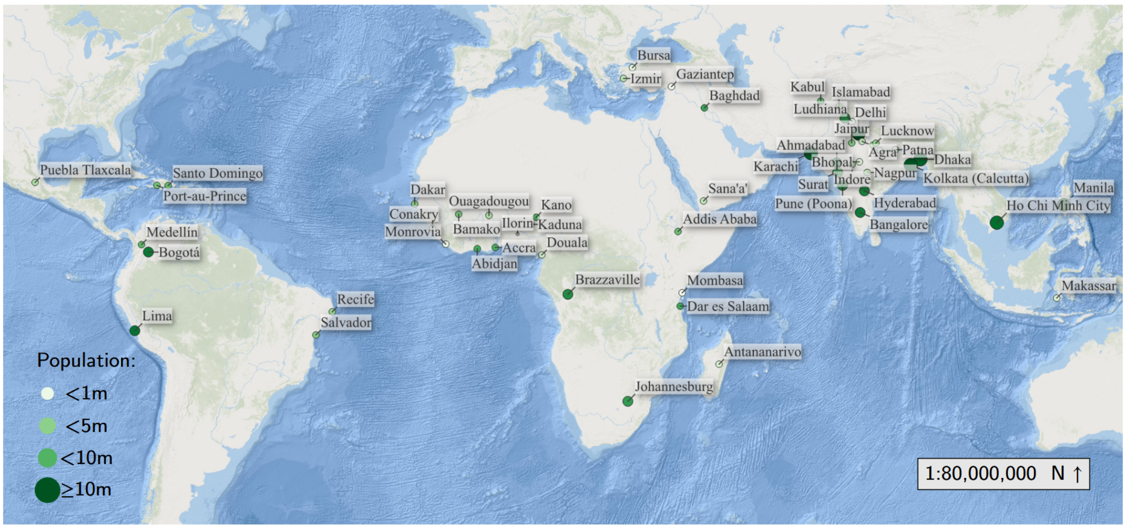

Figure 13 Location of the 55 cities across the Global South. The cities are scaled and colored by their population size.

mapping by remote sensing experts and aligned with the 2022 PlanetScope imagery. This mapping utilized additional sources such as Google Satellite imagery and, where accessible, Google Street View data. To ensure data accuracy, labels were limited to a maximum of five slum settlements per city, or roughly 4.6 square kilometers, with a focus on accuracy while acknowledging that many cities likely contain more slum areas than were mapped. To enhance the dataset, the LCZ data was reclassified into four distinct classes. This reclassification involved merging urban classes derived from LCZ classes 1 through 10, consolidating all non-built-up and vegetation classes from LCZ classes A(11) through F(16), and incorporating a water class. Subsequently, the slum data was amalgamated into this reclassified dataset, resulting in a comprehensive representation of the urban landscape.

The data was divided into smaller image tiles of 224×224 pixels, with an overlap of 45 pixels between tiles. Labels for these tiles were determined based on the majority class within each tile, with special consideration given to tiles containing at least 10% slum pixels, which were labeled as slum areas. This approach ensured that the dataset accurately reflected the diverse characteristics of urban environments, including both formal and informal settlements.

3.4.2. Transfer Learning Strategy Using Approximation of Uncertainty Estimation

The methodological approach can be seen in Figure 14, where transfer learning principles are leveraged. CNNs are first pretrained using data from four well-documented cities, Caracas, Mumbai, Nairobi, and Rio de Janeiro. This pretraining phase uses a large dataset of 143,188 image tiles, of which 21,448 were classified as slum areas. Following pretraining, the models were adapted to the target city's dataset through transfer learning, using a carefully balanced dataset of 100 image tiles per class. The slum samples were

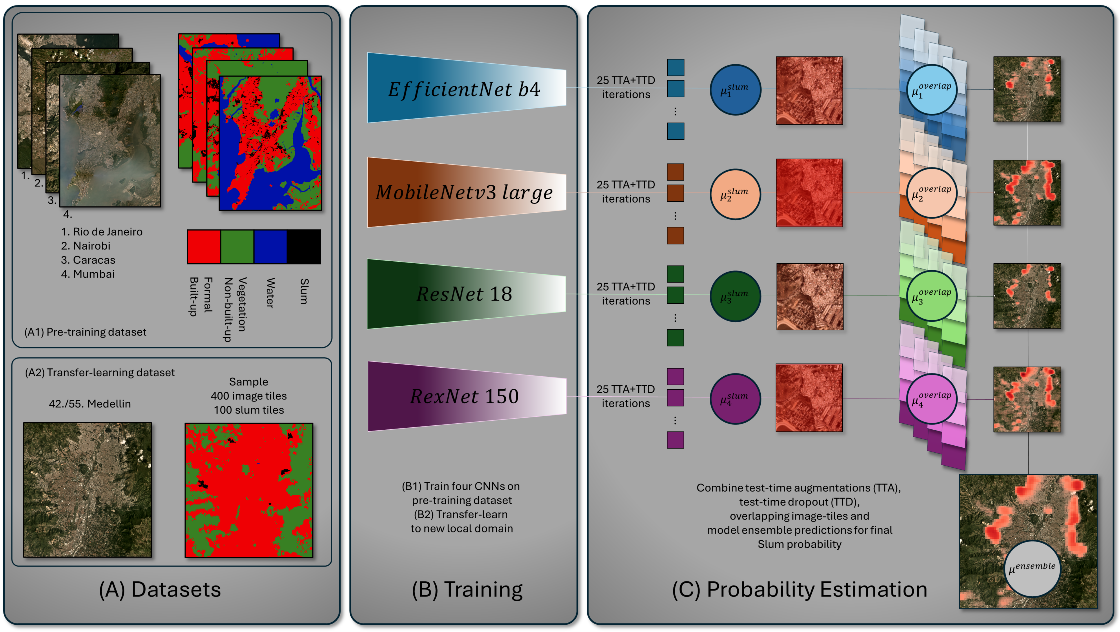

Figure 14 A simplified schematic overview of our approach to estimate slum probabilities on a large scale using Medellin, Colombia, as an example for one (42^{th}) of the 55 cities in our slum probability dataset.First, four CNNs are pre-trained on an initial imbalanced dataset, and these representations are subsequently transferred to a balanced city's dataset. The final slum probability is approximated using multiple methods, including test-time augmentations, test-time dropout, overlapping image tiles, and model ensembles.

selected from different areas to ensure geographic diversity. To enhance the reliability of predictions and account for uncertainty, approximation methods, incorporating test-time augmentations and dropout techniques are used. In addition an ensemble of four CNNs is included, enabling an aggregate of predictions that effectively enhance overall model variability.

Four CNNs were used to create a diverse model ensemble. Firstly, ResNet-18, as proposed by He et al. (2015), is a cornerstone in deep learning with 11.7 million parameters, balancing complexity and efficiency. Secondly, ReXNet-150, introduced by Han et al. (2021), comprises 9.8 million parameters, optimizing performance with effective channel dimension configuration. Thirdly, EfficientNet-B4, proposed by Tan and Le (2020), features 19.5 million parameters and scales depth, width, and resolution dimensions, offering substantial capacity for capturing intricate patterns. Lastly, MobileNetV3 Large, introduced by Howard et al. (2019), has 5.5 million parameters, excelling in resource-constrained environments with its compact yet powerful design. By incorporating these diverse CNN architectures, the goal is to benefit from their capabilities across a spectrum of tasks.

The tranfer-learning methodology, illustrated in Figure 14 begins by pretraining each CNN model initialized with ImageNet weights on a sizable but imbalanced remote sensing dataset. Figure 14 (A1) and (B1) shows the pre-training of CNNs using data from four well-known cities: Caracas, Mumbai, Nairobi, and Rio de Janeiro. These cities were selected for their well-documented spatial information from Stark et al. (2020, 2024b), enabling the creation of a large dataset consisting of 143,188 image tiles, with 21,448 tiles classified as slum areas.

Subsequently, transfer learning is employed to adapt these models to a specific city's dataset, which is carefully balanced with 100 image tiles per class, as illustrated in Figure 14 (A2) and (B2). For each city depicted in Figure 13, 100 samples per class were selected, totaling 400 samples. The classes include urban, vegetation, and water, which were randomly sampled. The slum samples are specifically sampled from different slum areas to ensure geographic diversity for transfer-learning, transfer-validation, and transfer-testing. It is important to note that while urban, vegetation, and water samples are drawn from the entire city, slum samples are limited to a few slum settlements. This means that due to random sampling, slum image tiles might be present in the urban, vegetation, and water categories, resulting in a class-balanced transfer-learning dataset but with potentially noisy labels. During the transfer-learning process, the entire CNN architecture remains trainable, with no layers being frozen.

Each transfer-learned model estimates the uncertainty of its predictions by averaging over 25 iterations (Gal and Ghahramani, 2016; Stark et al., 2024b). To gauge uncertainty, test-time augmentation is applied to address epistemic uncertainty (model uncertainty) by introducing various data augmentations to the test data, thereby reducing the model's lack of knowledge through consensus predictions. The same methods for data augmentation as described in Wang et al. (2019) are used. Test-time dropout addresses both epistemic uncertainty and aleatoric uncertainty (data uncertainty), capturing variability due to model uncertainty and intrinsic noise in the data (Ebel et al., 2023; Wang et al., 2019). For the dropout method, a value of 0.3 is applied, as shown in (Stark et al., 2024b).

For our experiments several steps regarding uncertainty approximations within the methodology are combined:

First, the logits of final layer L for $i = 1$ to 25 iterations are combined into an array of size $n \times i$, where $n = 4$ classes:

$$L = \begin{bmatrix} L_{11} & L_{12} & \dots & L_{1i} \\ L_{21} & L_{22} & \dots & L_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \dots & L_{ni} \end{bmatrix} \tag{3.1}$$

Next, a sigmoid function $\sigma(x)$ over the array to scale each network output to the range $[0, 1]$ is applied:

$$\sigma(L) = \begin{bmatrix} \sigma(L_{11}) & \sigma(L_{12}) & \dots & \sigma(L_{1i}) \\ \sigma(L_{21}) & \sigma(L_{22}) & \dots & \sigma(L_{2i}) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(L_{n1}) & \sigma(L_{n2}) & \dots & \sigma(L_{ni}) \end{bmatrix} \tag{3.2}$$

Finally, the mean is calculated for only the corresponding values to the 4^{th} class, defined as the slum class probability μ_{slum}.

$$\mu_{slum} = \frac{1}{i} \sum_{k=1}^{i} \sigma(L)_{4k} \tag{3.3}$$

After calculating the approximated slum probability μ_{slum}, each image tile is georeferenced to its original remote sensing data source. To address the inherent characteristics of remote sensing datasets and mitigate edge-related issues, a strategy of predicting on overlapping image tiles is employed.

Due to the use of overlapping image tiles, each tile overlaps with its neighboring tiles by 45 pixels in both the x and y directions. This overlap results in five overlapping predicted image tiles for each location in the original image. To generate a final prediction for each location, the mean probability is calculated from the five overlapping tiles. This process involves averaging the predicted probabilities, which helps to smooth out noise and improve the robustness of the predictions.

As a result, the final output is a set of image tiles with a size of 45 × 45 pixels, where each pixel represents the mean probability derived from the overlapping predictions. This approach leverages the redundancy provided by overlapping tiles to enhance the accuracy and reliability of the final probabilistic outputs.

The overlapping mean probability $\mu_{overlap}$ in equation 3.4 for each pixel (i, j) can be expressed as follows:

$$\mu_{overlap} = \frac{1}{5} \sum_{k=1}^{5} p_{i,j,k} \tag{3.4}$$

where $p_{i,j,k}$ is the predicted probability at the pixel (i, j) from the k-th overlapping tile, and k ranges from 1 to 5.

Our research adopts an additional approach of averaging model predictions across multiple models.The aim is to capture different sources of uncertainty stemming from variations in model architectures and initializations. By averaging these predictions, as shown in equation 3.5, the goal is to mitigate the uncertainty associated with individual models.

$$\mu_{ensemble} = \frac{1}{n} \sum_{j=1}^{n} \mu_4 \tag{3.5}$$

This strategy aims to enhance predictive stability and interpretability within ensemble learning frameworks, as shown in Figure 14 (C). Averaging predictions within each model reduces inherent variability and mitigates the influence of outliers or noisy predictions. Averaging within each model promotes greater stability, clearer interpretation of contributions, and stabilizes predictions despite high variability (Song et al., 2023).

3.4.3. Evaluation of Results and Discussion of Slum Variability

Table 11 presents the accuracy metrics for the four different classes, including the standard deviation of results from the 55 cities. These results are based on the transfer testing dataset. It is important to note that the testing dataset is noisy, as the mapped slum settlements in most of the 55 cities are incomplete and not fully represented. Nevertheless, Table 11 provides an indicator of the class-based accuracies of our method. The urban class achieved high accuracy with an F1 score of 95.27% ± 5.68%, indicating strong model performance in identifying formal built-up areas. Vegetation was also well-detected, with an F1 score of 92.06% ± 6.09%. However, the water class showed significant variability, with an F1 score of 73.98% ± 28.56%, suggesting that while water bodies were accurately identified when detected, some instances were missed.

The slum class presented the most significant challenges, with an F1 score of 47.41% ± 19.62%. Although the recall was high at 87.19% ± 21.50%, indicating that most slum areas were detected, the precision was much lower at 47.81% ± 29.81%. This discrepancy suggests that a considerable number of non-slum areas were incorrectly classified as slums, likely due to the incomplete nature of the reference dataset, where some correctly identified slums were not present in the reference data and were thus counted as false positives.

The study generated slum probability maps for all 55 cities, visualized in Figures 15 and 16. These maps reveal the varying probability of slum presence across different morphological urban areas. The visual analysis highlights the diversity in slum settlement sizes and probabilities, with some cities exhibiting small, concentrated slum pockets, while others display extensive slum areas. Cities such as Douala, Islamabad, and Johannesburg show

Table 11 Accuracy metrics for the four classes in our classification schema.

Class	F1 [%]	Precision [%]	Recall [%]
Built-up	95.27±5.68	95.42±2.17	95.63±8.46
Non built-up	92.06±6.09	94.74±4.70	89.84±8.25
Water	73.98±28.56	98.17±4.79	76.63±26.08
Slum	47.41±19.62	47.81±29.81	87.19±21.50

high slum probabilities, indicating widespread or more certain slum areas, while cities like Conakry, Ho Chi Minh City, and Delhi exhibit lower slum probabilities, reflecting either smaller or less certain slum areas.

The variability within and between cities underscores the complexity of slum classification, influenced by both the morphological characteristics of slums and the surrounding urban environment. This variability makes it challenging to uniformly describe slum areas across different cities, necessitating careful consideration of local contexts in slum mapping.

The analysis reveals that cities with high slum probabilities generally have extensive slum areas, while cities with lower probabilities might have smaller or less distinct slum settlements. The variability in slum probability distributions further emphasizes the need for tailored approaches to slum detection and classification in different urban contexts.

A significant challenge in this study was the incomplete reference dataset, which likely underrepresents the actual number of slum settlements. This limitation affected the precision of the model, as some areas identified as slums were not recognized in the reference data, leading to false positives. Addressing this issue requires improving the quality and completeness of slum mapping data to better align the model's predictions with reality.

Moreover, the variability in slum characteristics across different cities complicates the task of accurately identifying and categorizing slum areas. Slum morphology can vary widely, from informal makeshift housing to more permanent structures, necessitating a nuanced approach to classification that accounts for these differences. The study's findings highlight the importance of integrating HR data and comprehensive mapping efforts to enhance the accuracy and utility of slum detection models, ultimately supporting more effective urban planning and interventions. This research extends beyond traditional boundaries in slum mapping. Figure 17 demonstrates how the approach maps slum settlements across various geographical environments, identifying both typical and atypical slum morphologies. Different slum categories and their probabilities are detected, including areas where slums transition into formal settlements. This method offers an improvement over traditional binary classifications that only recognize typical slum characteristics.

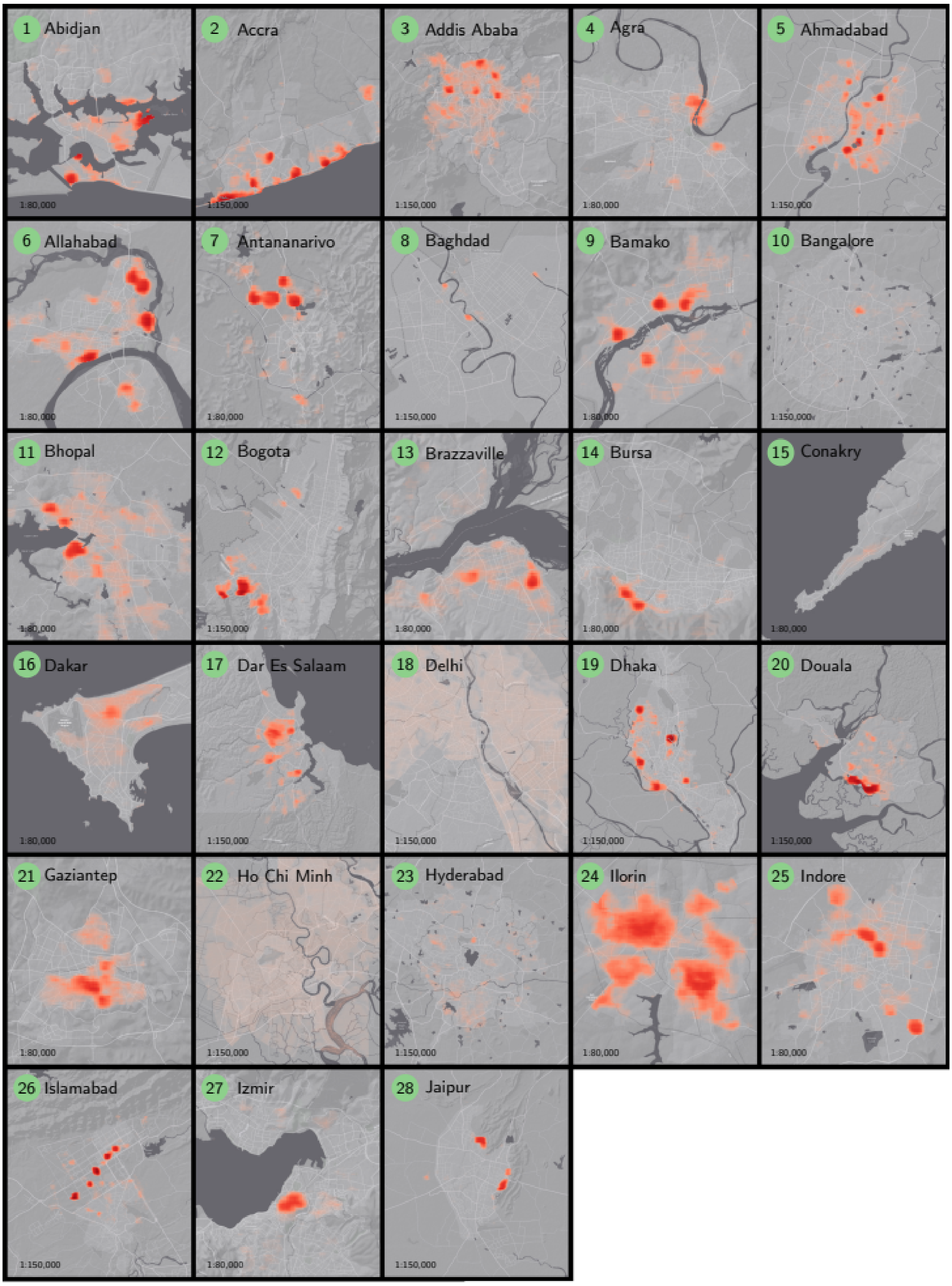

Figure 15 Slum probability maps for the cities 1 to 28.

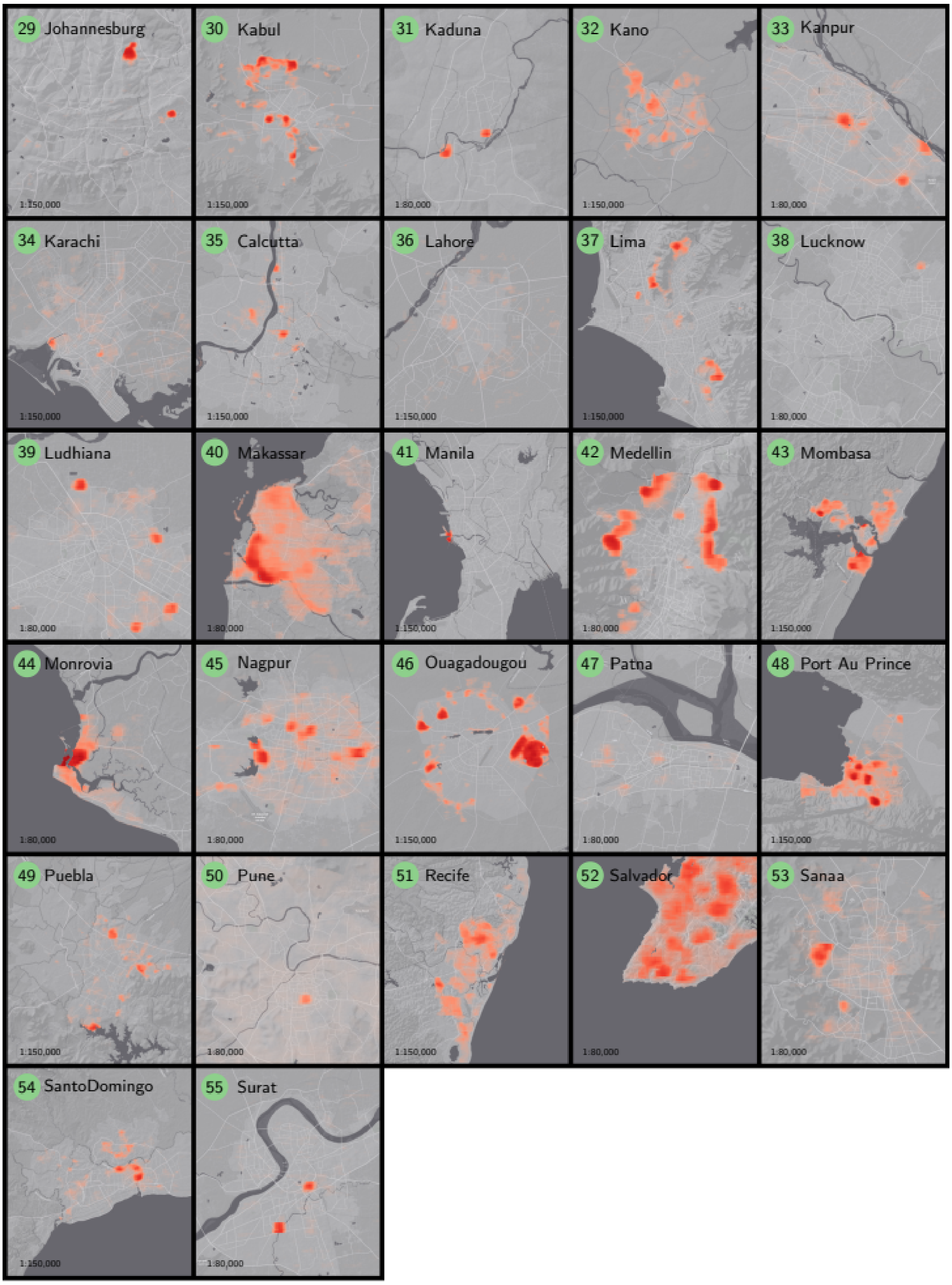

Figure 16 Slum probability maps for the cities 29 to 55.

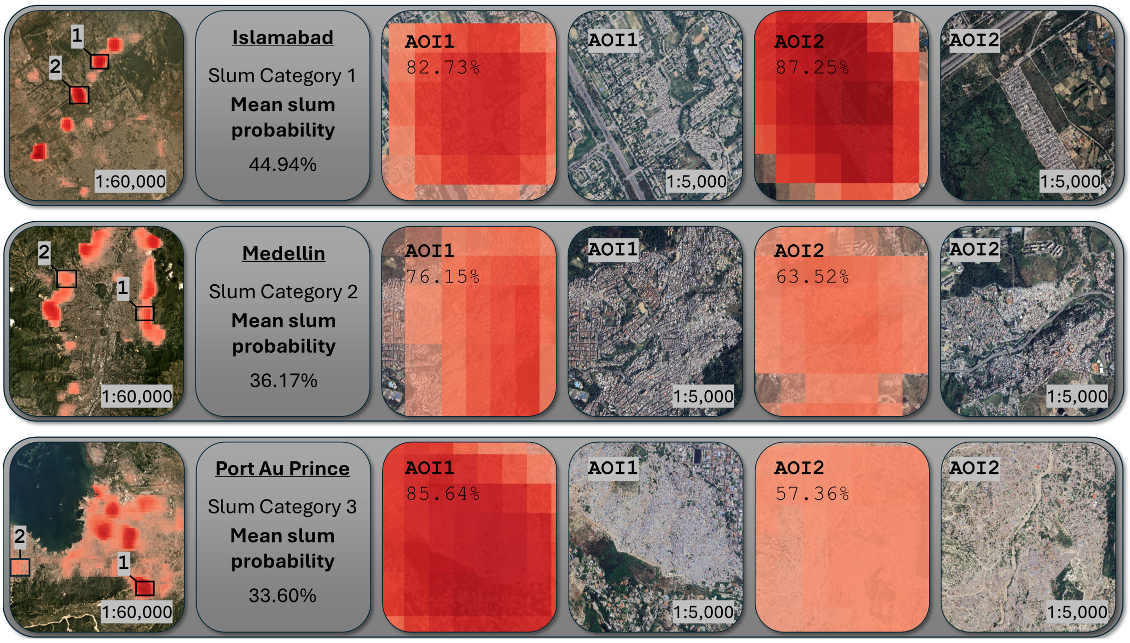

Figure 17 Examples from three cities, each representing a different slum category. The figure displays the city-wide slum probability on the left and two areas of interest (AOI) on the right. For each AOI, Google satellite imagery is shown. Slum probabilities are provided for the entire city and for each AOI.

In Islamabad, the capital of Pakistan with a population of approximately 1.2 million, slums can be characterised by typical slum morphologies (Rehman et al., 2022). A city-wide slum probability map (Figure 17) reveals numerous small slum settlements with high slum probability. In AOIs 1 and 2, these settlements have distinct borders separating them from formal settlements and non-built-up areas, resulting in high slum probability values.

Medellin, home to 2.6 million people, represents the second slum category, where slum settlements exhibit most morphological traits of slums but often include multi-story concrete buildings. Many slum areas in Medellin are vulnerable to landslides (Kühnl et al., 2023). The slum probability map (Figure 17) shows large areas with varying probabilities, indicating a gradual transition of slum areas into formal settlements, especially in AOI 1. In AOI 2, a slum settlement with lower slum probability lacks the more typical morphology of slums.

In Port Au Prince, Haiti's capital, over 60% of the population lives in low environmental quality conditions in densely populated areas (Joseph et al., 2014). The city's slum probability map shows low probability scores in many areas, with a few distinct high-probability slum settlements. This variability places Port Au Prince in the third slum category due to its intra-urban diversity. While AOI 1 exhibits typical slum characteristics, many neighborhoods, such as AOI 2, show atypical slum morphologies.

This research demonstrates the effective application of advanced machine learning and uncertainty-aware methods to map slum areas across 55 diverse cities. By utilizing transfer learning and ensemble predictions, the challenge of limited labeled data is overcome, achieving high accuracy in slum detection. The resulting slum probability maps provide valuable insights into urban poverty patterns, aiding policymakers and urban planners in addressing socio-economic disparities.

A coherent methodology was applied to a comprehensive slum dataset across the Global South, including probability estimates for each prediction. This approach offers a nuanced understanding of slum categories within each city, revealing intra- and inter-urban variability. These insights are crucial for tailoring interventions to specific urban needs, leading to more effective urban planning.

By integrating transfer learning with large-scale remote sensing data, the study enhances the understanding of urban environments and promotes sustainable development. The slum probability maps serve as valuable tools for addressing urban poverty and fostering equitable growth. This work highlights the potential of advanced machine learning in transforming urban analysis and addressing the complex challenges of cities in the Global South.

4. Discussion

4.1. Limitations of the Contributed Applications

4.1.1. Methodological Constraints in Mapping Slum Morphologies

Semantic segmentation has proven to be an effective tool for mapping slum morphologies in cities where typical slum structures are prevalent, such as Mumbai and Caracas (Fisher et al., 2022; Mahabir et al., 2018; Verma et al., 2019). In these cases, the distinct spatial characteristics of slums, including dense, informal settlements with irregular patterns, enable the segmentation algorithms to perform relatively well. However, when faced with more complex and atypical slum morphologies, such as those found in cities like Medellín and Port-au-Prince, where informal settlements coexist alongside formal housing structures, semantic segmentation encounters significant challenges. These mixed morphologies blur the boundary between formal and informal areas, making it difficult for conventional segmentation approaches to accurately delineate slums.

This is expected to become even more challenging as many urban areas in the Global South exhibit this gradual transition between informal and formal settlement structures. In many cities, typical slum morphologies do not exist as distinct clusters but rather as part of a continuum of different settlement types. This variability increases the difficulty of applying a semantic segmentation approach universally across many cities in the Global South. A purely semantic segmentation-based method is likely to struggle in consistently identifying slum areas in these complex urban environments, potentially leading to less robust and reliable results.

On the other hand, scene classification, while also limited in its ability to precisely differentiate between slum and formal settlements, offers an alternative perspective. In cities like Mumbai, where the distinction between informal and formal areas is stark, scene classification may not always yield the most accurate delineation of slum boundaries. However, when combined with uncertainty-aware methods, scene classification becomes more advantageous. Introducing uncertainty measures allows for a gradual change in the confidence of predicted classes, which better captures the transitional nature of slum-like settlement structures. This is particularly beneficial in cities where the boundaries between slums and formal settlements are not clearly defined. These aspects can be seen in Figure 18 for the city of Medellin, where two close up views of slum settlements are highlighted 18(2) and 18(3). By incorporating uncertainty, the scene classification outputs reflect a more nuanced understanding of the urban fabric, offering a significant improvement over standard semantic segmentation results in these contexts. In Figure 18(2), the slum is classified as having mixed slum morphologies. The gradual transition into a formal settlement further complicates the classification. Notably, only the scene classification approach

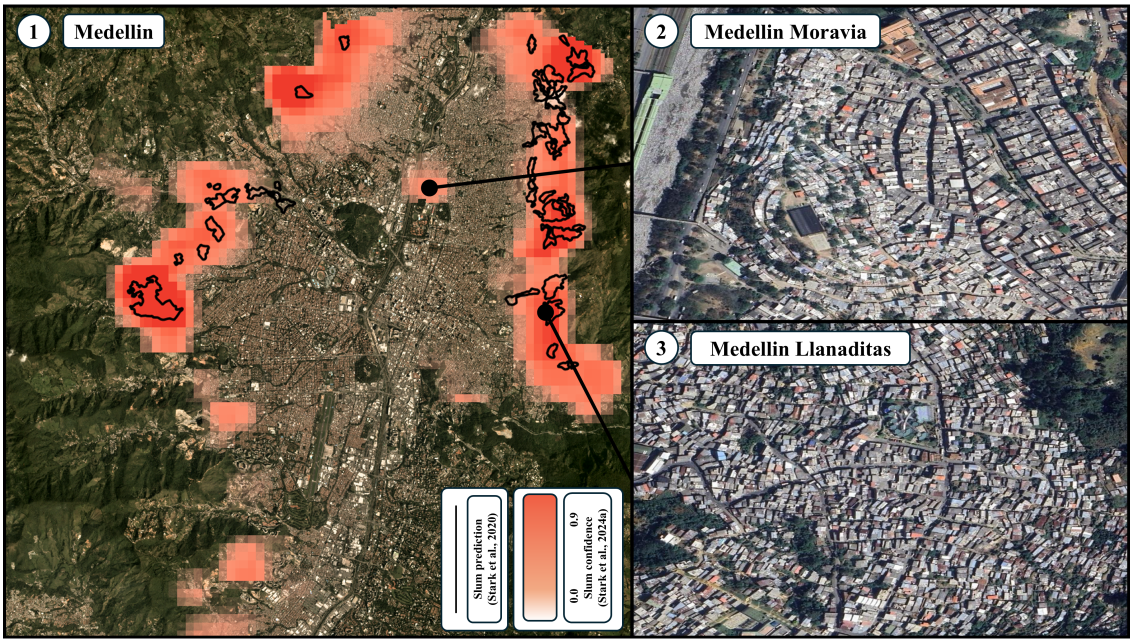

Figure 18 Comparing the results for semantic segmentation from Stark et al. (2020) to the slum confidence results from Stark et al. (2024a) in (1). (2) and (3) show two areas of interest where only the approach from Stark et al. (2024a) was able to detect the slum settlement in (2).

incorporating uncertainty estimations from Stark et al. (2024a) successfully detected the slum settlement, whereas the semantic segmentation approach failed to achieve accurate detection (Stark et al., 2020).

Lastly, although not explored in this study, there are methods available for mapping individual buildings within slum settlements, such as instance segmentation. Instance segmentation presents a distinct set of challenges, particularly in densely packed slum areas, where buildings are often clustered tightly together. Differentiating between individual structures becomes difficult, especially when buildings are constructed using multiple materials for a single roof. This complexity makes it challenging to accurately identify and segment individual buildings in typical slum environments. Nevertheless, recent studies, such as **?**$[]+\}(?![]*[.!?]), havebeguntoexplorethepotentialofinstancesegmentationinthesecontexts, pavingthe$

In summary, while semantic segmentation and scene classification each have their strengths, the complexity of slum morphologies across the Global South necessitates the exploration of more nuanced approaches, including uncertainty methods to achieve more accurate and robust mapping of slum settlements.

4.1.2. The Effects of Remote Sensing Data Resolution

Higher-resolution remote sensing data generally leads to more accurate results in mapping and analysis tasks as presented in **?**$[]+\}(?![]*[.!?])insection3.1.IfVHRdatawereavailablegloballyforal$

In contrast, globally available satellite platforms such as PlanetScope and Sentinel-2 offer a more standardized option for scientific studies. These platforms provide consistent metadata, which is crucial for making research findings more comparable and reliable. However, a trade-off exists: the geometric resolution of PlanetScope and Sentinel-2 is lower than that of VHR data. While this lower resolution might reduce the precision of certain analyses, the global coverage and availability of these datasets make them more practical for large-scale urban studies in the Global South. Therefore, despite the limitations in resolution, platforms like Sentinel-2 and PlanetScope are often the preferred choice for research due to their accessibility, consistency, and cost-effectiveness.

4.1.3. Unreliable Ground Truth Data and Labor-Intensive Reference Collection

Transfer learning has shown significant promise in the successful mapping of slum areas. However, for each city some slum sample are needed as even a small number of local samples have been demonstrated to significantly outperform standard classification from a more generalized model, previously trained on slum data from other geographical regions. Despite advancements in reducing the required number of samples, such as through few-shot learning methods (Stark et al., 2023), the collection of reference data remains essential. Although ongoing research aims to minimize the number of required samples, reference data is still a critical component of slum mapping.

The development of larger foundational models (Xiong et al., 2024) or the creation of expansive slum datasets (Thomson et al., 2020) in the future could greatly enhance inference accuracy and provide more robust slum mapping solutions. These advancements would be invaluable in overcoming current limitations associated with data collection.

A significant challenge in slum mapping is the discrepancy between reference data and ground truth data (Kraff et al., 2019). Slums are highly heterogeneous, with different definitions and categories depending on the geographic region. This variability makes it difficult to define a consistent morphological feature space for slums. As a result, there is often a mismatch between the reference data used to train the models and the actual ground truth of slum settlements. Additionally, ground truth data, where available, may be outdated due to the dynamic and rapidly evolving nature of slums (Gevaert et al., 2019). This temporal discrepancy introduces further uncertainty into the mapping process and highlights the need for ongoing refinement of reference datasets and the methods used to train slum-mapping models.

4.2. Significance of Contributions to Global Poverty Mitigation Initiatives

In chapter 1, the SDGs and the UN's Call for Better Data was introduced. This section explores how the contributions of this work can support, inform, and enhance these and other critical initiatives.

4.2.1. Impact Towards the Sustainable Developments Goals

As the first of the SDGs, Goal 1 underscores its fundamental importance by aiming to eradicate poverty in all its forms globally. SDG 1 encompasses seven key targets that contribute to this overarching goal. These targets include eradicating extreme poverty and reducing overall poverty by 2030, implementing social protection systems, ensuring equal access to economic resources and basic services, building resilience against climate-related and other shocks, mobilizing resources for poverty programs, and establishing policy frameworks that support pro-poor and gender-sensitive strategies for poverty eradication.

A critical factor influencing the success of these targets is the accuracy and reliability of the underlying data used to estimate the number of people living in extreme poverty. Inaccurate or outdated data can significantly undermine efforts to track and address poverty, leading to misinformed policy decisions and resource allocation (Kuffer et al., 2019). The report from the World Bank for the city Port-au-Prince exemplifies the potential impact of such data-related issues (D'Aoust et al., 2022). The report identified vulnerable city sectors, highlighting areas such as Martissant, Cité Soleil, western Carrefour, and regions along the Grise River and Kenscoff route as having the highest levels of vulnerability, predominantly in slum areas. In contrast, less vulnerable regions were identified in parts of Pétion-Ville, Pacot, and the sparsely populated northern and eastern areas of Carrefour.

When comparing these findings to the slum confidence map from section 3.4 of Port-au-Prince seen in Figure 19, a strong correlation is evident. However, the slum confidence map offers several advantages, including higher resolution and the inclusion of confidence scores, providing a more nuanced understanding of the spatial distribution of poverty. These confidence scores allow policymakers to prioritize interventions more effectively by identifying areas with uncertain data, thereby enabling more targeted and informed decision-making. In the two highlighted examples in Figure 19 it can also be seen that the proposed approach from Stark et al. (2024a) is able to identify slum settlements within the city center more clearly. This example highlights the importance of utilizing advanced methodologies and high-quality data to further support the targets of SDG 1.

Global efforts to eradicate extreme poverty have been severely disrupted by the COVID-19 pandemic and a series of major shocks from 2020 to 2022. The pandemic reversed decades of progress, increasing extreme poverty for the first time in years, and setting global progress back by three years. Recovery has been uneven, with low-income countries struggling the most, and achieving the goal of ending poverty by 2030 is now unlikely (United Nations, 2023b). By 2022, 9% of the world's population, 712 million people, were living in extreme poverty, with projections indicating that 590 million will still live in extreme poverty by 2030. The pandemic has also slowed progress on halving national poverty rates, with less than 30% of countries on track to meet this goal by 2030. Social protection for children remains inadequate, with 1.4 billion children lacking coverage in 2023. Furthermore, economic losses due to disasters continue to exceed $115 billion annu-

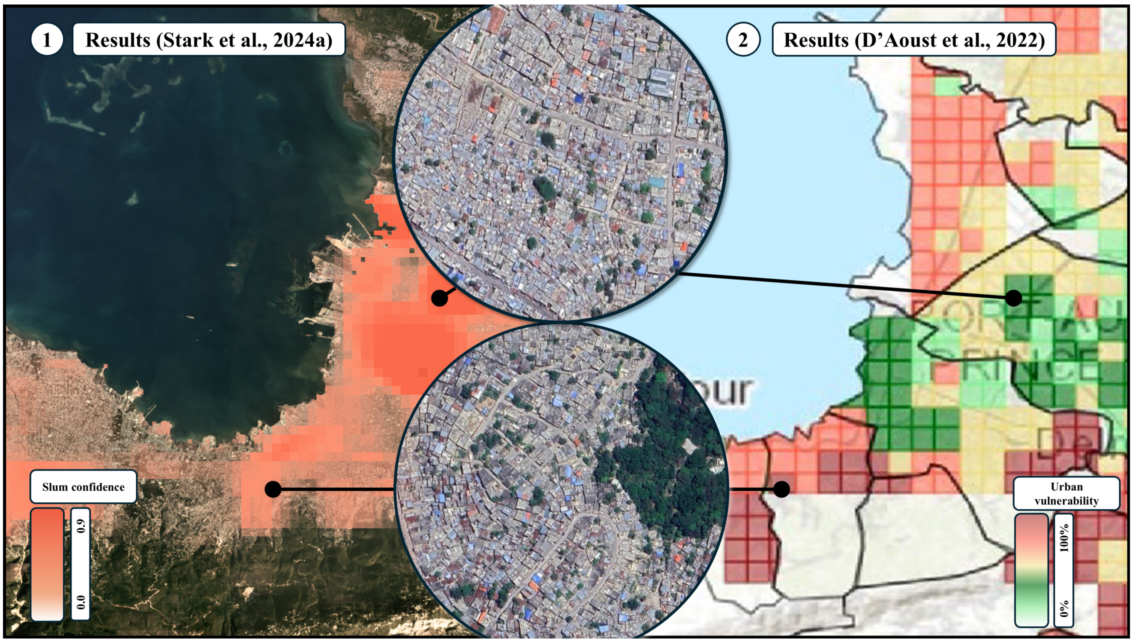

Figure 19 Comparing the results from the World Bank report from D'Aoust et al. (2022) of mapping urban vulnerability to the slum confidence results from (Stark et al., 2024a). Two area of interests of slums are highlighted in where only the results from Stark et al. (2024a) are able to detect slums within the city center.

ally, showing no signs of improvement (United Nations, 2023b). Government spending on essential services has remained stable, with a persistent gap between advanced economies and developing nations. This is why it is crucial to have updated, high-quality global data on slum settlements (United Nations, 2024). While global slum mapping is not yet fully realized, the methods presented in this work demonstrate a promising step forward. With further improvements, a comprehensive global slum map could become an invaluable tool for advancing the goals of SDG 1.

SDG 11 aims to make cities and human settlements inclusive, safe, resilient, and sustainable, directly addressing the challenges of slums and extreme urban poverty. Slum areas, often characterized by poor-quality housing and inadequate infrastructure, are particularly vulnerable to disasters due to unsafe building materials and lack of disaster-resilient planning. In addition, limited road safety and infrastructure exacerbate the risks faced by slum dwellers. Achieving SDG 11 requires reducing disaster risks, improving housing conditions, and ensuring that urban development supports safer, more resilient communities, especially in slum settlements where vulnerabilities are most acute (Kühnl et al., 2023).

Similarly SDG 6 aims to ensure clean water and sanitation for all, a goal particularly challenging in slum settlements. Slums often lack basic infrastructure, including proper water supply systems and sanitation facilities, making access to clean water and hygiene difficult. Overcrowding and poor waste management exacerbate water contamination risks, leading to heightened public health issues. Achieving SDG 6 in these regions necessitates substantial investment in sustainable infrastructure and the development of innovative solutions to ensure access to clean water and adequate sanitation in densely populated, informal urban settlements where conventional systems are challenging to implement. The results, particularly those derived from semantic segmentation, offer valuable insights into the spatial extent of slums, enabling more accurate estimates of slum populations. This information serves as an optimal input for designing efficient water distribution networks and supply infrastructure in urban areas (Friesen et al., 2017; Stark, 2018).

4.2.2. Poverty Data Repositories

The UN Call for Better Data aims to improve global efforts in monitoring and eliminating extreme poverty by leveraging big data for official statistics (United Nations, 2024). Statisticians worldwide will collaborate virtually through the UN Statistical Commission to advance this initiative. The key objectives include providing strategic direction for a global big data program, promoting practical applications of big data while addressing challenges such as methodology, legal concerns, and security, and enhancing capacity-building efforts. Additionally, the initiative advocates for the use of big data in policy-making and works to build public trust in its application for official statistics.

The IDEAMAPS Project is a global initiative aimed at creating a comprehensive slum repository to support NGOs and local governments (Kuffer et al., 2024; Thomson et al., 2020). This research network focuses on improving methods for mapping slum areas by generating citywide maps that highlight deprivations and assets. The project helps stakeholders use this data for urban upgrading, advocacy, and monitoring efforts. IDEAMAPS emphasizes the importance of data validation by city stakeholders, ensuring that the data is comparable across cities, regularly updated, and accessible to communities and local governments to promote equity and justice in urban planning and development.

All results presented in this dissertation are fully reproducible and made available through multiple GitHub repositories including some example data (Stark et al., 2024a, Stark et al., 2024b). This marks an important step towards making the methods accessible to the public, allowing others to replicate and build upon the work. Prediction maps are shared upon reasonable request within established communities, such as the "Slum Modeling Community of Practice". However, publicly sharing these predictions poses ethical challenges, which are discussed in the following section. In the future, when models are more refined or when sharing aggregated statistical values per city, the approach to data sharing could become more flexible.

4.3. Ethical Considerations in AI Driven Slum Mapping

The use of deep learning methods in urban remote sensing offers vast potential for analyzing and understanding cities at scale. However, it also presents significant ethical challenges. Deep learning models trained on remote sensing data, such as satellite or aerial imagery, can provide insights into various aspects of urban development, infrastructure, and environmental conditions. These insights can aid in urban planning, disaster management, and sustainable development. Nevertheless, issues related to privacy, surveillance, and data ownership must be carefully considered (Kochupillai et al., 2022; Zhu et al., 2022).

One major ethical concern is the potential for Artificial Intelligence (AI) to infringe on individual privacy. High-resolution imagery combined with AI can lead to unintended surveillance, where personal or sensitive information about individuals or communities is inadvertently exposed (Kamila and Jasrotia, 2023). There is also the risk of bias in AI models, as these models may disproportionately impact vulnerable populations if they are trained on biased or incomplete data (Jobin et al., 2019). For instance, urban remote sensing can reinforce existing inequalities if the AI systems are not designed to fairly represent diverse communities. Additionally, transparency and accountability are crucial decisions are made based on AI predictions (Kim et al., 2020). The outcomes of AI models should be explainable and accessible, ensuring that urban planning decisions driven by AI are made ethically and with public oversight (Höhl et al., 2024).

In the context of labeling and predicting living conditions using AI, ethical considerations become particularly sensitive. The process of classifying areas based on living standards risks stigmatizing communities, especially when predictions or classifications are not contextually accurate (Reijneveld et al., 2000). Such data can inadvertently reinforce stereotypes or cause harm by misjudging people's living environments. Therefore, predictions must be handled with great care, ensuring that the data does not lead to negative societal consequences or unfairly label communities (Jaber and Abbad, 2023).

While there is a desire to make these AI-generated insights widely available for research, urban planning, and policymaking, the ethical balance is delicate (Zhu et al., 2022). On the one hand, sharing data can drive important interventions and improvements in urban development. On the other hand, the risk of misrepresenting living conditions could result in policy decisions that negatively affect the very communities the data is meant to help (Corburn and Sverdlik, 2017; Owusu et al., 2021). Therefore, AI models need to be designed with safeguards that minimize bias, respect cultural and societal contexts, and ensure that outputs do not stigmatize individuals or communities. In the future, sharing only aggregated data or statistical values, rather than direct classifications of living conditions, could be a more ethically sound approach, helping to mitigate these risks while still providing valuable insights.

5. Conclusion and Outlook

5.1. Conclusion

This research sought to answer four key questions regarding the use of deep learning in remote sensing for large-scale slum mapping. First, an investigation was conducted into whether deep learning methods can be effectively applied to map slums with typical morphologies using satellite imagery. Our experiments demonstrated that FCNs, when applied to VHR imagery such as QuickBird, yield highly accurate results for typical slum areas, such as those found in Mumbai. The detailed spatial resolution of 0.5 meter allowed the network to capture the complex, irregular patterns characteristic of slum settlements, achieving strong segmentation outcomes. Transfer learning further enhanced these results, showing promise for the detection of slums using HR sensors like Sentinel-2, making large-scale slum mapping more feasible.

Secondly, an exploration was conducted into whether these methods could be scaled to map slums on a globally distributed level. Through transfer learning, a model was able to generalize across different geographic regions, providing robust slum predictions in cities with varying slum morphologies. Experiments showed that using a transfer-learned XFCN mapping accuracy improves significantly compared to training the XFCN within only one single city. This suggests that it is possible to leverage deep learning methods to map slums at a global scale by using HR imagery, such as PlanetScope data.

Thirdly, detecting the gradual transition between formal and informal settlements requires uncertainty estimates to distinguish between these fuzzy boundaries. By incorporating Monte Carlo dropout into the deep learning models the confidence of slum predictions could be assessed, which is crucial in areas where formal and informal features overlap. The uncertainty-aware approach proved essential in refining predictions, particularly in urban environments where slums share features with formal settlements, making classification more complex. The introduction of uncertainty not only increased the robustness of the results but also highlighted areas where the model was less confident, providing valuable information for further analysis or ground validation.

Finally, the methodology was applied to a large-scale dataset, mapping slums across 55 diverse cities. The combination of transfer learning and uncertainty-aware methods produced slum probability maps that offer detailed insights into urban poverty patterns. These maps provide a critical tool for policymakers and urban planners, enabling them to address socio-economic disparities with more targeted interventions. The successful application of our methods across a wide geographic range underscores the potential of deep learning methods in transforming urban analysis, offering salable, accurate, and ethically sound solutions for slum mapping on a global scale.

In summary, this dissertation demonstrated that deep learning methods, combined with remote sensing datasets, can effectively map slums across various geographical settings. The presented approaches could handle the complexities of global-scale mapping and provide uncertainty estimates to navigate the blurry boundaries between formal and informal settlements. By applying these methods to a wide variety of globally distributed cities, the foundation for future work in large-scale urban poverty mapping is laid, offering valuable tools for sustainable development and equitable urban planning.

5.2. Outlook

To build upon the results of this thesis and address the ongoing need for slum mapping (United Nations, 2023b), further advancements in deep learning for detecting slums are necessary. These developments will help create applications that can scale globally. Given the rapid advancements in deep learning, several potential directions can be explored for slum mapping using remote sensing data. These include expanding the models to handle diverse urban environments on a global scale, integrating more complex data sources, and improving the robustness of predictions in diverse and challenging urban contexts, such as slums. The following outlines potential areas for future research and applications in this domain.

Deep learning methods are rapidly evolving, model architectures often achieve higher accuracies in benchmark datasets by utilizing larger models or optimizing parameter efficiency, all while maintaining reasonable processing times (Brown et al., 2020; Tan and Le, 2020). However, these advancements typically rely on large datasets, which are often scarce in the context of slum mapping. Despite this, some emerging techniques could still prove beneficial for slum mapping. These methods have the potential to enhance accuracy and generalization even with limited data, offering promising methods for future research in this field.

The rise of Vision Transformers (ViTs) presents a potential improvement over standard CNN architectures for slum mapping. Unlike CNNs, which rely on convolutional layers to capture spatial hierarchies, ViTs use self-attention mechanisms to process entire images at once, capturing global context more effectively. This can lead to better performance in certain tasks, especially with larger datasets (Maurício et al., 2023; Wei et al., 2022). Given their success in benchmark datasets, it would be interesting to explore whether similar achievements could be realized in slum mapping, particularly in addressing the need for diverse and expansive datasets, as discussed earlier. Incorporating ViTs could offer a promising direction for improving the accuracy and generalization of slum mapping models.

Several promising research directions can enhance the scalability and effectiveness of slum mapping using deep learning and remote sensing data. Active learning, for instance, could allow models to become more efficient by selecting the most informative data points for training, significantly reducing the need for large labeled datasets (Geiß et al., 2017; Tuia et al., 2009). This approach, coupled with more advanced transfer learning and few-shot learning techniques, could further improve the accuracy and adaptability of models in new geographic regions, even when limited training data is available (Stark et al., 2023). These methods are particularly relevant for applications in slum mapping, where labeled datasets are often sparse, and urban structures can vary dramatically across different cities.

Another exciting avenue for future research is the application of self-supervised learning strategies, which could enable large-scale slum mapping in the real world without relying heavily on labeled data. By allowing models to learn from vast amounts of unlabeled imagery, self-supervised learning could facilitate the creation of global slum maps with less human intervention and greater scalability (Li et al., 2021; Tao et al., 2023; Wang et al., 2022b).

Ethical considerations remain crucial in the development of deep learning based slum mapping tools. Explainable AI (XAI) must be prioritized to ensure that models are transparent, responsible, and free from bias (Jobin et al., 2019; Owusu et al., 2021; Zhu et al., 2022). Given the sensitivity of slum mapping, where the misclassification of communities can lead to stigmatization, ensuring fairness and accountability in AI predictions is essential. This would foster trust among stakeholders and make the results of these models more ethically sound and usable in real-world applications.

Lastly, making public datasets available for slum mapping poses ethical challenges, especially when georeferenced data might expose vulnerable communities. However, providing unreferenced visual data or aggregated statistical data derived from the results could offer valuable insights while maintaining privacy and ethical standards. The future of slum mapping holds immense potential. With improvements in AI transparency and data sharing, along with scalable learning techniques, the tools developed through this research could transform urban analysis, ultimately leading to smarter, more equitable solutions for poverty alleviation.

Bibliography

Abdollahnejad, A., Panagiotidis, D., Shataee Joybari, S., and Surový, P. (2017). Prediction of Dominant Forest Tree Species Using QuickBird and Environmental Data. *Forests*, 8(2):42.

Ajami, A., Kuffer, M., Persello, C., and Pfeffer, K. (2019). Identifying a Slums' Degree of Deprivation from VHR Images Using Convolutional Neural Networks. *Remote Sensing*, 11(11):1282.

Aravena Pelizari, P., Geiß, C., Groth, S., and Taubenböck, H. (2023). Deep multitask learning with label interdependency distillation for multicriteria street-level image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204:275–290.

Barros Filho, M. and Sobreira, F. (2008). Accuracy of lacunarity algorithms in texture classification of high spatial resolution images from urban areas. In *XXI congress of international society of photogrammetry and remote sensing*, pages 417–422.

Baud, I., Kuffer, M., Pfeffer, K., Sliuzas, R., and Karuppannan, S. (2010). Understanding heterogeneity in metropolitan India: The added value of remote sensing data for analyzing sub-standard residential areas. *International Journal of Applied Earth Observation and Geoinformation*, 12(5):359–374.

Baud, I. S. A., Pfeffer, K., Sridharan, N., and Nainan, N. (2009). Matching deprivation mapping to urban governance in three Indian mega-cities. *Habitat International*, 33(4):365–377.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Chen, D., Tu, W., Cao, R., Zhang, Y., He, B., Wang, C., Shi, T., and Li, Q. (2022). A hierarchical approach for fine-grained urban villages recognition fusing remote and social sensing data. *International Journal of Applied Earth Observation and Geoinformation*, 106:102661.

Chen, Q., Cheng, Q., Wang, J., Du, M., Zhou, L., and Liu, Y. (2021). Identification and Evaluation of Urban Construction Waste with VHR Remote Sensing Using Multi-Feature Analysis and a Hierarchical Segmentation Method. *Remote Sensing*, 13(1):158.

Chen, T.-H. K., Qiu, C., Schmitt, M., Zhu, X. X., Sabel, C. E., and Prishchepov, A. V. (2020). Mapping horizontal and vertical urban densification in Denmark with Landsat time-series from 1985 to 2018: A semantic segmentation solution. *Remote Sensing of Environment*, 251:112096.

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. Technical report. arXiv:1610.02357 [cs] type: article.

Corburn, J. and Sverdlik, A. (2017). Slum Upgrading and Health Equity. *International Journal of Environmental Research and Public Health*, 14(4):342.

Dabra, A. and Kumar, V. (2023). Evaluating green cover and open spaces in informal settlements of Mumbai using deep learning. *Neural Computing and Applications*, 35(16):11773–11788.

D'Aoust, O., Gunneman, J., Patel, K. V., and Tassot, C. (2022). *Cash in the city: the case of Port-au-Prince*. World Bank.

Debray, H., Kraff, N. J., Zhu, X. X., and Taubenböck, H. (2023). Planned, unplanned, or in-between? A concept of the intensity of plannedness and its empirical relation to the built urban landscape across the globe. *Landscape and Urban Planning*, 233:104711.

Doda, S., Kahl, M., Ouan, K., Obadic, I., Wang, Y., Taubenböck, H., and Zhu, X. X. (2024). Interpretable deep learning for consistent large-scale urban population estimation using Earth observation data. *International Journal of Applied Earth Observation and Geoinformation*, 128:103731.

Dovey, K. and Kamalipour, H. (2017). Informal/Formal Morphologies. Routledge.

Dufitimana, E. and Niyonzima, T. (2023). Leveraging the Potential of Convolutional Neural Network and Satellite Images to Map Informal Settlements in Urban Settings of the City of Kigali, Rwanda. *Rwanda Journal of Engineering, Science, Technology and Environment*, 5(1).

Duque, J. C., Patino, J. E., Ruiz, L. A., and Pardo-Pascual, J. E. (2015). Measuring intra-urban poverty using land cover and texture metrics derived from remote sensing data. *Landscape and Urban Planning*, 135:11–21.

Ebel, P., Garnot, V. S. F., Schmitt, M., Wegner, J. D., and Zhu, X. X. (2023). Uncrtaints: Uncertainty quantification for cloud removal in optical satellite time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2086–2096.

Escalante, B. (2012). *Remote Sensing: Applications*. BoD – Books on Demand.

Esch, T., Taubenböck, H., Heldens, W., Thiel, M., Wurm, M., Geiß, C., and Dech, S. (2010a). Urban remote sensing–how can earth observation support the sustainable development of urban environments? In *Proceedings*, pages 1–11.

Esch, T., Thiel, M., Schenk, A., Roth, A., Muller, A., and Dech, S. (2010b). Delineation of Urban Footprints From TerraSAR-X Data by Analyzing Speckle Characteristics and Intensity Information. *IEEE Transactions on Geoscience and Remote Sensing*, 48(2):905–916.

European Union (2023). Poverty and social exclusion. `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Poverty_statistics`. Accessed: 2024-08-14.

Filho, M. B. and Sobreira, F. (2005). Assessing texture pattern in slum across scales: an unsupervised approach.

Fisher, T., Gibson, H., Liu, Y., Abdar, M., Posa, M., Salimi-Khorshidi, G., Hassaine, A., Cai, Y., Rahimi, K., and Mamouei, M. (2022). Uncertainty-Aware Interpretable Deep Learning for Slum Mapping and Monitoring. *Remote Sensing*, 14(13):3072.

Frazier, A. E. and Hemingway, B. L. (2021). A Technical Review of Planet Smallsat Data: Practical Considerations for Processing and Using PlanetScope Imagery. *Remote Sensing*, 13(19):3930.

Friesen, J., Kraff, N. J., and Taubenböck, H. (2024). The Spatiotemporal Dynamics of Morphological Slums in Mumbai, India. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:13824–13836.

Friesen, J., Rausch, L., and Pelz, P. F. (2017). Providing water for the poor - towards optimal water supply infrastructures for informal settlements by using remote sensing data. In *2017 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4.

Friesen, J., Taubenböck, H., Wurm, M., and Pelz, P. F. (2018). The similar size of slums. *Habitat International*, 73:79–88.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. pages 1050–1059. PMLR.

Galeon, F. (2011). Determining formalities of settlement clusters using fractal dimensions. Federation Internationale des Geometres.

Geiß, C., Thoma, M., Pittore, M., Wieland, M., Dech, S. W., and Taubenböck, H. (2017). Multitask Active Learning for Characterization of Built Environments With Multisensor Earth Observation Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(12):5583–5597.

Gevaert, C. M., Kohli, D., and Kuffer, M. (2019). Challenges of mapping the missing spaces. In *2019 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4. ISSN: 2642-9535.

Giada, S., De Groeve, T., Ehrlich, D., and Soille, P. (2003). Information extraction from very high resolution satellite imagery over Lukole refugee camp, Tanzania. *International Journal of Remote Sensing*, 24(22):4251–4266.

Gopalakrishnan, K., Khaitan, S. K., Choudhary, A., and Agrawal, A. (2017). Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection. *Construction and Building Materials*, 157:322–330.

Gosteva, A. A., Matuzko, A. K., and Yakubailik, O. E. (2019). Detection of changes in urban environment based on infrared satellite data. *IOP Conference Series: Materials Science and Engineering*, 537(6):062051.

Gram-Hansen, B. J., Helber, P., Varatharajan, I., Azam, F., Coca-Castro, A., Kopackova, V., and Bilinski, P. (2019). Mapping Informal Settlements in Developing Countries using Machine Learning and Low Resolution Multi-spectral Data. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 361–368, New York, NY, USA. Association for Computing Machinery.

Han, D., Yun, S., Heo, B., and Yoo, Y. (2021). Rethinking Channel Dimensions for Efficient Model Design. Technical report. arXiv:2007.00992 [cs] type: article.

Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. Technical report. arXiv:1512.03385 [cs] type: article.

Hofmann, P. et al. (2001). Detecting informal settlements from ikonos image data using methods of object oriented image analysis-an example from cape town (south africa). *Jürgens, C.(Ed.): Remote Sensing of Urban Areas/Fernerkundung in urbanen Räumen*, pages 41–42.

Höhl, A., Obadic, I., Fernández-Torres, M.-A., Oliveira, D., and Zhu, X. X. (2024). Recent trends challenges and limitations of explainable ai in remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 8199–8205.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., and Adam, H. (2019). Searching for MobileNetV3. Technical report. arXiv:1905.02244 [cs] type: article.

Huang, J., Lu, X. X., and Sellers, J. M. (2007). A global comparative analysis of urban form: Applying spatial metrics and remote sensing. *Landscape and Urban Planning*, 82(4):184–197.

Huang, Y., Zhang, F., Gao, Y., Tu, W., Duarte, F., Ratti, C., Guo, D., and Liu, Y. (2023). Comprehensive urban space representation with varying numbers of street-level images. *Computers, Environment and Urban Systems*, 106:102043.

Hunter, L. M. (2005). Migration and Environmental Hazards. *Population and Environment*, 26(4):273–302.

Ibrahim, M. R., Titheridge, H., Cheng, T., and Haworth, J. (2019). predictSLUMS: A new model for identifying and predicting informal settlements and slums in cities from street intersections using machine learning. *Computers, Environment and Urban Systems*, 76:31–56.

Jaber, F. and Abbad, M. (2023). A realistic evaluation of the dark side of data in the digital ecosystem. *Journal of Information Science*, page 01655515231205499.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Joseph, M., Wang, F., and Wang, L. (2014). Gis-based assessment of urban environmental quality in port-au-prince, haiti. *Habitat Int.*, 41:33–40.

Kamila, M. K. and Jasrotia, S. S. (2023). Ethical issues in the development of artificial intelligence: recognizing the risks. *International Journal of Ethics and Systems*, ahead-of-print(ahead-of-print).

Kim, B., Park, J., and Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, 134:113302.

Kit, O., Lüdeke, M., and Reckien, D. (2012). Texture-based identification of urban slums in Hyderabad, India using remote sensing data. *Applied Geography*, 32(2):660–667.

Kochupillai, M., Kahl, M., Schmitt, M., Taubenböck, H., and Zhu, X. X. (2022). Earth Observation and Artificial Intelligence: Understanding emerging ethical issues and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):90–124.

Kohli, D., Warwadekar, P., Kerle, N., Sliuzas, R., and Stein, A. (2013). Transferability of Object-Oriented Image Analysis Methods for Slum Identification. *Remote Sensing*, 5(9):4209–4228.

Kraff, N. J., Taubenböck, H., and Wurm, M. (2019). How dynamic are slums? EO-based assessment of Kibera's morphologic transformation. In *2019 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4. ISSN: 2642-9535.

Kraff, N. J., Wurm, M., and Taubenböck, H. (2020). Uncertainties of Human Perception in Visual Image Interpretation in Complex Urban Environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:4229–4241.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Kuffer, M., Abascal, A., Engstrom, R., Thomson, D. R., Tregonning, G., Shonowo, A., Zhao, Q., de Albuquerque, J. P., Elias, P., Onyambu, F. C., and Kabaria, C. (2024). IDEAMAPS: Modelling Sub-Domains of Deprivation with EO and AI. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 1562–1566. ISSN: 2153-7003.

Kuffer, M., Barros, J., and Sliuzas, R. V. (2014). The development of a morphological unplanned settlement index using very-high-resolution (VHR) imagery. *Computers, Environment and Urban Systems*, 48:138–152.

Kuffer, M., Orina, F., Sliuzas, R., and Taubenböck, H. (2017). Spatial patterns of slums: Comparing African and Asian cities. In *2017 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4.

Kuffer, M., Persello, C., Pfeffer, K., Sliuzas, R., and Rao, V. (2019). Do we underestimate the global slum population? In *2019 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4. ISSN: 2642-9535.

Kuffer, M., Pfeffer, K., and Sliuzas, R. (2016). Slums from Space—15 Years of Slum Mapping Using Remote Sensing. *Remote Sensing*, 8(6):455.

Kuffer, M., Sliuzas, R., Pfeffer, K., and Baud, I. (2015). The utility of the co-occurrence matrix to extract slum areas from vhr imagery. In *2015 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4.

Kühnl, M., Sapena, M., Wurm, M., Geiß, C., and Taubenböck, H. (2023). Multitemporal landslide exposure and vulnerability assessment in medellín, colombia. *Nat. Hazards*, 119(2):883–906.

Leao, S. and Leao, D. (2011). Targeting housing problems through urban texture analysis. In *Proceedings of the 12th International Conference on Computers in Urban Planning and Urban Management, Lake Louise, AB, Canada*, pages 5–8.

Lesiv, M., See, L., Laso Bayas, J. C., Sturn, T., Schepaschenko, D., Karner, M., Moorthy, I., McCallum, I., and Fritz, S. (2018). Characterizing the Spatial and Temporal Availability of Very High Resolution Satellite Imagery in Google Earth and Microsoft Bing Maps as a Source of Reference Data. *Land*, 7(4):118.

Li, J. and Liu, Z. (2019). Multispectral Transforms Using Convolution Neural Networks for Remote Sensing Multispectral Image Compression. *Remote Sensing*, 11(7):759.

Li, K., Wan, G., Cheng, G., Meng, L., and Han, J. (2020). Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307.

Li, W., Chen, H., and Shi, Z. (2021). Semantic Segmentation of Remote Sensing Images With Self-Supervised Multitask Representation Learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:6438–6450.

Lipton, M. (1980). Migration from rural areas of poor countries: The impact on rural productivity and income distribution. *World Development*, 8(1):1–24.

Liu, R., Kuffer, M., and Persello, C. (2019). The Temporal Dynamics of Slums Employing a CNN-Based Change Detection Approach. *Remote Sensing*, 11(23):2844.

Liu, X., Clarke, K., and Herold, M. (2006). Population Density and Image Texture. *Photogrammetric Engineering & Remote Sensing*, 72(2):187–196.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. Technical report. arXiv:1411.4038 [cs] type: article.

Lu, D., Hetrick, S., and Moran, E. (2010). Land Cover Classification in a Complex Urban-Rural Landscape with QuickBird Imagery. *Photogrammetric Engineering & Remote Sensing*, 76(10):1159–1168.

Lumban-Gaol, Y. A., Rizaldy, A., and Murtiyoso, A. (2023). Comparison of Deep Learning Architectures for the Semantic Segmentation of Slum Areas from Satellite Images. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume XLVIII-1/W2-2023, pages 1439–1444. Copernicus.

Lunga, D., Arndt, J., Gerrand, J., and Stewart, R. (2021). ReSFlow: A Remote Sensing Imagery Data-Flow for Improved Model Generalization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:10468–10483.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., and Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:166–177.

Mahabir, R., Croitoru, A., Crooks, A. T., Agouris, P., and Stefanidis, A. (2018). A Critical Review of High and Very High-Resolution Remote Sensing Approaches for Detecting and Mapping Slums: Trends, Challenges and Emerging Opportunities. *Urban Science*, 2(1):8.

Maurício, J., Domingues, I., and Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences*, 13(9):5521.

Maxwell, A. E., Bester, M. S., and Ramezan, C. A. (2022). Enhancing Reproducibility and Replicability in Remote Sensing Deep Learning Research and Practice. *Remote Sensing*, 14(22):5760.

Mboga, N., Persello, C., Bergado, J. R., and Stein, A. (2017). Detection of Informal Settlements from VHR Images Using Convolutional Neural Networks. *Remote Sensing*, 9(11):1106.

Melander, E. and Ãberg, M. (2007). The Threat of Violence and Forced Migration: Geographical Scope Trumps Intensity of Fighting. *Civil Wars*, 9(2):156–173.

Mishkin, D., Sergievskiy, N., and Matas, J. (2017). Systematic evaluation of convolution neural network advances on the Imagenet. *Computer Vision and Image Understanding*, 161:11–19.

Mou, L., Ghamisi, P., and Zhu, X. X. (2017). Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3639–3655.

Mou, L., Hua, Y., and Zhu, X. X. (2020). Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):7557–7569.

Müller, I., Taubenböck, H., Kuffer, M., and Wurm, M. (2020). Misperceptions of Predominant Slum Locations? Spatial Analysis of Slum Locations in Terms of Topography Based on Earth Observation Data. *Remote Sensing*, 12(15):2474.

Najmi, A., Gevaert, C. M., Kohli, D., Kuffer, M., and Pratomo, J. (2022). Integrating Remote Sensing and Street View Imagery for Mapping Slums. *ISPRS International Journal of Geo-Information*, 11(12):631.

Novack, T. and Kux, H. J. (2010). Urban land cover and land use classification of an informal settlement area using the open-source knowledge-based system InterIMAGE. *Journal of Spatial Science*, 55(1):23–41.

OPHI (2023). Global multidimensional poverty index. `https://ophi.org.uk/multidimensional-poverty-index/`. Accessed: 2024-08-14.

Owen, K. K. and Wong, D. W. (2013). An approach to differentiate informal settlements using spectral, texture, geomorphology and road accessibility metrics. *Applied Geography*, 38:107–118.

Owusu, M., Kuffer, M., Belgiu, M., Grippa, T., Lennert, M., Georganos, S., and Vanhuysse, S. (2021). Geo-Ethics in Slum Mapping. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 5700–5703. ISSN: 2153-7003.

Owusu, M., Nair, A., Jafari, A., Thomson, D., Kuffer, M., and Engstrom, R. (2024). Towards a scalable and transferable approach to map deprived areas using Sentinel-2 images and machine learning. *Computers, Environment and Urban Systems*, 109:102075.

Pan, W. and Du, J. (2021). Towards sustainable urban transition: A critical review of strategies and policies of urban village renewal in Shenzhen, China. *Land Use Policy*, 111:105744.

Persello, C. and Kuffer, M. (2020). Towards uncovering socio-economic inequalities using vhr satellite images and deep learning. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 3747–3750. IEEE.

Persello, C. and Stein, A. (2017). Deep fully convolutional networks for the detection of informal settlements in vhr images. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2325–2329.

Prabhu, R., Parvathavarthini, B., and Alaguraja, A. R. (2021). Integration of deep convolutional neural networks and mathematical morphology-based postclassification framework for urban slum mapping. *Journal of Applied Remote Sensing*, 15(1):014515.

Pradhan, B., Jebur, M. N., Shafri, H. Z. M., and Tehrany, M. S. (2016). Data Fusion Technique Using Wavelet Transform and Taguchi Methods for Automatic Landslide Detection From Airborne Laser Scanning Data and QuickBird Satellite Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1610–1622.

Qiu, C., Mou, L., Schmitt, M., and Zhu, X. X. (2019). Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 154:151–162.

Qiu, C., Schmitt, M., Geiß, C., Chen, T.-H. K., and Zhu, X. X. (2020). A framework for large-scale mapping of human settlement extent from Sentinel-2 images via fully convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 163:152–170.

Rehman, M. F. U., Aftab, I., Sultani, W., and Ali, M. (2022). Mapping temporary slums from satellite imagery using a semi-supervised approach. *IEEE Geosci. Remote. Sens. Lett.*, 19:1–5.

Reijneveld, S. A., Verheij, R. A., and Bakker, D. H. d. (2000). The impact of area deprivation on differences in health: does the choice of the geographical classification matter? *Journal of Epidemiology & Community Health*, 54(4):306–313.

Rhinane, H., Hilali, A., Berrada, A., and Hakdaoui, M. (2011). Detecting Slums from SPOT Data in Casablanca Morocco Using an Object Based Approach. *Journal of Geographic Information System*, 03(03):217.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Rußwurm, M. and Körner, M. (2018). Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders. *ISPRS International Journal of Geo-Information*, 7(4):129.

Rüther, H., Martine, H. M., and Mtalo, E. G. (2002). Application of snakes and dynamic programming optimisation technique in modeling of buildings in informal settlement areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, 56(4):269–282.

Schmitt, A., Sieg, T., Wurm, M., and Taubenböck, H. (2018). Investigation on the separability of slums by multi-aspect TerraSAR-X dual-co-polarized high resolution spotlight images based on the multi-scale evaluation of local distributions. *International Journal of Applied Earth Observation and Geoinformation*, 64:181–198.

Schuegraf, P., Stiller, D., Tian, J., Stark, T., Wurm, M., Taubenböck, H., and Bittner, K. (2024). Ai-based building instance segmentation in formal and informal settlements. In *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 1558–1561.

Sedova, B. and Kalkuhl, M. (2020). Who are the climate migrants and where do they go? Evidence from rural India. *World Development*, 129:104848.

Senecal, J. J., Sheppard, J. W., and Shaw, J. A. (2019). Efficient Convolutional Neural Networks for Multi-Spectral Image Classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. ISSN: 2161-4407.

Seoh, R. (2020). Qualitative Analysis of Monte Carlo Dropout. Technical report. arXiv:2007.01720 [cs, stat] type: article.

Shahtahmassebi, A. R., Li, C., Fan, Y., Wu, Y., lin, Y., Gan, M., Wang, K., Malik, A., and Blackburn, G. A. (2021). Remote sensing of urban green spaces: A review. *Urban Forestry & Urban Greening*, 57:126946.

Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. Technical report. arXiv:1409.1556 [cs] type: article.

Soille, P. and Pesaresi, M. (2002). Advances in mathematical morphology applied to geoscience and remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 40(9):2042–2055.

Song, L., Estes, A. B., and Estes, L. D. (2023). A super-ensemble approach to map land cover types with high resolution over data-sparse african savanna landscapes. *Int. J. Appl. Earth. Obs. Geoinf.*, 116:103152.

Stark, T. (2018). Using deep convolutional neural networks for the identification of informal settlements to improve a sustainable development in urban environments. Master's thesis, Technische Universität München.

Stark, T., Wurm, M., Debray, H., Zhu, X. X., and Taubenböck, H. (2024a). Uncertainty Aware Slum Mapping in 55 Heterogeneous Cities. Technical Report 4893813, Rochester, NY.

Stark, T., Wurm, M., Zhu, X. X., and Taubenböck, H. (2020). Satellite-Based Mapping of Urban Poverty With Transfer-Learned Slum Morphologies. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:5251–5263.

Stark, T., Wurm, M., Zhu, X. X., and Taubenböck, H. (2023). Detecting challenging urban environments using a few-shot meta-learning approach. In *2023 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4. ISSN: 2642-9535.

Stark, T., Wurm, M., Zhu, X. X., and Taubenböck, H. (2024b). Quantifying Uncertainty in Slum Detection: Advancing Transfer Learning With Limited Data in Noisy Urban Environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:4552–4565.

Stasolla, M. and Gamba, P. (2008). Spatial indexes for the extraction of formal and informal human settlements from high-resolution sar images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1(2):98–106.

Stewart, I. D. and Oke, T. R. (2012). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93(12):1879 – 1900.

Su, J. and Hu, Q. (2004). Fast residential area extraction algorithm in high resolution remote sensing image based on texture analysis. *Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Istanbul, Turkey*, pages 12–23.

Sulik, J. J. and Edwards, S. (2010). Feature extraction for Darfur: geospatial applications in the documentation of human rights abuses. *International Journal of Remote Sensing*, 31(10):2521–2533.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going Deeper With Convolutions. pages 1–9.

Tacoli, C., McGranahan, G., and Satterthwaite, D. (2015). *Urbanisation, rural-urban migration and urban poverty*. JSTOR.

Tan, M. and Le, Q. V. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Technical report. arXiv:1905.11946 [cs, stat] type: article.

Tao, C., Qi, J., Guo, M., Zhu, Q., and Li, H. (2023). Self-Supervised Remote Sensing Feature Learning: Learning Paradigms, Challenges, and Future Works. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–26.

Tarozzi, A. and Deaton, A. (2009). Using Census and Survey Data to Estimate Poverty and Inequality for Small Areas. *The Review of Economics and Statistics*, 91(4):773–792.

Taubenböck, H., Wurm, M., Esch, T., and Dech, S. (2015). *Globale urbanisierung*. Springer.

Taubenböck, H., Debray, H., Qiu, C., Schmitt, M., Wang, Y., and Zhu, X. X. (2020). Seven city types representing morphologic configurations of cities across the globe. *Cities*, 105:102814.

Taubenböck, H. and Kraff, N. J. (2014). The physical face of slums: a structural comparison of slums in Mumbai, India, based on remotely sensed data. *Journal of Housing and the Built Environment*, 29(1):15–38.

Taubenböck, H., Kraff, N. J., and Wurm, M. (2018). The morphology of the Arrival City - A global categorization based on literature surveys and remotely sensed data. *Applied Geography*, 92:150–167.

Taubenböck, H., Weigand, M., Esch, T., Staab, J., Wurm, M., Mast, J., and Dech, S. (2019). A new ranking of the world's largest cities—Do administrative units obscure morphological realities? *Remote Sensing of Environment*, 232:111353.

Temba, P., Nero, M. A., Botelho, L. M. R., and Lopes, M. E. C. (2015). Building vectorization inside a favela utilizing lidar spot elevation. In *Earth Observing Systems XX*, volume 9607, pages 32–39. SPIE.

Thomas, I., Tannier, C., and Frankhauser, P. (2008). Is there a link between fractal dimension and residential environment at a regional level? *Cybergeo: European Journal of Geography*.

Thomson, D. R., Kuffer, M., Boo, G., Hati, B., Grippa, T., Elsey, H., Linard, C., Mahabir, R., Kyobutungi, C., Maviti, J., Mwaniki, D., Ndugwa, R., Makau, J., Sliuzas, R., Cheruiyot, S., Nyambuga, K., Mboga, N., Kimani, N. W., de Albuquerque, J. P., and Kabaria, C. (2020). Need for an Integrated Deprived Area "Slum" Mapping System (IDEAMAPS) in Low- and Middle-Income Countries (LMICs). *Social Sciences*, 9(5):80.

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M. F., and Emery, W. J. (2009). Active Learning Methods for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232.

UN-Habitat (2020). Annual report.

UNDP (2023). Multidimensional poverty index. `https://hdr.undp.org/en/mpi`. Accessed: 2024-08-14.

United Nations (2000). 2015 millennium development goals report.

United Nations (2023a). Extreme poverty definition. `https://www.un.org/sustainabledevelopment/poverty/`. Accessed: 2024-08-14.

United Nations (2023b). The sustainable development report.

United Nations (2023c). Transforming our world: the 2030 agenda for sustainable development.

United Nations (2024). Better data, more inclusive development.

Valous, N. A., Sun, D.-W., Allen, P., and Mendoza, F. (2010). The use of lacunarity for visual texture characterization of pre-sliced cooked pork ham surface intensities. *Food Research International*, 43(1):387–395.

Verma, D., Jana, A., and Ramamritham, K. (2019). Transfer learning approach to map urban slums using high and medium resolution satellite imagery. *Habitat International*, 88:101981.

Wahbi, M., El Bakali, I., Ez-zahouani, B., Azmi, R., Moujahid, A., Zouiten, M., Alaoui, O. Y., Boulaassal, H., Maatouk, M., and El Kharki, O. (2023). A deep learning classification approach using high spatial satellite images for detection of built-up areas in rural zones: Case study of Souss-Massa region - Morocco. *Remote Sensing Applications: Society and Environment*, 29:100898.

Walker, R. (2023). *Poverty and the World Order: The Mirage of SDG 1.* Sustainability Matters. Agenda Publishing.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45.

Wang, J., Yang, M., Chen, Z., Lu, J., and Zhang, L. (2022a). An MLC and U-Net Integrated Method for Land Use/Land Cover Change Detection Based on Time Series NDVI-Composed Image from PlanetScope Satellite. *Water*, 14(21):3363.

Wang, X.-R., Hui, E. C.-M., and Sun, J.-X. (2017). Population migration, urbanization and housing prices: Evidence from the cities in china. *Habitat International*, 66:49–56.

Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., and Zhu, X. X. (2022b). Self-Supervised Learning in Remote Sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):213–247.

Weeks, J. R., Hill, A., Stow, D., Getis, A., and Fugate, D. (2007). Can we spot a neighborhood from the air? Defining neighborhood structure in Accra, Ghana. *GeoJournal*, 69(1):9–22.

Wei, H.-P., Deng, Y.-Y., Tang, F., Pan, X.-J., and Dong, W.-M. (2022). A Comparative Study of CNN- and Transformer-Based Visual Style Transfer. *Journal of Computer Science and Technology*, 37(3):601–614.

Weng, Q. and Quattrochi, D. A. (2006). *Urban Remote Sensing.* CRC Press.

Whang, S. E., Roh, Y., Song, H., and Lee, J.-G. (2023). Data collection and quality challenges in deep learning: a data-centric AI perspective. *The VLDB Journal*, 32(4):791–813.

WHO (2023). Health and poverty. `https://www.who.int/news-room/fact-sheets/detail/poverty-and-health`. Accessed: 2024-08-14.

World Bank (2023). International poverty line. `https://www.worldbank.org/en/topic/poverty/overview`. Accessed: 2024-08-14.

Wurm, M., Droin, A., Stark, T., Geiß, C., Sulzer, W., and Taubenböck, H. (2021). Deep Learning-Based Generation of Building Stock Data from Remote Sensing for Urban Heat Demand Modeling. *ISPRS International Journal of Geo-Information*, 10(1):23.

Wurm, M. and Taubenböck, H. (2018). Detecting social groups from space – Assessment of remote sensing-based mapped morphological slums using income data. *Remote Sensing Letters*, 9(1):41–50.

Wurm, M., Taubenböck, H., Weigand, M., and Schmitt, A. (2017). Slum mapping in polarimetric SAR data using spatial features. *Remote Sensing of Environment*, 194:190–204.

Xiong, Z., Wang, Y., Zhang, F., and Zhu, X. X. (2024). One for all: Toward unified foundation models for earth vision.

Yin, J., Dong, J., Hamm, N. A. S., Li, Z., Wang, J., Xing, H., and Fu, P. (2021). Integrating remote sensing and geospatial big data for urban land use mapping: A review. *International Journal of Applied Earth Observation and Geoinformation*, 103:102514.

Young, A. (2013). Inequality, the Urban-Rural Gap, and Migration. *The Quarterly Journal of Economics*, 128(4):1727–1785.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115.

Zhu, X. X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Häberle, M., Hua, Y., Huang, R., Hughes, L., Li, H., Sun, Y., Zhang, G., Han, S., Schmitt, M., and Wang, Y. (2019). So2Sat LCZ42: A Benchmark Dataset for Global Local Climate Zones Classification. Technical report. arXiv:1912.12171 [cs, eess] type: article.

Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F. (2017). Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36.

Zhu, X. X., Wang, Y., Kochupillai, M., Werner, M., Häberle, M., Hoffmann, E. J., Taubenböck, H., Tuia, D., Levering, A., Jacobs, N., Kruspe, A., and Abdulahhad, K. (2022). Geoinformation Harvesting From Social Media Data: A community remote sensing approach. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):150–180.

A. Appendix Publications

A.1. Semantic Segmentation of Slums in Satellite Images Using Transfer Learning on Fully Convolutional Neural Networks

Reference: Wurm, M., Stark, T., Zhu, X. X., Weigand, M., & Taubenböck, H. (2019). Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. ISPRS journal of photogrammetry and remote sensing, 150, 59-69.

ISPRS Journal of Photogrammetry and Remote Sensing 150 (2019) 59–69

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks

Michael Wurm[a,*], Thomas Stark[b], Xiao Xiang Zhu[b,c], Matthias Weigand[a,d], Hannes Taubenböck[a]

[a] German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Oberpfaffenhofen 82234, Germany
[b] Technical University of Munich (TUM), Signal Processing in Earth Observation (SiPEO), 80333 Munich, Germany
[c] German Aerospace Center (DLR), Remote Sensing Technology Institute (IMF), Oberpfaffenhofen 82234, Germany
[d] University of Würzburg, Department for Remote Sensing, 97074 Würzburg, Germany

ARTICLE INFO

Keywords:
Slums
FCN
Convolutional neural networks
Deep learning
Transfer learning

ABSTRACT

Unprecedented urbanization in particular in countries of the global south result in informal urban development processes, especially in mega cities. With an estimated 1 billion slum dwellers globally, the United Nations have made the fight against poverty the number one sustainable development goal. To provide better infrastructure and thus a better life to slum dwellers, detailed information on the spatial location and size of slums is of crucial importance. In the past, remote sensing has proven to be an extremely valuable and effective tool for mapping slums. The nature of used mapping approaches by machine learning, however, made it necessary to invest a lot of effort in training the models. Recent advances in deep learning allow for transferring trained fully convolutional networks (FCN) from one data set to another. Thus, in our study we aim at analyzing transfer learning capabilities of FCNs to slum mapping in various satellite images. A model trained on very high resolution optical satellite imagery from QuickBird is transferred to Sentinel-2 and TerraSAR-X data. While free-of-charge Sentinel-2 data is widely available, its comparably lower resolution makes slum mapping a challenging task. TerraSAR-X data on the other hand, has a higher resolution and is considered a powerful data source for intra-urban structure analysis. Due to the different image characteristics of SAR compared to optical data, however, transferring the model could not improve the performance of semantic segmentation but we observe very high accuracies for mapped slums in the optical data: QuickBird image obtains 86–88% (positive prediction value and sensitivity) and a significant increase for Sentinel-2 applying transfer learning can be observed (from 38 to 55% and from 79 to 85% for PPV and sensitivity, respectively). Using transfer learning proofs extremely valuable in retrieving information on small-scaled urban structures such as slum patches even in satellite images of decametric resolution.

1. Introduction

Poverty is considered one of the major challenges for our society in the upcoming decades, making it the number one issue of the Sustainable Development Goals as defined by the United Nations (UN, 2017). In urban areas, slums are the most visible, distinct manifestation of poverty (Amnesty International, 2016). Unprecedented processes of urbanization over the past decades have transformed mankind into an urban species with two thirds of the global population being expected to live in urban areas by the year 2050 (UN, 2015). This rural-urban migration is especially intense in mega cities of the global south, such as Mumbai in India which grew at a pace of up to 300,000 inhabitants per year (Burdett and Rhode, 2010). Since formal urban development cannot keep up with this pace of rural-urban migrants, many new urban dwellers are forced to find their new homes in settlements of informal nature with poor living conditions, lack of basic services such as access to safe water and sanitation facilities. Today, these *slums* are home to almost an estimated billion dwellers on a global scale (UN Habitat, 2015). In some cities, the share of slum dwellers accounts for up to 42% of the city's total population in official numbers (and a significantly higher number in estimations) such as it is the case for Mumbai (Taubenböck and Wurm, 2015). Various strategies for dealing with slums have been developed by local authorities, however a recent change can be observed towards a strategy of integrating the 'invisible city' into governing structures is today for many cities the accepted way to deal with those informal areas since the presence of slums cannot be neglected anymore (Wurm and Taubenböck, 2018). Thus, the derivation of reliable, spatial information on the size and location of slum

[*] Corresponding author.
E-mail address: michael.wurm@dlr.de (M. Wurm).

https://doi.org/10.1016/j.isprsjprs.2019.02.006
Received 25 May 2018; Received in revised form 10 December 2018; Accepted 11 February 2019
Available online 21 February 2019

areas by mapping approaches has gained much of interest over the past.

1.1. Morphological characteristics of slums from a remote sensing perspective

As it can be observed for many applications in the context of urban remote sensing, the turn of the millennium marks an important date with the advent of very high resolution satellites providing images at resolutions of 1 m or better. Especially for the discrimination of very small, heterogeneous objects such as buildings within the urban environment, high image resolutions are of crucial importance. Thus, also in the context of slum mapping, an increased interest in the utilization of VHR satellite images can be observed since then. This goes in parallel with the advent of more sophisticated image analysis techniques such as object-based image analysis or, recently, deep learning methods. Thus, in the following we review previous works on remote sensing-based slum mapping based on different methods and image features in the light of the complex nature of slum morphology.

From a synoptic perspective, urban poverty finds its physical expression in many different ways which usually do not follow a strict and universal concept (Taubenböck et al., 2018; Kuffer et al., 2017). However, some forms of urban poverty in particular can be directly associated with the morphology of the built environment, though (Sandborn and Engstrom, 2016; Jean et al., 2016; Wurm and Taubenböck, 2018). Most commonly, organic, irregular arrangements of buildings are associated with slum areas, as well as low building heights, poor construction materials and a generally high building density in often hazardously exposed areas (Baud et al., 2010; Kuffer et al., 2016a; Graesser et al., 2012; Jain, 2007). These characteristic morphologic features are extensively exploited in remote sensing-based image analysis for slum mapping. Since recently thorough studies on the state of slum mapping have been released (Kuffer et al., 2016a; Mahabir et al., 2018), we only summarize below past research efforts based on significant cornerstones in methods or data. While generally, very high mapping accuracies are achieved by visual image interpretation (Wurm and Taubenböck, 2018; Taubenböck et al., 2018) or knowledge-based methods using object-based image analysis (OBIA) relying on tuned parameters (Kuffer et al., 2014; Baud et al., 2010), large-area mapping of slums is usually based on machine learning algorithms which aim at generalizing specific semantic knowledge in the images based on labeled elements and image descriptors to provide transferability of the learned knowledge into unknown areas. One key feature in the identification of slums is their sharp contrast in their physical appearing compared to formal developed urban areas. Therefore, contextual image features such as the grey-level-co-occurrence-matrix (GLCM) was used extensively in slum mapping in combination with machine learning techniques such as random forests (e.g. Kuffer et al., 2016b; Graesser et al., 2012; Wurm et al., 2017; Owen and Wong, 2013) or support vector machines (Huang et al., 2015). Besides the extensive use of VHR optical data, only few studies were dedicated to the exploitation of actively acquired data, e.g. such as dual-polarized X-band SAR data from TerraSAR-X (Wurm et al., 2017; Schmitt et al., 2018). Only recently, the current trend in machine learning for semantic segmentation of images has been taken up by the application of deep learning for the detection of slums in VHR images confirming current trends in deep learning methods to outperform state-of-the-art machine learning techniques (Persello and Stein, 2017). The next subsequent step to learning and applying a network on the same data set is to transfer a pretrained network to sensors of different resolutions. Thus, deeper networks consisting of more hidden layers need to be considered (Oquab et al., 2014).

1.2. Transfer learning for semantic segmentation using convolutional neural networks

Generally, most machine learning methods work well because human-designed representations and features are used to optimize weights for an accurate prediction. Representation learning attempts to automatically learn good features or representations, which works well for small problems. In contrast, manually designed features are often over-specified, incomplete, and are very time-consuming for design and validation. Deep learning algorithms attempt to automatically learn multiple levels of representations exclusively from its input data, without the need of additional user input (Zhu et al., 2017). Besides its effectiveness, this can be regarded as one of the reasons for the big success of deep learning in machine learning since the task of training and prediction is facilitated. Recent advances in the field have proven deep learning a very successful set of tools, sometimes even able to surpass human ability to solve highly computational tasks (Zhu et al., 2017). Especially for image representations, convolutional neural networks have proven to excel at extracting mid- and high level abstract features from raw images. Recent studies indicate that the feature representations learned by CNNs are greatly effective in large scale image recognition (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014), object detection (Girshick et al., 2016) and semantic segmentation (Long et al., 2015).

Image segmentation aims at understanding an image at pixel level, i.e. each pixel of an image is assigned a semantic class. Initially, images of a fixed size were required for classification, but soon fully convolutional networks (FCNs) without fully connected layers popularized CNN architectures for dense predictions of images of any size and significantly increased speed (Long et al., 2015). Apart from fully connected layers, one of the main challenges using CNNs for semantic segmentation are the ‘pooling layers’. They increase the field of view and are able to aggregate the context while discarding the location information. However, semantic segmentation requires the exact alignment of class maps and thus, needs the spatial information to be preserved. This issue can be tackled by encoder-decoder architectures where an encoder gradually reduces the spatial dimension with pooling layers and a decoder which gradually recovers the object details and spatial dimension using transposed/fractionally strided convolutions. While FCNs can learn the interpolation during the decoding process, upsampling produces coarse segmentation maps because of loss of information during pooling. Therefore, skip connections are introduced from higher resolution feature maps.

In Long et al. (2015), the authors describe the key observation that fully connected layers in classification networks can be viewed as convolutions with kernels that cover their entire input regions. This is equivalent to evaluating the original classification network on overlapping input patches but is much more efficient because computation is shared over the overlapping regions of patches. In remote sensing, the use of deep learning brings up new challenges, since satellite image analysis raises some unique issues that need to be considered, e.g. geolocation of satellite images, sensor specifics (resolution, incidence angles, data quality etc.) or the big data challenge (Zhu et al., 2017).

In the context of remote sensing, scene classification of satellite images, which aims to automatically assign a semantic label to each pixel in an image, has recently been an active research topic in the field of VHR satellite images. Generally, scene classification can be divided into two steps: *feature extraction* and *classification*. With growing numbers of images, training a complicated non-linear classifier is very time consuming. Hence, to extract a holistic and discriminative feature representation is the most significant part for scene classification. Traditional approaches are mostly based on the Bag-of-Visual-Words model (Sivic and Zisserman, 2003; Zhu et al., 2016), but their potential for improvement was limited by the ability of experts to design the feature extractor and the expressive power encoded. In contrast, deep learning architectures have been successfully applied to the problem of scene classification of high-resolution satellite images outperforming state-of-the-art image classifiers (Zou et al., 2015; Penatti et al., 2015; Castelluccio et al., 2015; Mou et al., 2017).

As deep learning is a multi-layer feature learning architecture, it can learn more abstract and discriminative semantic features with growing depth. Thus, it has been shown that it can achieve far better classification performances compared to mid-level approaches (Zhu et al., 2017). Training of neural networks, is usually performed using pretrained networks on large image datasets, e.g., COCO (Lin et al., 2014), Pascal VOC (Everingham et al., 2010) or ImageNet (Deng et al., 2009) which in general reach impressive accuracies (Hu et al., 2015; Zou et al., 2015). Expanding the three channel input limitations of traditional deep learning algorithms (Kemker et al., 2018; Marmanis et al., 2018) use specific architectures to use elevation information and multispectral imagery to boost performance in semantic segmentation frameworks. Training networks from scratch is an extremely elaborate and time consuming method which is usually employed only if the data has completely different characteristics compared to internet images, for example hyperspectral images (Mou et al., 2018; Pan et al., 2018) or SAR (Gong et al., 2017; Hughes et al., 2018).

In general, transfer learning builds upon learned knowledge from one dataset to improve learning in another dataset. More specifically, it can be described as a method which aims to improve learning the target predictive function $f_T(\cdot)$ in the new target dataset D_T using the knowledge learned in the source dataset D_S. As described by Pan and Yang (2010), transfer learning can be divided into three categories: (a) In inductive transfer learning the target task is different from the source task, no matter if the source or target datasets are the same or not. In this case labeled data is required to induce the target learning task. (b) For transductive transfer learning both target and source learning tasks are the same while their datasets are different. In this situation no labeled data in the target dataset are required. Lastly, (c) unsupervised transfer learning is used when the target and source tasks are different, and no labeled data is available in both source and target datasets.

Selection of transfer learning strategies not only depends on the availability of existing labels in both source and target datasets and the similarity of the source and target dataset but also if weights learned in the source task can be adjusted or shared in the target task. Transfer learning can be achieved using multiple strategies. Multi-task learning has been used to improve object detection accuracy by transferring knowledge from one object class to another using a support vector machine's (SVM) discriminative training framework for HOG template models (Aytar and Zisserman, 2011) or using a hierarchical classification model that allows rare objects to borrow statistical strength from related objects (Salakhutdinov et al., 2011). Two multi-task classifiers are used to obtain a more robust classifier for object detection in videos (Ma et al., 2014). In hyperspectral remote sensing domain adaption technology can be applied to share knowledge between different geographical domains when using support vector machines (Sun et al., 2012) or random forest classifiers with transfer component analysis (Xia et al., 2017). Impressive results could be observed using unsupervised feature representation using pretrained CNNs for scene classification in very high resolution remote sensing imagery (e.g. Castelluccio et al., 2015; Hu et al., 2015). Inductive transfer learning enables to further improve the learning task where backpropagation successfully re-weights labeled data from natural image datasets, e.g. ImageNet to solve new problems in remote sensing datasets (e.g. Maggiori et al., 2017; Marmanis et al., 2016; Nogueira et al., 2017; Kang et al., 2018). Therefore, in this study inductive transfer learning of a FCN is used due to relative large labeled datasets where the fine tuning of weights during backpropagation aims to achieve best possible results.

1.3. Transferring deep features between various remote sensing data sets

In slum mapping, in particular approaches using remotely sensed data from satellite images with varying characteristics were used extensively for assessing image processing and analysis techniques (Kuffer et al., 2016a; Mahabir et al., 2018). Both scientific meta-studies state that while previous work on remote sensing-based slum mapping has acknowledged the advances of recent machine learning techniques for locating slums in satellite images, they lack transferability between various data sets. Costs for the large-area availability of very high resolution (VHR) optical satellite imagery at a geometric resolution of 1 m and below are a limiting factor and thus, multi-sensor approaches with data sets of varying origins are proposed.

In this study we want to address these identified issues by using state-of-the-art machine learning techniques from the family of convolutional neural networks (CNN) which need no tuning of parameters and have therefore better capabilities for transferring a trained network to another data set, as long as the training data set is sufficiently large and representative. Specifically, we want to explore the capabilities of this process of '*transfer learning*' to adopt a pretrained CNN from VHR optical Quickbird imagery to be applied to satellites with larger mapping areas but lower geometric resolution such as Sentinel-2. Further, in a second experiment we want to assess the capabilities of transfer learning from optical imagery to active SAR imagery from TerraSAR-X.

The remainder of this article is structured as follows: in the following Section 2 we present the methodological framework of fully convolutional networks (FCN), transfer learning for slum mapping and used data sets among the experimental set-up. In Section 3 we present the results and discussion of the performed experiments, while Section 4 concludes the paper.

2. Methods and experimental set-up

2.1. Method: The fully convolutional network FCN-VGG19

FCNs, first introduced by Long et al. (2015) allow for semantic segmentation to train end-to-end and pixel-to-pixel for the prediction of dense outputs from arbitrary sized input images. Learning and inference are performed on the entire image by dense feedforward computation and backpropagation. Within the network upsampling layers enable a pixelwise prediction and learning with subsampled pooling. For our experiments, we use the CNN based on the classification architecture VGG19 by the Visual Geometry Group of Oxford University (Simonyan and Zisserman, 2014). The CNN relies on rather small receptive fields of 3×3 pixels which are convolved with the input at every pixel. In this way a stack of two 3×3 convolutional layers has an effective receptive field of 5×5. Consequently, four layers have a 9×9 effective receptive field. This strategy has the advantage of incorporating four nonlinear rectification layers instead of a single one, making the decision function more discriminative. Furthermore, it decreases the number of parameters: $4(3^2C^2) = 36C^2$ produces less trainable weights than a single 9×9 convolutional layer: $9^2C^2 = 81C^2$.

To adapt the CNN-VGG19 architecture to an FCN some modifications are required: The final classification layer is discarded and replaced with a 1×1 convolution and with the channel dimension of the number of used classes. Further, deconvolutional layers are introduced for bilinear upsampling of the coarse outputs to pixel-dense outputs. In this case, upsampling through deconvolutional layers means using transpose convolutions. This operation simply reverses the forward and backward passes of the convolution. Upsampling is performed for end-to-end learning by backpropagation from a pixelwise loss (Long et al., 2015).

A graphical representation of the used FCN-VGG19 architecture is depicted in Fig. 1. It shows that the FCN uses skips, which combines the final prediction layer with lower level layers with finer strides. Fusing fine layers and coarse layers lets the model make local predictions that respect a global structure. The FCN fuses the upsampled output of the VGG19 network architecture with predictions computed on top of the third and fourth pooling layer.

2.2. Method: Transfer learning approach

Training the FCN was performed using an inductive transfer

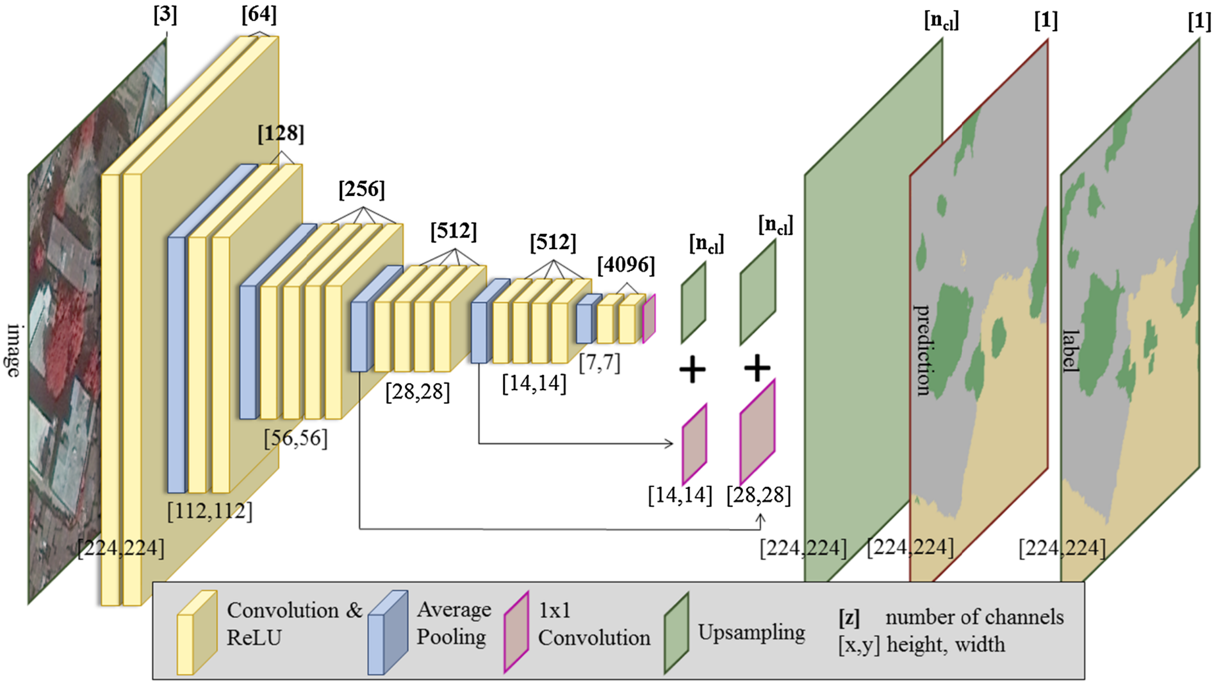

Fig. 1. Architecture of the FCN-VGG19 adapted from Long et al. (2015) which learns to combine high level information with fine, low level information using skips from the third and fourth pooling layer. Hidden layers are equipped with rectified linear units (ReLUs) and the number of channels for the convolutional layers increases with the depth of the network. During training the input image is a fixed size of 224 × 224 pixels, while receptive fields for all filters are 3 × 3 pixels throughout the whole network. This configuration allows the FCN to learn approximately 140 million parameters. Prediction is performed using upsampling layers with four channels for the all classes [n_{cl}] in the reference data. Upsampling layers are fused with 1 × 1 convolutions of the third and fourth pooling layers with the same channel dimension [x,y,n_{cl}]. The final upsampling layer predicts fine details using fused information from the last convolutional layer, third and fourth pooling layer upsampled at stride 8.

Table 1
Characteristics of satellite images for testing transfer learning techniques for the FCN-VGG19.

	GSD	Scene size	Bands/Polarization	Date	Incidence Angle	Image tiles
QuickBird	0.5 m	103 km^2	blue, green, red, nir	Nov 17, 2008	16.6°	7487
Sentinel-2	10 m	781 km^2	blue, green, red, nir	Nov 19, 2017	4.8°	219
TerraSAR-X	6 m	242 km^2	HH/VV	Sep 29, 2013	33.7°	2113
	6 m	242 km^2	VV/VH	Dec 11, 2013	33.7°	
	6 m	308 km^2	HH/VV	Oct 10, 2013	34.7°	
	6 m	308 km^2	VV/VH	Dec 04, 2013	34.7°	

learning approach (cf. Section 1.3). When given a source domain dataset D_S and a learning task T_S, a target domain dataset D_T and learning task T_T aims to improve the learning of the target predictive function $f_T(\cdot)$ in D_T using the knowledge in D_S and T_S where $T_S \neq T_T$ (Pan and Yang, 2010). In this case the target domain dataset D_T and the learning task $T_T$benefit from using the knowledge learned in the source domain dataset D_S. We present two groups of experiments. In our first approach weights from a vgg19 CNN which was pretrained on the ImageNet dataset are transfer learned for 100 epochs with all weights available for tuning during the backpropagation algorithm on all three remote sensing datasets where the source domain is the ImageNet dataset $D_S^{ImageNet}$ and the target domain is QuickBird, Sentinel-2 and TerraSAR-X imagery (FCN QB, FCN S2, FCN TX). Instance transfer allows to relabel weights from the source domain to the target domain and ensures adapting the backpropagation algorithm to improve the target learning task. Table 1 indicates a small dataset in the target domain for Sentinel-2 D_T^{S2} with only 219 image tiles and also in the TerraSAR-X target domain D_T^{TX} with only 2113 image tiles. A small target domain in $D_T^{S2,TX}$is usually insufficient for finding good feature representations between the source learning task $T_S^{ImageNet}$ and the target learning task $T_T^{S2,TX}$ for which reason a second group of transfer learning experiments was performed. It aims to reduce differences between the source and target domain where both domains are based on satellite images. Thus, the FCN trained on the QuickBird dataset (FCN QB) from the first group of experiments acts as a new source domain D_S^{QB} for the second group. The target learning task for Sentinel D_T^{S2} benefits from a better feature representation since both datasets D_S^{QB} and D_T^{S2} are optical remote sensing images. In the same way, the experiment is performed for the TerraSAR-X target domain D_T^{TX}. For both transfer learning experiments all trainable variables of the FCN are available during backpropagation to ensure adapting all parameters for the different resolutions and image sensing methods of the remote sensing data.

2.3. Material: Satellite images for slum mapping

For our experiments, space-borne satellite images of three different sensors (QuickBird, Sentinel-2, TerraSAR-X) with entirely different specifications are investigated. Since we aim at testing the capabilities of transfer learning of pretrained models between different images, we briefly introduce the used satellite images for our experiments below (Table 1). In general, our main image data set is from QuickBird. For transfer learning we use Sentinel-2 and TerraSAR-X.

QuickBird: was the first VHR commercial space-borne sensor with a sub-meter resolution of 0.5 m in the panchromatic band. The four

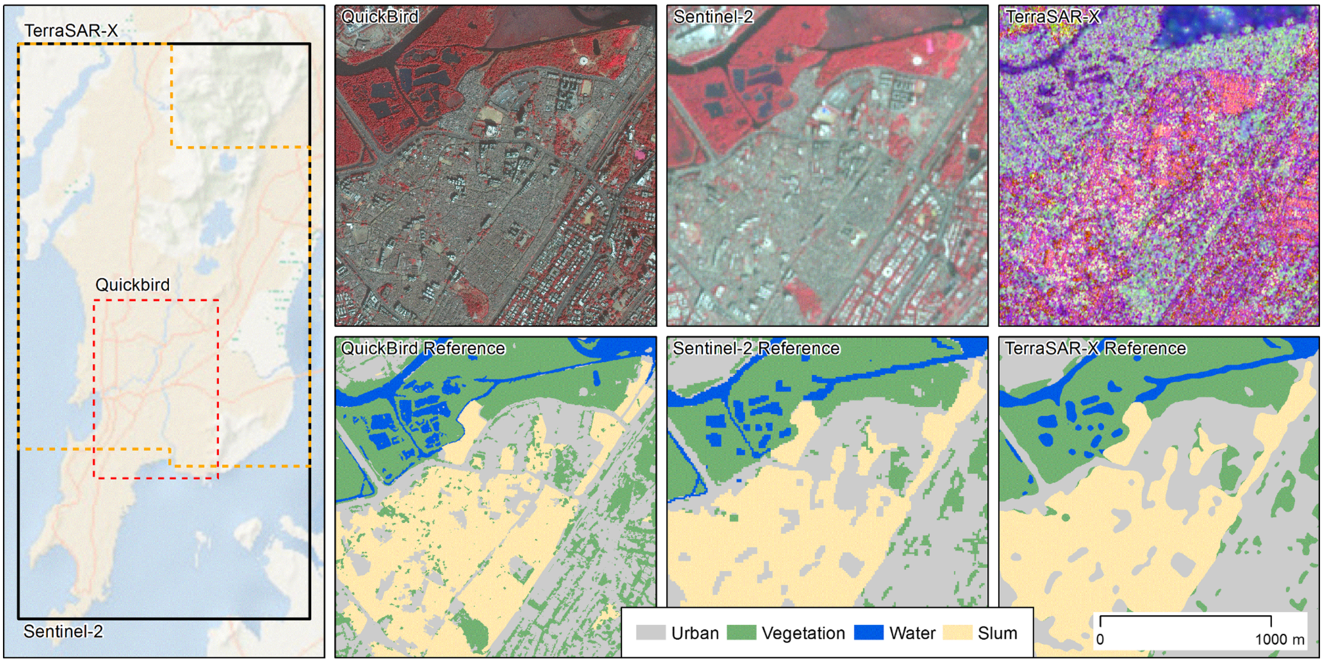

Fig. 2. Composites and reference labels for all datasets: QuickBird and Sentinel-2 in false color and TerraSAR-X as PCA composite for a subset of central Mumbai.

multispectral bands *blue*, *green*, *red* and *nir* are acquired at 2 m resolution. Scenes usually have a swath width of ∼17 km.

Sentinel-2: is the high resolution optical sensor of the European Copernicus Programme with 12 spectral and thermal bands at varying resolutions. The *blue*, *green*, *red* and *nir* bands are acquired at 10 m resolution. The swath width is 290 km.

TerraSAR-X: is an active SAR sensing system with various imaging modes of polarizations and resolution. For the commonly used stripmap mode, dual and cross polarized images are acquired at a ground sampling distance (GSD) of 6 m. The swath width is 11 km.

Satellite images are split into image tiles of 224 × 224 pixels with an overlap of 28 pixels to increase the amount of input data and to counter classification problems near edges. Since semantic segmentation performs classification of the entire images, four semantic classes are defined which cover the entire scenes: 'urban', 'vegetation', water' and 'slums'. For training and evaluation, fully labeled images are created for each data set (Fig. 2). Labeling of reference data is based on a multi-step image analysis procedure through a combination of hierarchical, knowledge-based and object-based classification, machine learning and visual image interpretation: in a first step, image objects are generated through a combined workflow of quad-tree and multiresolution image segmentation methods. Further, spectral and spatial image features are calculated for each image object and basic landcover classes such as water and vegetation are classified using a random forest classifier based on visually derived training objects. In a subsequent step, slum patches are derived by visual image interpretation from image analysts and cross-validated. The reference map is controlled by a stratified spatial random sample of 800 test points over the image with a resulting overall accuracy of 93% and a kappa value of 0.91. Accuracy for the slum class is reported with sensitivity of 92% and a positive prediction value of 95%. For the transfer learning experiments, the reference map was adapted to the geometric resolution of each target image data set.

2.4. Experimental set-up

The FCNs are trained on an Nvidia Titan X GPU using the 'adam optimizer' (Kingma and Ba, 2014) and a batch size of two image tiles. All FCNs use fixed learning rates of 10^{-5} and a dropout value of 15%. The training methodology for the FCNs was as following: *first*, a pretrained model is initially trained for 100 epochs on all three datasets (QuickBird, Sentinel-2 and TerraSAR-X) to set-up the FCN. *Second*, two transfer learning experiments are conducted: the pretrained QuickBird-FCN is transferred on Sentinel-2 and TerraSAR. The implementation of the FCN is based on the TensorFlow™ framework of Shekkizhar (2017).

Performance of the FCN is evaluated within a 4-fold cross validation procedure where each scene is split into four equal data strips. Out of the four data strips, three strips are used as training samples which are randomly shuffled after each epoch and the remaining strip is used for validation. The cross-validation process is repeated four times, with each of the four strips used exactly once for validation. Finally, the four results of the folds are mosaicked to produce a single output covering the entire scene with each strip being the result of one of the four classification experiments and thus allowing for assessment of independent results.

For quantitative assessment of the accuracy of the outputs of semantic segmentation, some commonly accepted performance measures are used: *First*, overall measures assess the general performance and *second*, class-specific measures reveal specific insights. The kappa index is applied as a measure to define to what extent the classification outcome differs from a random result with ranges between 0 and 1; where 0 corresponds to a completely random result and 1 corresponds to a completely nonrandom result. The overall accuracy (OA) and intersection over union (IoU; also known as Jaccard Index) are calculated in addition. OA is generated from an error matrix between the classification map and the reference map and allows for a general assessment of the agreement between the two maps; however, OA can be subject to a strong bias for very imbalanced semantic class distributions.

Class-specific accuracy measures are calculated to assess the proportion of correctly classified pixels from the reference (sensitivity) and the fraction of correctly classified pixels from the output (positive prediction value; PPV). These multiple standard measures are used for comparison with other classification experiments and are, much like OA, subject to well-known biases due to class-imbalance. Therefore, IoU is used in deep learning such as PASCAL VOC and CITYSCAPES challenge (Long et al., 2015). This accuracy measure compares the similarity between two maps and is calculated by the sum of true positives divided by the sum of true positives, false positives and false negatives

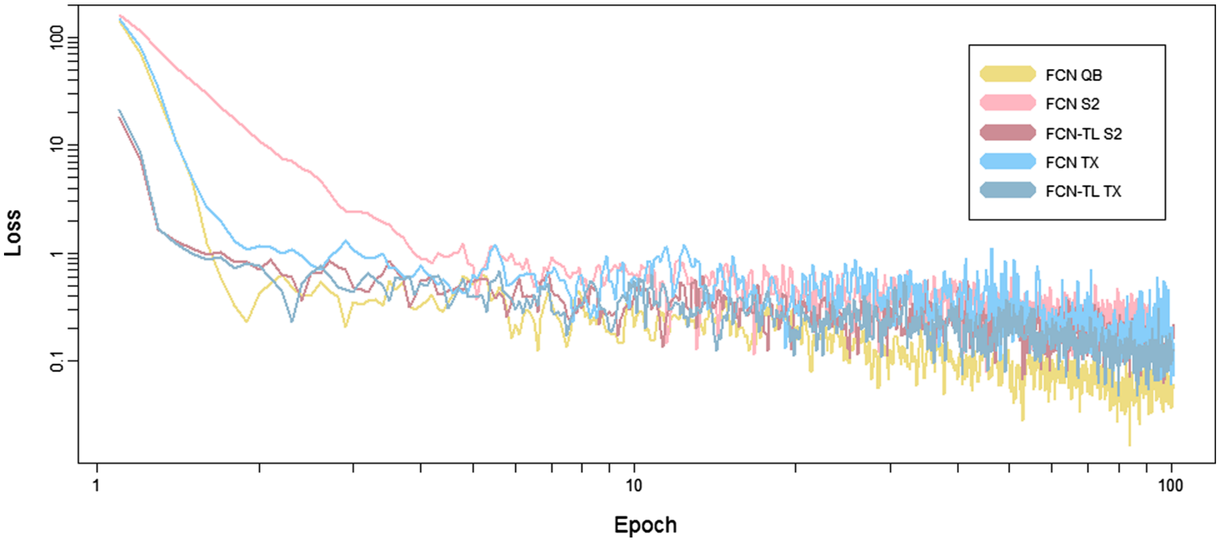

Fig. 3. Logarithmic learning curves for training five FCNs. The x-axis shows all FCNs trained for an equal duration of 100 epochs. The y-axis shows the cross entropy loss computed during training.

over the whole data set. It can be viewed as a precise indicator to the success of a classifier.

Besides the above introduced pixel-based performance evaluation strategies, a patch-based accuracy assessment is applied to account for a dependency of the slum patch area and the accuracy of the FCN. In this way, slum patch sizes are grouped into three size-based classes: smaller than 5 ha, 5–25 ha and larger than 25 ha. Accuracy assessment is performed for each slum patch size and analyzed (see Section 3.3).

3. Results and discussion

In this section, the capabilities of deep learning for slum mapping in different remotely sensed data sets with varying characteristics are analyzed subject to the quantitative results of the performed semantic segmentation experiments. Performance of the FCN is first evaluated for all four semantic classes in general and second for the slum class in particular. In total, five experiments were performed in two groups:

(1) training a pretrained model on the high resolution QuickBird image (*FCN QB*), on Sentinel-2 (*FCN S2*) and on TerraSAR-X (*FCN TX*).
(2) transfer learning of the pretrained FCN on Quickbird to Sentinel-2 (*FCN-TL S2*) and TerraSAR-X (*FCN-TL TX*).

Training the FCN is performed using a sparse softmax cross entropy loss function within TensorFlow™ to measure the performance of the model. The loss is a summation of the errors made for each example during the training stage, which implies how well or poorly a certain model behaves after each iteration of optimization. The respective loss curves are presented in Fig. 3 where all five FCNs indicate an interpretation on how well the model performs for the training datasets. All networks show convergence towards zero with some minimal jitter between 0.01 and 0.5. Both transfer learned FCNs (*FCN-TL S2 and FCN-TL TX*) reach a low loss value much faster than the pretrained FCNs, while the FCN trained on Sentinel-2 data takes considerably longer to converge against zero.

Semantic segmentation based on the FCN is performed on all total scenes (cf. extents in Fig. 2) according to the above described experimental set-up (cf. Section 2.4). A graphical depiction of the results for the same subset of a central area in Mumbai is depicted in Fig. 4. Visual interpretation of the results indicates very fine-structured patches for *FCN QB* as it is also the case in the reference data set. For that reason high accuracies are to be expected for the QB data set. As regards with the Sentinel-2 data, the effects of transfer learning become clearly visible: from large-structured patches of the results for *FCN S2*, a major increase in granularity using the transfer learning approach *FCN-TL S2* is observed: even at a geometric resolution of 10 m, small fractions of vegetation and slum patches are successfully detected. For TerraSAR-X (*FCN TX*), no significant alteration of the classification result is observed through transfer learning.

3.1. Overall accuracies

Quantitative results in terms of overall performance for the semantic segmentation are presented in Table 2 for all five experiments. With regards to overall measures, all five experiments obtained considerable accuracies with Kappa values between 0.72 and 0.85. The best performing set-up is reported, as expected, for QuickBird (*FCN-QB*). The Kappa value (0.85) and the Overall Accuracy (90.62%) show a very high agreement. This is followed by the Sentinel-2 experiment (*FCN-TL S2*) with the same Kappa value (0.85) and marginally lower OA (89.64%). Interestingly, highest IoU (87.43%) is reported for Sentinel-2 (*FCN–TL S2*) which can be considered as being mostly related to the substantially larger area of interest for Sentinel-2 (cf. Fig. 2) and the respectively larger shares of water bodies (cf. Table 3) which impact significantly the overall measures in general and the IoU in particular.

Transfer learning from the ImageNet domain $D_S^{ImageNet}$ to the remote sensing domains $D_T^{QB,S2,TX}$ performs well for the QuickBird learning task. This can be accounted for by a sufficient quantity of training data in D_T^{QB} (cf. Table 1). The second transfer task $D_T^{S2,TX}$ with less training data performs significantly poorer. Two possible reasons can explain this aspect: for the Sentinel-2 target learning task there is just not enough data available for a good knowledge transfer from $D_S^{ImageNet}$ to D_T^{S2}. The same accounts the for transfer learning task to TerraSAR-X data including another difficulty of a stark difference in feature representation of optical image data in $D_S^{ImageNet}$ and radar data in D_T^{TX}.

As regards with the performance of transfer learning against the performance of pre-trained networks, we observe remarkable differences among the transfer between QB/S2 and QB/TX: the transfer learning approach could significantly increase all overall performance measures for S2; however, no relevant change in accuracy is observed for the transfer between QuickBird and TerraSAR-X data. In fact, accuracy is even marginally lower for the transfer learning approach in this particular setting. We interpret this effect by difficulties of the network in transferring the learned model from optical features to SAR image features (cf. Hughes et al., 2018). Thus, no additional improvement of the model can be achieved.

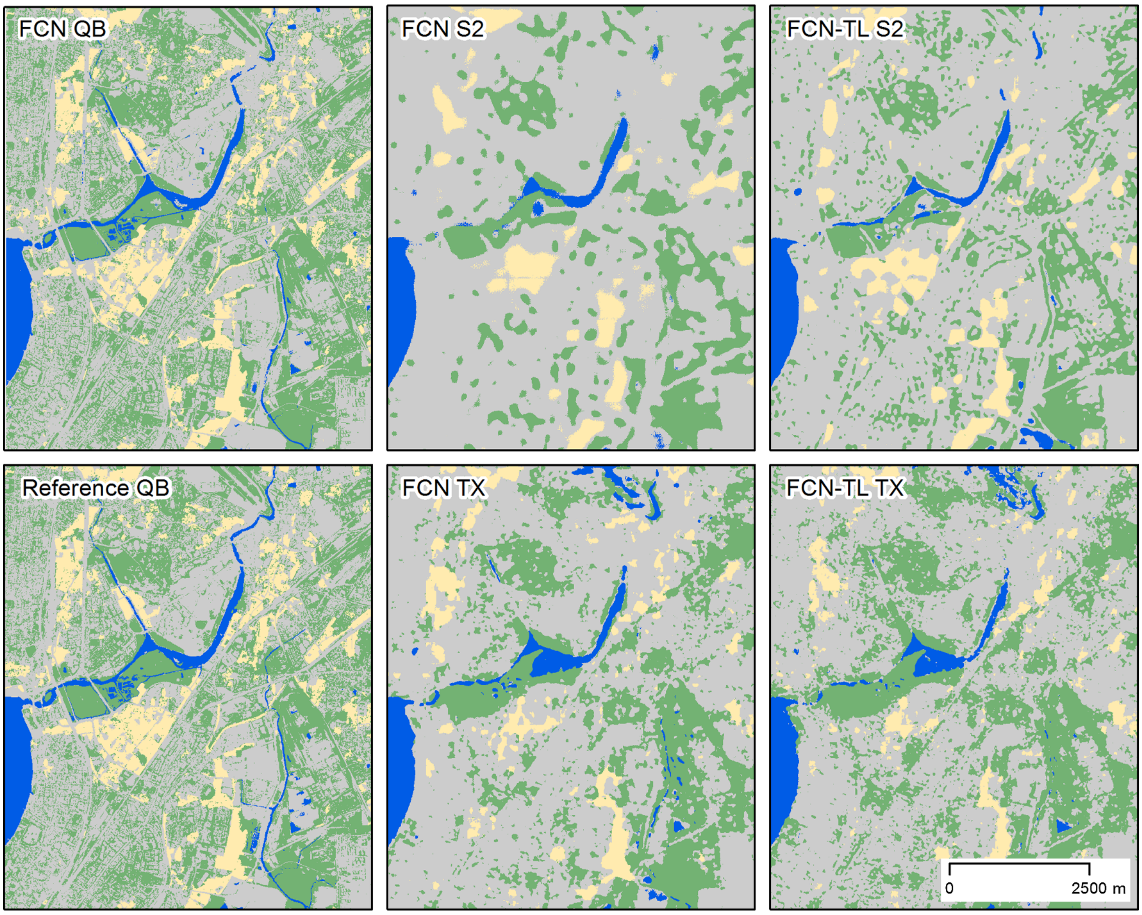

Fig. 4. Results of the semantic segmentation for the five experiments on the three data sets: QuickBird [QB], Sentinel-2 [S2] and TerraSAR-X [TX] on pre-trained FCNs and transfer learned FCNs [FCN-TL].

Table 2
Performance Evaluation of the FCN For all Classes. OA: Overall Accuracy; IoU: Intersection over Union; TL: Transfer Learned.

Approach	Kappa	OA (%)	IoU (%)
FCN-QB	**0.85**	**90.62**	84.12
FCN-S2	0.81	86.71	83.94
FCN-TL S2	**0.85**	89.64	**87.43**
FCN-TX	0.73	80.68	73.96
FCN-TL TX	0.72	80.03	73.02

Transfer learning from the QuickBird domain D_S^{QB} to the Sentinel-2 domain D_T^{S2} improves performance for all accuracy measurements significantly due to the similar feature representation in both the source and the target domain. Performance when using transfer learning techniques from the QuickBird domain D_S^{QB} to TerraSAR-X 2 domain D_T^{TX} stagnates or decreases to about 1–2% in the accuracy measurements. Prior studies have already pointed out this observation when dealing with SAR data (Zhu et al., 2017). We can confirm these issues where the upper limit of SAR classification accuracy is reached when only 2113 image tiles are available. The knowledge transfer is too difficult when transfer learning from either ImageNet or QuickBird to SAR data due to the significantly different image information representation

3.2. Class-based accuracies

While overall performance measures allow for a general assessment of the conducted experiments, detailed interpretation of class-based performance evaluation shed more light on the segmentation results. Thus, class-specific performance measures are presented in Table 3. With respect to the individual semantic classes, we observe the following: by far the highest accuracies in all performance measures for the classes ‘urban’ and ‘slum’ are obtained by QuickBird (*FCN-QB*). For

Table 3
Performance Evaluation of the FCN for the Individual Semantic Classes for the total scenes. IoU: Intersection over Union; TL: Transfer Learned; PPV: Positive Prediction Value; Sens: Sensitivity; A: area (percentage of scene coverage). Best results are marked in bold.

	Urban			Vegetation			Water			Slum		
Approach	Sens (%)	PPV (%)	IoU (%)	Sens (%)	PPV (%)	IoU (%)	Sens (%)	PPV (%)	IoU (%)	Sens (%)	PPV (%)	IoU (%)
FCN-QB	**91.37**	**90.34**	**83.24**	92.90	95.35	88.88	**90.78**	**90.97**	**83.28**	**85.70**	**88.39**	**77.02**
FCN-S2	87.47	75.87	68.43	96.42	98.44	94.97	85.35	89.72	77.75	38.21	78.82	35.51
FCN-TL S2	87.62	82.00	73.49	**97.47**	**98.57**	**96.12**	90.14	90.61	82.44	55.47	85.25	51.23
FCN-TX	84.29	83.13	71.99	93.86	94.03	88.59	78.46	75.65	62.63	51.64	72.50	46.27
FCN-TL TX	85.78	80.21	70.80	93.49	93.58	87.85	75.82	75.64	60.94	43.64	78.43	38.42

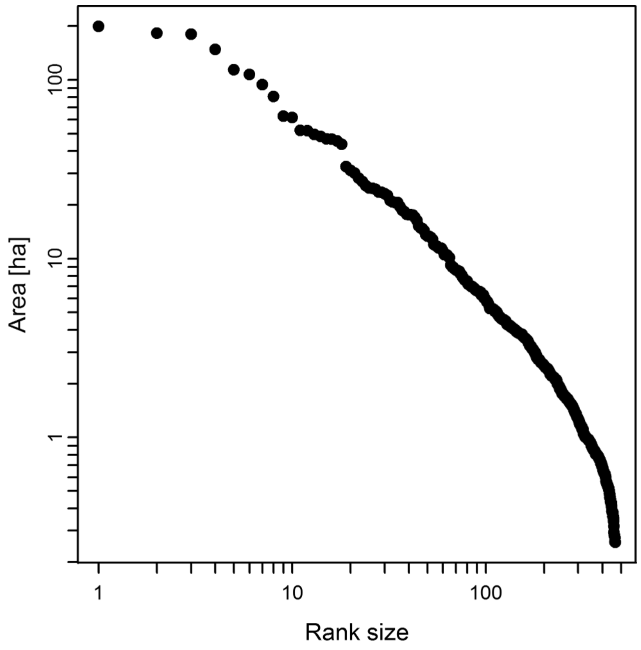

Fig. 5. Rank size distribution of slum patch sizes in Mumbai in a loglog plot.

Table 4
Proportions of number of slum patches and area for three size-based classes.

	Small slums [< 5 ha]	Medium slums [5–25 ha]	Large slums [> 25 ha]
Patches	84.63%	13.62%	1.75%
Area	26.10%	36.40%	37.50%

Table 5
Sensitivity measurement as a function of varying slum patch size.

Approach	Small slums [< 5 ha]	Medium slums [5–25 ha]	Large slums [> 25 ha]
FCN-QB	78.57	83.63	88.39
FCN-S2	09.32	28.19	47.18
FCN-TL S2	24.67	50.64	62.46
FCN-TX	31.26	47.36	55.34
FCN-TL TX	20.78	37.98	48.36

the ‘vegetation’ class, Sentinel-2 (*FCN-TL S2*) obtained best results, most likely due to the aggregation of information in Sentinel-2 and the consequential less small-structured vegetation fraction. Accuracies for the water class are quite similar between QuickBird (*FCN-QB*) and Sentinel-2 (*FCN-TL S2*) with only marginal differences. While for the urban class, QuickBird (*FCN-QB*) performs considerably better than Sentinel-2 (*FCN-TL S2*). The effect for the ‘slum’ class is most striking: the small-scaled buildings and their very organic arrangements are best segmented by the sensor with the highest geometric resolution being also capable of identifying individual buildings or shacks. Both, positive prediction value (88.4%) as well as sensitivity (85.7%) reach very high accuracies, i.e. the majority of slum areas as classified in the reference data set could be detected and only very few false positives occur. These effects are underpinned by high very IoU values (77%) which can be seen a very conservative measure of accuracy.

Comparing the results for pretraining and transfer learning, we observe a significant gain in accuracy in all semantic classes for Sentinel-2. Especially the performance of slums is increased remarkably making the effect of transfer learning in this case extremely valuable. As already reported in literature (Hughes et al., 2018), no positive effect is observed for TerraSAR-X data. Here, almost all classes are better represented by the pretraining approach (*FCN-TX*) than the transfer learning approach (*FCN-TL TX);* however, with one exception: PPV of slums is increased. If considering only the slum class, however, very competitive results in comparison to Sentinel-2 are obtained (55.47% vs. 51.64%).

All in all, we can state the following:

(1) the pretrained network on QuickBird performs very well in classifying heterogeneous urban environments.
(2) transfer learning for Sentinel-2 can significantly improve the results.
(3) for TerraSAR-X performance is reported lower than for the optical data.
(4) Transfer learning for TerraSAR-X could not improve the performance.

3.3. *The impact of slum patch size*

As stressed already in prior studies slum patch sizes vary significantly within cities (e.g. Wurm et al., 2017). Friesen et al. (2018) found that slum patch size distribution in several mega cities in the world follow very closely Zipf’s law and can be analyzed via rank size distribution (Zipf, 1941). The case for Mumbai is presented in Fig. 5. We observe a majority of small slums with areas below 5 ha and only a handful of large slums above 25 ha. Their respective contribution to the total slum area is, however, inverse, as presented in Table 4.

Based on these observations, we additionally perform a patch size-based accuracy assessment for the specific class of ‘slums’ to analyze the impact of slum patch size on the resulting classification performance. Both, a visual comparison for all approaches, and a quantitative assessment of sensitivity are conducted (Table 5). Small slum patches (< 5 ha) are presented in Fig. 6 with very good slum mapping capabilities for QuickBird (*FCN-QB:* 78.57%). Further, a significant increase of sensitivity for Sentinel-2 between pretrained and transfer learned is observed (9.32 vs. 24.67%). Prior discussed effects for TerraSAR-X images are also observed for the smallest group of patches: decreasing sensitivity between pretrained and transfer learned (31.26 vs. 20.78%). Both, Sentinel-2 and TerraSAR-X, however, perform very poor for this smallest group of patch sizes which is to be expected at image resolutions of 10 m and 6 m, respectively.

Medium-sized slum patches are presented in Fig. 7. Here the same trend is identified as for small patches: highest sensitivity is obtained by QuickBird (*FCN-QB*: 83.63%) and transfer learning significantly improves slum patch detection for Sentinel-2 against pre-training (28.19 vs. 50.64%). Again, a decrease is measured for the approach using TerraSAR-X (47.36 vs. 37.98%).

Finally, results for large slum patches (Fig. 8) are reported highest for all performed experiments. In QuickBird 88.39% of the reference slum pixels are detected (*FCN-QB*). For Sentinel-2, again, transfer learning significantly enhances mapping capabilities (47.18 vs. 62.46%) and a decrease in a performance is observed for TerraSAR-X (48.36 vs. 55.34%). Summarizing these observations, a strong effect of slum patch size on the detection rate is reported for all experiments (cf. Wurm et al., 2017).

4. Conclusion

In this paper, we perform a series of experiments to analyze the capabilities of fully convolutional neural networks for semantic segmentation of slums for the example of Megacity Mumbai using satellite images with different characteristics. As a result, we observe the following effects:

(1) very high geometric resolution of 0.5 m in QuickBird imagery allows for the best results of all experiments.
(2) transfer learning of a pre-trained network from QuickBird to

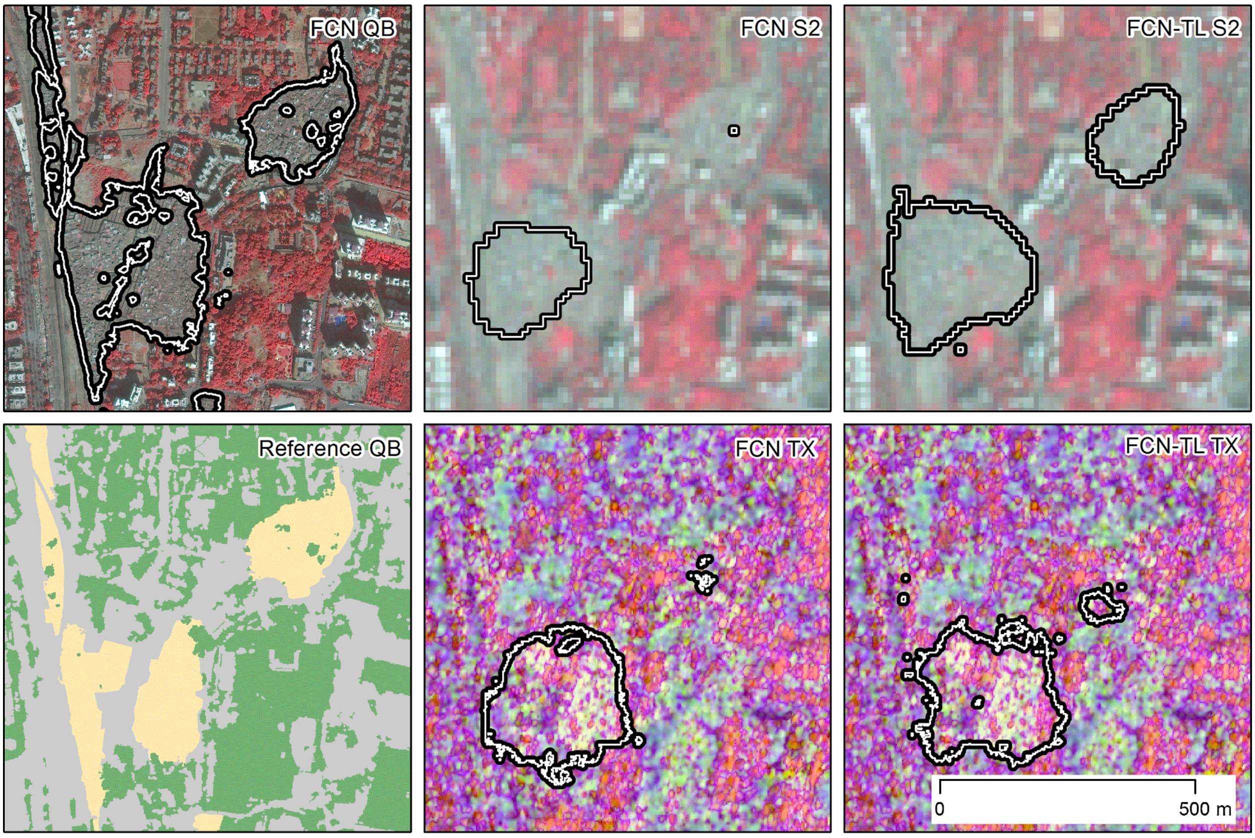

Fig. 6. Comparative alignment of small slum patches [< 5 ha] showing differences in segmentation results obtained by pre-trained FCNs and transfer learned FCNs (FCN-TL) on QuickBird, Sentinel-2 and TerraSAR-X images. Slum patches in the reference map are depicted in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

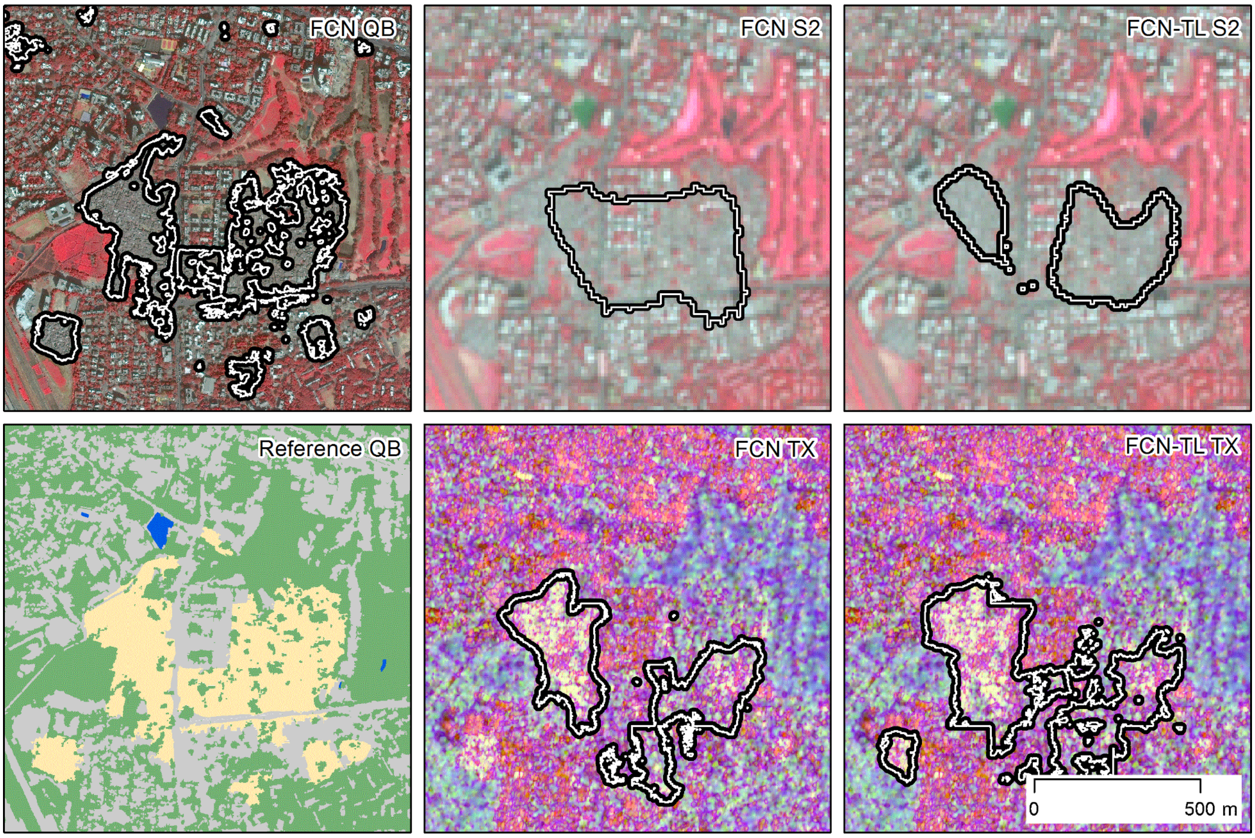

Fig. 7. Comparative alignment of medium sized slums [5 ha–25 ha] showing differences in segmentation results obtained by pre-trained FCNs and transfer learned FCNs (FCN-TL) on QuickBird, Sentinel-2 and TerraSAR-X images. Slum patches in the reference map are depicted in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

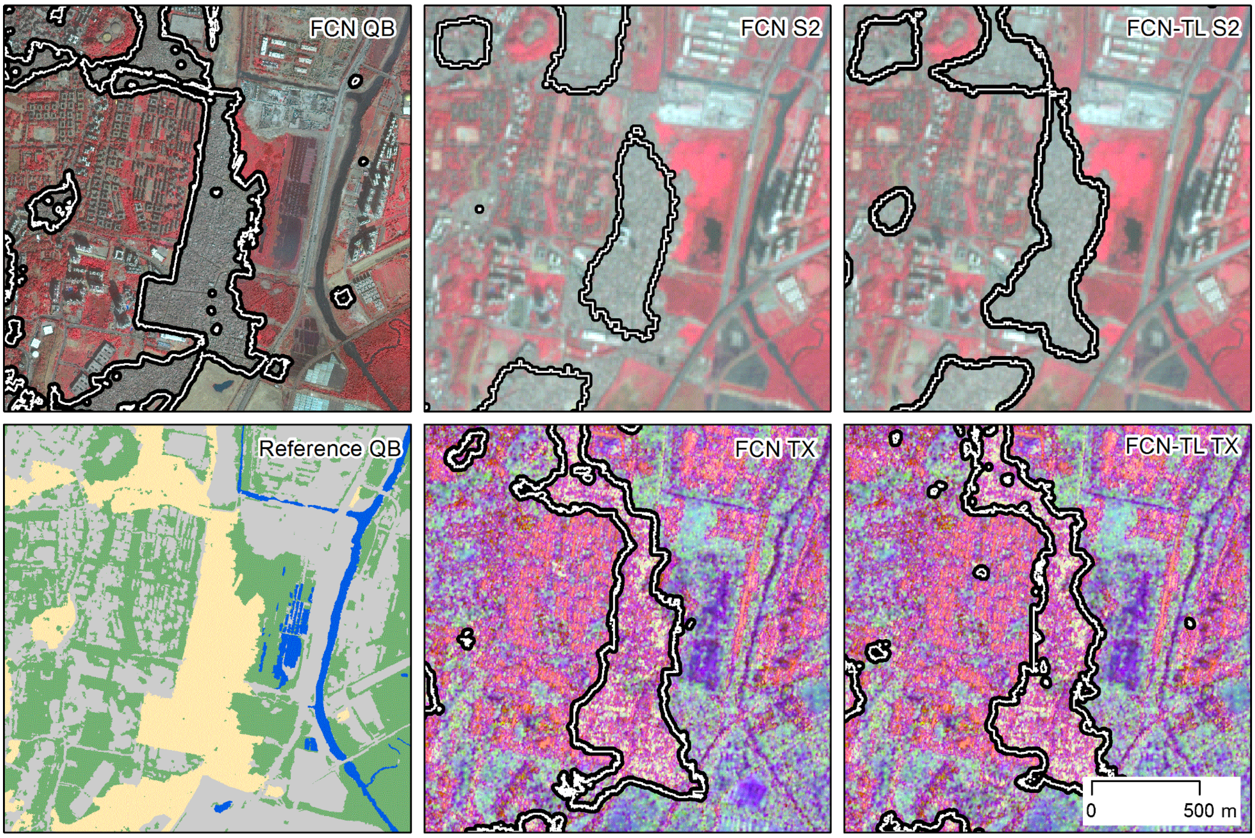

Fig. 8. Comparative alignment of a large slum patch [> 25 ha] showing differences in segmentation results obtained by pretrained FCNs and transfer learned FCNs (FCN-TL) on QuickBird, Sentinel-2 and TerraSAR-X images. Slum patches in the reference map are depicted in yellow. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Sentinel-2 images significantly improves the segmentation results. This makes medium resolution sensors at 10 m GSD an opportunity for very large-area mapping of slums for entire countries or sub-continents.

(3) for active satellite imagery such as TerraSAR-X, the transfer learning approach does not improve the results, but even decrease the performance. We relate this observation to the fact that the network is not able to transfer the learned image features from optical imagery to the SAR representation of urban structures.

(4) Further, we observe a strong effect of slum patch size for being detected by the segmentation approaches. While this effect is smallest for high resolution QuickBird imagery which already performs at a very high level: from 79.57% for < 5 ha to 88.39% for > 25 ha, an increase from 9.32 to 47.18% in sensitivity is obtained for Sentinel-2 pretrained (*FCN-S2*) and from 24.67 to 62.46% for Sentinel-2 transfer learned (*FCN-TL S2*). The same effect is also observed for TerraSAR-X: from 31.26 to 55.34% for pre-trained (*FCN-TX*) and 20.78 to 48.36% for transfer learned, respectively (*FCN-TL TX*).

Finally, segmentation outcomes are extremely promising and encouraging for further experiments using transfer learning and fully convolutional networks for slum mapping in satellite imagery. Further experiments need to focus on large-area approaches and the transfer between different geographical regions. This challenging task needs to address the morphological representations of slums in different cultural areas as shown by Taubenböck et al. (2018), since the physical nature of slums is represented by a large variety of morphological structures.

Funding

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No [ERC-2016-StG-714087], Acronym: So2Sat).

References

Amnesty International, 2016. Eine Milliarde Menschen Leben in Slums. https://www.amnesty.de/mit-menschenrechten- gegen-armut/wohnen-wuerde/eine-milliarde-menschen-weltweit-leben-slums.

Aytar, Y., Zisserman, A., 2011. Tabula rasa: Model transfer for object category detection. In: IEEE International Conference on Computer Vision (ICCV), pp. 2252–2259.

Baud, I., Kuffer, M., Pfeffer, K., Sliuzas, R.V., Karuppannan, S., 2010. Understanding heterogeneity in metropolitan India: The added value of remote sensing data for analyzing sub-standard residential areas. Int. J. Appl. Earth Obs. Geoinf. 12, 359–374.

Burdett, R., Rhode, P., 2010. Living in the urban age. In: Living in the Endless City. Phaidon, pp. 8–43.

Castelluccio, M., Poggi, G., Sansone, C., Verdoliva., L., 2015. Land use classification in remote sensing images by convolutional neural networks. arXiv preprint arXiv:1508.00092.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. IEEE Computer Vision and Pattern Recognition (CVPR).

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88 (2), 303–338.

Friesen, J., Taubenböck, H., Wurm, M., Pelz, P.F., 2018. The similar size of slums. Habitat Int. 73, 79–88.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2016. Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 38 (1), 142–158.

Gong, M., Yang, H., Zhang, P., 2017. Feature learning and change feature classification based on deep learning for ternary change detection in SAR images. ISPRS J. Photogramm. Remote Sens. 129, 212–225. https://doi.org/10.1016/j.isprsjprs.2017.05.001.

Graesser, J., Cheriyadat, A., Vatsavai, R.R., Chandola, V., Long, J., Bright, E., 2012. Image based characterization of formal and informal neighborhoods in an urban landscape. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5 (4), 1164–1176. https://doi.org/10.1109/JSTARS.2012.2190383.

Hu, F., Xia, G.S., Hu, J., Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 7 (11), 14680–14707.

Huang, X., Liu, H., Zhang, L., 2015. Spatiotemporal detection and analysis of urban villages in mega city regions of China using high-resolution remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. 53, 3639–3657.

Hughes, L., Schmitt, M., Mou, L., Wang, Y., Zhu, X., 2018. Identifying corresponding patches in SAR and optical images with a pseudo-siamese CNN. IEEE Geosci. Remote Sens. Lett. 15 (5), 784–788.

Jain, S., 2007. Use of IKONOS satellite data to identify informal settlements in Dehradun, India. Int. J. Remote Sens. 28, 3227–3233.

Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S., 2016. Combining satellite imagery and machine learning to predict poverty. Science 353 (6301), 790–794. https://doi.org/10.1126/science.aaf7894.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X., 2018. Building instance classification using street view images. ISPRS J. Photogramm. Remote Sens. 145 (Part A), 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006.

Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS J. Photogramm. Remote Sens. 145, 60–77. https://doi.org/10.1016/j.isprsjprs.2018.04.014.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kuffer, M., Barros, J., Sliuzas, R., 2014. The development of amorphological unplanned settlement index using very-high-resolution (VHR) imagery. Comput. Environ. Urban. Syst. 48, 138–152. https://doi.org/10.1016/j.compenvurbsys.2014.07.012.

Kuffer, M., Pfeffer, K., Sliuzas, R., 2016a. Slums from space – 15 years of slum mapping using remote sensing. Remote Sens. 8 (6), 455. https://doi.org/10.3390/rs8060455.

Kuffer, M., Pfeffer, K., Sliuzas, R., Baud, I., 2016b. Extraction of slum areas from VHR imagery using GLCM variance. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (5), 1830–1840. https://doi.org/10.1109/JSTARS.2016.2538563.

Kuffer, M., Pfeffer, K., Sliuzas, R., Baud, I., van Maarseveen, M., 2017. Capturing the Diversity of deprived areas with image-based features: the case of Mumbai. Remote Sens. 9 (4), 384. https://doi.org/10.3390/rs9040384.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: common objects in context. In: European Conference on Computer Vision. Springer, Cham, pp. 740–755.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Ma, Z., Yang, Y., Nie, F., Sebe, N., Yan, S., Hauptmann, A.G., 2014. Harnessing lab knowledge for real-world action recognition. Int. J. Comput. Vision 109 (1–2), 60–73.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 55 (2), 645–657.

Mahabir, R., Croitoru, A., Crooks, A.T., Agouris, P., Stefanidis, A., 2018. A Critical review of high and very high-resolution remote sensing approaches for detecting and mapping slums: trends, challenges and emerging opportunities. Urban Sci. 2 (1), 8. https://doi.org/10.3390/urbansci2010008.

Marmanis, D., Datcu, M., Esch, T., Stilla, U., 2016. Deep learning earth observation classification using ImageNet pretrained networks. IEEE Geosci. Remote Sens. Lett. 13 (1), 105–109.

Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS J. Photogramm. Remote Sens. 135, 158–172. https://doi.org/10.1016/j.isprsjprs.2017.11.009.

Mou, L., Ghamisi, P., Zhu, X.X., 2018. Unsupervised spectral–spatial feature learning via deep residual Conv–Deconv network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 56 (1), 391–406.

Mou, L., Zhu, X., Vakalopoulou, M., Karantzalos, K., Paragios, N., Le Saux, B., Moser, G., Tuia, D., 2017. Multi-temporal very high resolution from space: Outcome of the 2016 IEEE GRSS Data Fusion Contest. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10 (8), 3435–3447.

Nogueira, K., Penatti, O.A., dos Santos, J.A., 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recogn. 61, 539–556.

Owen, K.K., Wong, D.W., 2013. An approach to differentiate informal settlements using spectral, texture, geomorphology and road accessibility metrics. Appl. Geogr. 38, 107–118. https://doi.org/10.1016/j.apgeog.2012.11.016.

Oquab, M., Bottou, L., Laptev, I., Sivic, J., 2014. Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1717–1724.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.

Pan, B., Shi, Z., Xu, X., 2018. MugNet: deep learning for hyperspectral image classification using limited samples. ISPRS J. Photogramm. Remote Sens. 145, 108–119. https://doi.org/10.1016/j.isprsjprs.2017.11.003.

Penatti, O.A., Nogueira, K., dos Santos, J.A., 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 44–51.

Persello, C., Stein, A., 2017. Deep fully convolutional networks for the detection of informal settlements in VHR images. IEEE Geosci. Remote Sens. Lett. 14 (12), 2325–2329. https://doi.org/10.1109/LGRS.2017.2763738.

Salakhutdinov, R., Torralba, A., Tenenbaum, J., 2011. Learning to share visual appearance for multiclass object detection. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR, pp. 1481–1488.

Sandborn, A., Engstrom, R.N., 2016. Determining the relationship between census data and spatial features derived from high-resolution imagery in Accra, Ghana. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 9 (5), 1970–1977. https://doi.org/10.1109/JSTARS.2016.2519843.

Schmitt, A., Sieg, T., Wurm, M., Taubenböck, H., 2018. Investigation on the separability of slums by multi-aspect TerraSAR-X dual-co-polarized high resolution spotlight images based on the multi-scale evaluation of local distributions. Int. J. Appl. Earth Obs. Geoinform. 64, 181–198. https://doi.org/10.1016/j.jag.2017.09.006.

Shekkizhar, S., 2017. FCN.tensorflow. GitHub https://github.com/shekkizh/FCN.tensorflow.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. In: Proc. International Conference on Learning Representations.

Sivic, J., Zisserman, A., 2003. Video Google: A text retrieval approach to object matching in videos. Computer Vision. Proceedings. Ninth IEEE International Conference on.

Sun, Z., Wang, C., Li, P., Wang, H., Li, J., 2012. Hyperspectral image classification with SVM-based domain adaption classifiers. In: IEEE International Conference on Computer Vision in Remote Sensing (CVRS), pp. 268–272.

Taubenböck, H., Wurm, M., 2015. Ich weiß, dass ich nichts weiß – Bevölkerungsschätzung in der Megacity Mumbai. In: Taubenböck, Wurm, Esch, Dech (Eds.), Globale Urbanisierung. Perspektive aus dem All: 171–178. Springer Spektrum.

Taubenböck, H., Kraff, N.J., Wurm, M., 2018. The morphology of the Arrival City - A global categorization based on literature surveys and remotely sensed data. Appl. Geogr. 92, 150–167. https://doi.org/10.1016/j.apgeog.2018.02.002.

United Nations, 2017. The sustainable Development Goals Report. https://unstats.un.org/sdgs/files/report/2017/TheSustainableDevelopmentGoalsReport2017.pdf.

UN, 2015. The world urbanization prospects. The 2014 revision. http://esa.un.org/unpd/wup/FinalReport/WUP2014-Report.pdf.

UN Habitat 2015: Slum Almanac 2015-2016. https://unhabitat.org/slum-almanac-2015-2016/#.

Wurm, M., Taubenböck, H., Weigand, M., Schmitt, A., 2017. Slum mapping in polarimetric SAR data using spatial features. Remote Sens. Environ. 194, 190–204. https://doi.org/10.1016/j.rse.2017.03.030.

Wurm, M., Taubenböck, H., 2018. Detecting social groups from space –Assessment of remote sensing-based mapped morphological slums using income data. Remote Sens. Lett. 9 (1), 41–50. https://doi.org/10.1080/2150704X.2017.1384586.

Xia, J., Yokoya, N., Iwasaki, A., 2017. Ensemble of transfer component analysis for domain adaptation in hyperspectral remote sensing image classification. In: In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4762–4765.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep Learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 5 (4), 8–36.

Zhu, Q., Zhong, Y., Zhao, B., Xia, G.S., Zhang, L., 2016. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. IEEE Geosci. Remote Sens. Lett. 13 (6), 747–751.

Zipf, G.K., 1941. National unity and disunity - the nation as a bio-social organism. Bloomington, Indiana. https://babel.hathitrust.org/cgi/pt?id=mdp.39015057175484;view=1up;seq=5.

Zou, Q., Ni, L., Zhang, T., Wang, Q., 2015. Deep learning based feature selection for remote sensing scene classification. IEEE Geosci. Remote Sens. Lett. 12 (11), 2321–2325.

A.2. Satellite-Based Mapping of Urban Poverty With Transfer-Learned Slum Morphologies

Reference: Stark, T., Wurm, M., Zhu, X. X., & Taubenböck, H. (2020). Satellite-based mapping of urban poverty with transfer-learned slum morphologies. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13, 5251-5263.

Satellite-Based Mapping of Urban Poverty With Transfer-Learned Slum Morphologies

Thomas Stark, Michael Wurm, Xiao Xiang Zhu, *Senior Member, IEEE*, and Hannes Taubenböck

***Abstract*—In the course of global urbanization, poverty in cities has been observed to increase, especially in the Global South. Poverty is one of the major challenges for our society in the upcoming decades, making it one of the most important issues in the Sustainable Development Goals defined by the United Nations. Satellite-based mapping can provide valuable information about slums where insights about the location and size are still missing. Large-scale slum mapping remains a challenge, fuzzy feature spaces between formal and informal settlements, significant imbalance of slum occurrences opposed to formal settlements, and various categories of multiple morphological slum features. We propose a transfer learned fully convolutional Xception network (XFCN), which is able to differentiate between formal built-up structures and the various categories of slums in high-resolution satellite data. The XFCN is trained on a large sample of globally distributed slums, located in cities of Cape Town, Caracas, Delhi, Lagos, Medellin, Mumbai, Nairobi, Rio de Janeiro, São Paulo, and Shenzhen. Slums in these cities are greatly heterogeneous in its morphological feature space and differ to a varying degree to formal settlements. Transfer learning can help to improve segmentation results when learning on a variety of slum morphologies, with high $F1$ scores of up to 89%.**

***Index Terms*—Fully convolutional network (FCN), remote sensing, slum mapping, transfer learning, urban poverty, Xception.**

I. Introduction

More than 600 million people live in extreme poverty, according the Sustainable Development Goals Report [1]. The credibility of these statistics, however, is in doubt [2], as a systematic global inventory of slums is nonexistent. Although methods for mapping urban poverty in earth observation data have improved tremendously over the past few years, the location of many smaller and lesser-known slum settlements is still unknown to policy makers and NGOs [3]. In the Global South especially, the process of rapid urbanization can overstrain sustainable city planning [4]; in other words, cities are failing to provide the necessary living spaces for their population. The consequence is the development of informal makeshift shelters, resulting in highly dynamic patterns in the urban living spaces of the poor. The perpetual migration into the cities, combined with insufficient housing for low-income groups triggers the formation of these informal settlements, where people looking for job opportunities in the city can find a place to live [1], [5]. Prominent slums like Dharavi in Mumbai and Kibera in Nairobi cannot be denied by authorities and are often tolerated by the local government, but slum dwellers living in smaller and more unknown slums represent a "hidden society"—They often fear eviction and relocation because they are located in endangered areas and are exposed to natural hazards or because city governments wish to upgrade these areas [6], [7].

Squatter settlements, favelas, huts, villas miseria, bidonvilles, urban villages, slums, informal settlement, and many other names are typically used, depending on the global location, to refer to urban poor areas. In general, all these names emphasize negative characteristics and imply nonaffiliation from a city's point of view [8]. Additionally, all terms for poor urban areas, while generally understood, contain ambiguities in their morphological appearance, ranging from very deprived areas to lesser ones [7], [9]. This diversity can, to some extent, be described by regional differences, cultural context, and the building material available for construction.

In this study, urban poverty areas are addressed on a large scale, including highly variable morphological slum features from 10 cities in the Global South. Thus, a uniform definition of the exact urban morphology of poverty is infeasible. While there are many discussions on the characterizations and nomenclature of urban poverty, in the context of this study, we refer to all urban poverty areas, with different physical morphologies compared to formal settlements, by the term *slums* for naming purposes.

Mapping these settlements is not a trivial task and certain challenges have to be addressed. The first challenge can be described as interurban variability, where morphological slum features can change depending on their particular geographical location. But these morphological slum features are conceptually fuzzy, do not have international consensus, and are, thus, very difficult to describe. The examples in Fig. 1 reveal that morphologic appearances of poverty can be different in every city, ranging from very dense low-rise shacks in Mumbai [see Fig. 1(a)] to three-story buildings in Medellin [see Fig. 1(d)]. A second challenge, complicating the already complex task

Manuscript received April 7, 2020; revised July 9, 2020; accepted August 15, 2020. Date of publication August 24, 2020; date of current version September 11, 2020. This work was supported in part by the European Research Council within the European Union Horizon 2020 Research and Innovation program under Grant 714087-So2Sat and in part by the BMBF Project inform@risk (FZK: 03G0883C). *(Corresponding author: Thomas Stark.)*

Thomas Stark is with the Chair of Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: t.stark@tum.de).

Michael Wurm and Hannes Taubenböck are with the German Remote Sensing Data Center, German Aerospace Center, 82234 Oberpfaffenhofen, Germany (e-mail: Michael.Wurm@dlr.de; Hannes.Taubenboeck@dlr.de).

Xiao Xiang Zhu is with the Chair of Signal Processing in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Oberpfaffenhofen, Germany (e-mail: xiaoxiang.zhu@dlr.de).

Digital Object Identifier 10.1109/JSTARS.2020.3018862

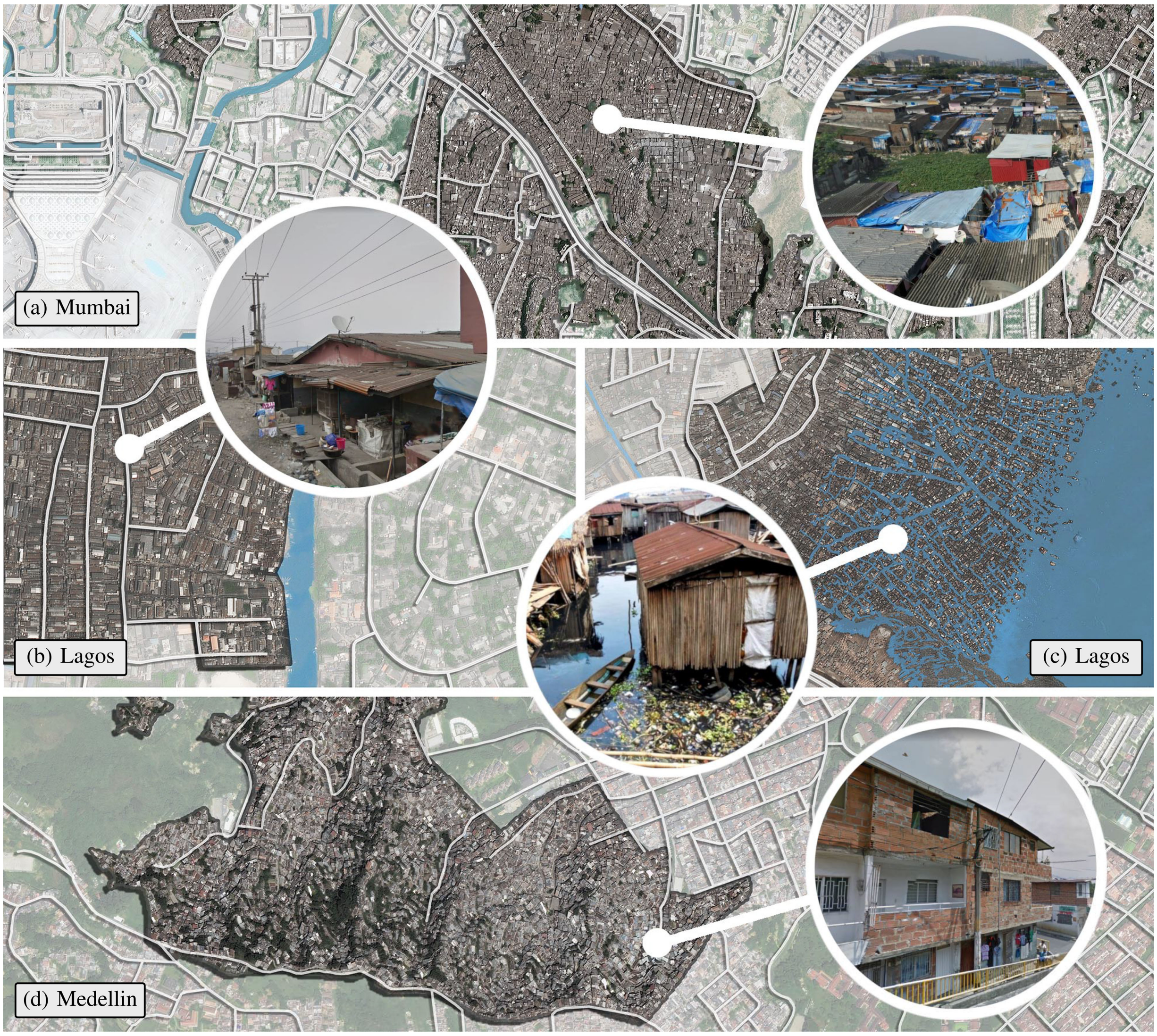

Fig. 1. Comparison of the inter- and intraurban variability of slums. Image (a) shows a typical slum in Mumbai, India, consisting of very densely built shacks. The images (b) and (c) in the middle show two very different slums in Lagos, Nigeria: poverty areas in the city's periphery as well as the downtown floating slum of Makoko in the Lagoon of Lagos. Image (d) depicts a slum in Medellin, Colombia, with three-story buildings made of concrete. Images from Google Street View provide additional close-up information on the local built-up structure.

of interurban variability, shows that slums can also feature an intraurban variability within the same city [8], [10]. These varying intraurban morphological slum features can be seen in the middle of Fig. 1(b) and (c). Although the slum areas in Lagos are located within the same city, their morphological appearance is inherently different. The very dense swimming shacks of the Makoko slum in Lagos [see Fig. 1(c)] and the less dense slums in the peripheral area with an almost regular road network shown in Fig. 1(b) demonstrate intraurban variability.

Fig. 1 also shows that deprived poverty settlements often come with some variation in the previously mentioned slum features. Fuzzy borders and similar morphological features on formal built-up structures can lead to a complex super state of the affiliation with a slum category. According to the work in[8], characteristic features for slums are settlements of incredible density, complex building structures, and significantly different appearances from their formal counterparts. In [3], slums are interpreted in five dimensions of their morphologic appearance: complex building geometry, high building density, irregular or nonexistent road network, roofing material, and site characteristics. These slums are described by the morphological appearance and can contain a variation of their aforementioned features. Additionally, in all examples in Fig. 1, the street layout is highlighted, making the difference between an orderly planned road structure in the formal settlements, a more irregular layout, or even a nonexistent road network more visible in the slum areas. Thus, besides the morphology of individual buildings, the street network can be seen as a key feature for differentiating between formal settlements from slums.

In this study, we aim to address the challenge of large-scale slum mapping featuring varying slum morphologies in the context of an applicable mapping approach. Thus, 10 globally distributed cities are selected: Cape Town (South Africa), Caracas (Venezuela), Delhi and Mumbai (India), Lagos (Nigeria), Medellin (Colombia), Nairobi (Kenya), Rio de Janeiro and São Paulo (Brazil), and Shenzhen (China), featuring different cultural regions, topographies, and building morphologies. The perception of people and the spatial structure is subjective [11]. This is also the case with slums, i.e.: What can be called a slum, since the boundaries to formal settlements are often fuzzy. We apply the categorization of slums as presented in [8], as we seek to integrate various morphologies into our mapping experiments. In [8], slums are grouped into multiple representations using five variables that describe their morphologies. The most extreme slum morphologies, meaning high building densities, nonuniform building orientation, high heterogeneity of the slum buildings themselves, very small building sizes, and low-rise building heights, can be found in the slums of Mumbai, Caracas, and Nairobi. This first category of slums, which is referred to as C_1, reflects stark morphological differences from formal settlements and correspond to the greatest possible physical assumption of a morphological slum. A second category of slums C_2 can be formed if the slum morphology deviates in a small capacity from the features of C_1. These slum types can be found in Delhi, Medellin, Lagos, and to an extreme in the urban villages of Shenzhen: There, slums are still very dense and disregard orderly building alignments, but their building heights are often more than one story high and feature a variation of regular and irregular road layouts in the slum settlements. In some cases, morphological slum features deviate more significantly from the typical assumption of the complex state of slum settlements. This third category C_3 of slums can be found in Cape Town, Rio de Janeiro, and São Paulo. In these cities, slum settlements can sometime even share urban morphologies found in their formal counterparts. The Township Victoria Merge in Cape Town and the Favela Paraispolis in São Paulo feature a regular road layout and less heterogeneous building alignments, making these areas difficult to categorize as C_1 or C_2. Here, the morphology of the slums is a mixture of the slum features typical of the first two groups and formal settlement structures.

The aim of this article is to systematically test transfer learning techniques using a fully convolutional network (FCN) to map slums of varying morphologic appearances from knowledge learned in different geographical and cultural settings. By using a large-scale globally distributed dataset of slums, the FCN is better able to generalize and, thus, is able to map slums in areas where this was previously not possible on high-resolution remote sensing data. We want to analyze the extent of interurban variability of slum settlements on a global scale and understand if it is possible to learn from features of varying morphological poverty representations. For this task, we specifically design a fully convolutional Xception network (XFCN) to train on multichannel remote sensing data. In this study, the XFCN is tested on its transfer learning capabilities of different slum categories, for comparative studies of the Xception model in regards to other convolutional neural networks (CNNs), we suggest the following papers [12]–[14]. As an additional option, auxiliary data in the form of the road layout from Open Street Map can be used as an extended input layer to support the model in its learning task.

The remainder of this article is structured as follows: In Section II, background on poverty mapping and the state of the art of semantic segmentation is reviewed. In Section III, the methodology of our proposed approach using a XFCN is presented. In Section IV, the remote sensing and auxiliary datasets including preprocessing steps are shown and the experimental setup is introduced. In Section V, the results of all experiments are shown. In Section VI, the results of all experiments are discussed with respect to their implication on poverty mapping. Finally, Section VII concludes this article.

II. Background and Related Work

Deprived poverty settlements feature a characteristic structural type in many cities of the Global South. Various approaches to detecting slums, ranging from machine learning techniques to object-based solutions, are presented in Section II-A. In the past five years, deep learning procedures for semantic segmentation of slums have been able to surpass traditional mapping methods in their ability to achieve mapping accuracies. These techniques for pixelwise classification are presented in Section II-B.

A. *Mapping Urban Poverty With Satellite Data*

To describe physical slum characteristics using remote sensing data, the morphological features of urban poverty need to be well understood. Thus, the data must be able to represent the physical properties of slum settlements. For example, since many slum buildings are considerably below 100 m^2 and slum areas often only have a size of 1 ha [10], [15], [16], the related images for their identification require a high spatial resolution. Moreover, roof surfaces are frequently not homogeneous in shape and color; when using high-resolution data, some of the roof pixels will consist of mixed roofing materials. Thus, a specific geometric resolution is needed to capture the morphological poverty features. At the same time, when talking about mapping poverty in multiple globally distributed cities, data availability also needs to be taken into consideration. This favors both the Copernicus mission Sentinel-2 and Planet Labs data from the PlanetScope satellite as optical sensor solutions, since both products are globally available. In [17], Sentinel-2 data were used to map slums and [18] compared Sentinel-2 data and very high resolution data. Both studies conclude that while mapping urban poor areas are possible in high-resolution 10-m ground sampling distance, it is a very limiting factor, especially considering mapping smaller slum patches. Given this circumstance, PlanetScopes 3-m geometric resolution strikes a perfect balance between data availability and high spatial resolution.

In the related scientific literature on slum mapping, various methods have been presented. In [17] and [19], the studies aimed at identifying complete slum patches using a combination of machine learning and textural feature engineering methods. Other work has been done using socioeconomic data and spatial features to determine income levels of slum settlements on a neighborhood level [9], [20], [21]. In [22], only the street network was used to predict slum areas in a combination of traditional machine learning and artificial neural networks. In

[7], [8], and [16], poor urban areas were analyzed on the level of individual buildings using an object-based approach to identify the varieties of slums and their temporal changes.

In the past five years, using deep learning techniques has become a popular trend, as it has been shown that mapping accuracies improved rule-based approaches significantly for mapping slum patches [18]. In [23] and [24], nighttime light intensities were used as a proxy for poor urban areas to transfer learn a CNN to high-resolution remote sensing data. In [25]–[27], fully convolutional neural networks (FCNs) were used to map slums on either high-resolution or very high-resolution data, whereas Wurm *et al.* [18] and Stark *et al.* [28] used different transfer learning techniques to map slums between different satellite sensors in the same city and between geographically separated cities, respectively. The authors concluded that not only more data, but also a novel deep learning architecture and more rigorous regularization is necessary for robust segmentation of slums on a large scale.

B. Semantic Segmentation Using Deep Learning

Semantic segmentation means understanding an image at a pixel level. While traditional CNN aim to classify a whole image patch, FCNs classify each pixel of an image, offering more information about the area and shape of the target class. First introduced in [29], FCNs replace the fully connected layers of a standard CNN with convolutional layers and dilated convolutions for upsampling to the original input dimensions. In the past five years, more advanced methods for semantic segmentation using deep learning techniques have been explored. Improvements in the backbone architecture as well as the upsampling phase can have been reported. Both U-Net [30] and SegNet [31] improved upsampling techniques, introducing long distance skip connections and convolutions during the upsampling phase, for semantic segmentation. While the original FCN in [29] used vgg16 architecture [32], today deeper and more efficient backbone models are available. GoogLeNet [33] and its Inception versions [34], [35] introduced deeper and more advanced implementations using network in network approaches, whereas ResNet variants [36] introduced skip connections and heavy batch normalization. Currently, not only the depth of the network but also its efficiency is major factor to be taken into consideration. While recently, the trend has been to go deeper with convolutions, networks like Xception [12], and EfficientNets [37] can outscore deeper variants while having fewer parameters to train.

Specific improvements for semantic segmentation in remote sensing data could be achieved in [38], where relation-augmented FCNs are used, in [39], with a gated graph CNN and structured feature embeddings, and in [40], by fusing very high-resolution data with auxiliary data. Training a CNN from scratch requires a significant amount of data and processing power. It is also very time consuming [41], which is why fine-tuning or transfer learning approaches are often used in order to handle less training data or transfer knowledge from a source domain to a target domain. Fine-tuning a CNN from a large dataset, such as ImageNet [42], Coco [43], or PascalVOC [44], was very popular in the first stages of adapting deep learning techniques into to the remote sensing domain [41], [45], but feature transformation from often low-quality natural images to multichannel remote sensing data means sacrificing valuable data information in the spectral and radiometric resolution of the satellite images [41]. Therefore, training a CNN from scratch specifically on remote sensing data often yields better results [46]–[49]. To take full advantage of the data richness present in remote sensing data, training from scratch offers great potential in learning high-quality feature representation when enough data and computational power are available.

III. Proposed Approach

CNNs pretrained on natural images most often limit the depth of the input image to just three channels, and thus, the high-quality multispectral data of remote sensing imagery are neglected. To exploit the full spectral depth of optical satellite sensors, CNNs can be trained on multispectral data from scratch on any number of input channels, but training these networks can be very computationally expensive [41]. Specific architectures can strike a balance on being as deep as possible, while at the same time, an efficient approach of implementing convolutions can save parameters, making the model more light weight and easier to train. Both these effects are present in the Xception [12] network, which is an evolution of the Inception models [33]–[35]. We propose using a modified Xception network as the backbone architecture to create a FCN, where a fully convolutional flow for segmentation follows the exit flow of the Xception network.

A. Backbone Architecture

The Xception network gets its name from the modules that make up the backbone architecture. The main idea behind these modules is to decouple cross-channel and spatial correlations to shrink the parameter size of the model. The Xception module is an evolution of the modules that are present in the Inception networks and take this principle to the extreme, hence its name. Fig. 2 shows an Xception module in detail. First, a depth/channel-wise 3×3 convolution is performed on all input dimensions; afterward, a pointwise 1×1 convolution maps the data to the desired output space. Thus, compared with conventional convolutions, we do not need to perform convolution across all output channels. This means that a number of connections are fewer and the model is lighter.

The Xception architecture is a linear stack of depthwise separable convolution layers with residual connections. This makes the model very easy to define and modify. The complete architecture, depicted in Fig. 3, consists of multiple entities. The entry flow is split into multiple blocks. The first block employs a 2-D convolution at stride 2 and valid padding, whereas the second 2-D convolution uses same padding and no stride, reducing the input dimension from $299 \times 299 \times n_{\text{dim}}$ to $147 \times 147 \times 64$. The remaining blocks use a similar sequence of two Xception modules, where the second module is accompanied by a max

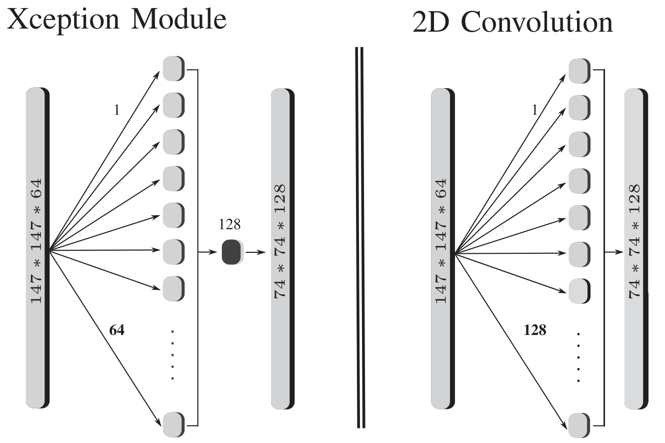

Fig. 2. Xception module in comparison to a standard 2-D convolution for the first depthwise separable convolution within the XFCN. After a depthwise convolution on the number of input parameters, a pointwise convolution follows, resulting in the desired number of output features.

pooling operation, which is fused with a residual connection from the input tensor of the previous Xception module at a stride of 2. The middle flow successively employs three Xception modules eight times while keeping its tensor dimension constant at $19 \times 19 \times 728$. Finally, within the exit flow, two blocks of each two Xception modules round up the Xception backbone architecture, where the first of the two blocks is fused with a residual skip connection. During the complete XFCN, all convolutions are a combination of batch normalization, a ReLU activation function, and a dropout layer. In total, the XFCN consists of 41 convolutional layers, including residual skip connections in the backbone.

B. Upsampling

The decoder of the XFCN uses an upsampling approach similar to the original FCN [29]. In our XFCN, five dilated convolutions are used to upscale the output of the exit flow with its dimension of $10 \times 10 \times 2048$ back to the original input height and width dimension. A softmax classifier is used to produce a single prediction tensor with a size of $299 \times 299 \times 1$. The decoder uses four long-distance skip connections fused with the fitting counterpart of the entry flow to preserve low-level features and a padding of two to ensure a fine-grained upsampling performance, as seen in the upscale flow of Fig. 3.

IV. Data and Experimental Setup

The XFCN introduced in Section III is specifically set up to map slums in high-resolution remote sensing data. In areas of low slum coverage especially, a transfer learning approach is necessary to train the XFCN on multidimensional remote sensing data. In this section, we present the remote sensing data used in this article, the sampling methods employed to create a large-scale dataset for transfer learning purposes, and the experimental setup of the XFCN.

A. Data Preprocessing and Data Sampling

For our experiments, we deployed high-resolution PlanetScope data from Planet Labs, Inc., [50]. With its 3-m resolution, resampled from a 3.7-m ground sampling distance, a daily global coverage, and a four-channel blue, green, red, and near infrared (B, G, R, NIR) composite, the data fit the needs of a large-scale poverty mapping approach in every respect. Beyond the spectral bands, we included the normalized difference vegetation index (NDVI) as an additional feature that increases number of the input images to five channels. Table I indicates in detail all PlanetScope datasets we used in this study. All datasets are surface reflectance 16-b data from the original PlanetScope data. Each band is min–max normalized to a float32 range of $0 - 1$ to create an evenly distributed dataset suitable for our deep learning framework.

The reference data for all 10 cities consist of manually mapped polygons for each PlanetScope scene. The reference data were created by multiple remote sensing experts to ensure consistency between all test sites. Additionally, the reference data were compared to ground truth data of poverty areas according to census tracts, when this was available. In cases where no official census data were available, or the ground truth data were outdated, the reference was created based on Bing aerial imagery and Google Street View images. The area of each city's dataset is limited by the PlanetScope scene and can be seen in Table I. All slums larger than 1 ha within the PlanetScope scene are included in the dataset. All slums, while featuring various different morphologies, were delineated in a coherent manner to ensure consistency when transfer learning between each city's dataset.

As an additional data source, we used the road network in Open Street Map to create an auxiliary layer for the input data tensor (B, G, R, NIR, NDVI, OSMp). To cope with inconsistencies in the street network between cities and the road categories, only paved roads, accessible by automobile, were selected, indicating major and residential usage. Foot and dirt paths were excluded from the OSM road network to create a coherent and unified data layer across all 10 datasets. Using only these roads, we calculated a binary logarithm ($\log_2$) proximity to each road. This not only shows the distance from each pixel to the nearest street, it also gives insights about the general shape of the road network, which can serve as a useful indicator of settlement structures [10], [22].

The input data-cube is split into a $299 \times 299 \times n_{\text{dim}}$ image patch to match the input dimension of the XFCN. The image patches are split with a large overlap of 199 pixels in both x and y directions to increase the datasets volume. To further increase the dataset size and its slum sample proportion, we make use of data augmentation on the image patches used for training. A variation of image translation, dropout, and gamma adjustments in [51] is used to increase the original data by a factor of four; each of these augmented image patches is then rotated three times by 90°. The augmenters are listed in Table II and are chosen based on successful training techniques from the work in [52] and [53].

Table I provides insight about the dataset used for training the XFCN. Ten cities in the Global South are selected, three in Africa

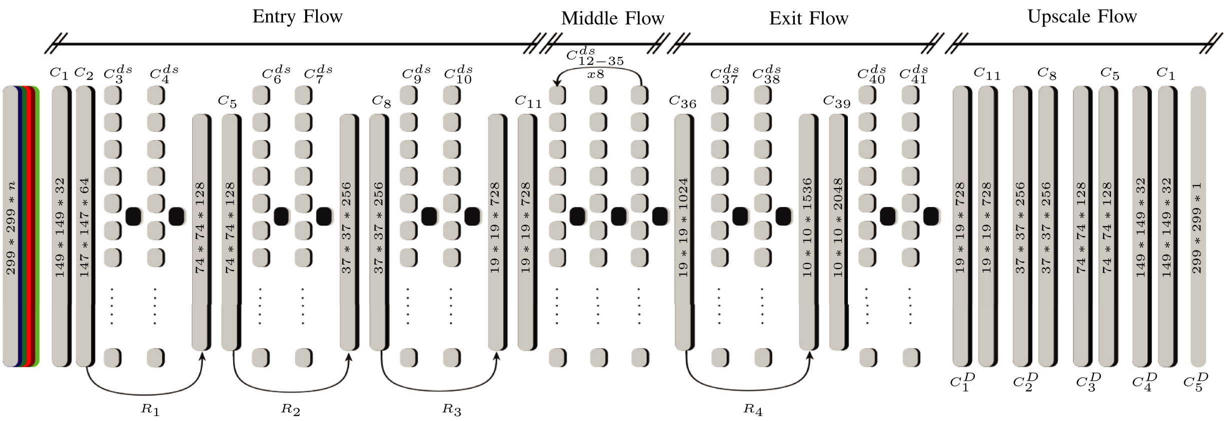

Fig. 3. Architecture of the XFCN. The Xception backbone is slightly changed to allow for a multidimensional input and more rigorous regularization. After the exit flow, a fully convolutional flow follows. All convolutional blocks are a combination of standard 2-D convolutions C_n or depthwise separable convolutions C_n^{ds} in combination with batch normalization, dropout, and ReLU activation functions. The XFCN features residual skip connection throughout the whole network (R_n), and during the upscale flow the dilated convolutions (C_n^D) are fused with the long distance skip connections from the entry flow.

TABLE I
Overview of the Datasets Used for Training the XFCN

Test site	Caracas CA	Mumbai MU	Nairobi NA	Delhi DE	Lagos LA	Medellin ME	Shenzhen SH	Cape Town CT	Rio RI	São Paulo SP
Slums category	C_1	C_1	C_1	C_2	C_2	C_2	C_2	C_3	C_3	C_3
PlanetScope Area $[km^2]$	357	$1,379$	211	852	230	59	$1,471$	356	$2,086$	$3,764$
Number of slums	104	452	47	232	51	49	$1,872$	70	404	905
Total area of slums $[km^2]$	30.4	41.3	8.2	5.3	15.0	4.1	46.2	6.1	26.4	51.3
Mean size of slums $[ha]$	29.2	9.1	17.5	2.3	29.4	8.4	2.5	8.3	6.5	4.4
Slum size dispersion $[ha]$	72.8	20.6	43.6	5.3	53.7	9.9	3.0	10.3	11.0	6.1
Training data										
Number of augmented training patches	10,902	19,109	2,300	2,162	3,090	1,565	23,909	2,117	10,722	18,822
Slum sample proportion [%]	38.2	26.4	22.7	7.3	46.1	24.1	22.7	19.8	19.4	16.9
Training steps										
$XFCN_{city}$	34,067	59,715	7,187	6,756	9,656	4,890	74,715	6,616	33,506	58,818
$XFCN_{LSP}$	288,044	283,453	230,991	231,336	314,897	291,096	221,211	266,333	288,662	260,818
$XFCN_{LSP}^{TF}$	23,848	41,801	5,031	4,729	6,759	3,423	52,300	4,630	23,454	41,173

The table shows information on each city's dataset, the training data, and finally, the total number of training steps for each experiment of the XFCN model is shown.

(Lagos, Nairobi, Cape Town), three in Asia (Delhi, Mumbai, Shenzhen), and four in Latin America (Caracas, Medellin, Rio de Janeiro, São Paulo). Ten cities are chosen due to their varying morphologic slum features, providing a comprehensive morphologic poverty feature set to learn diverse slum representations. All 10 cities are categorized by their morphological features from Section I into the categories C_{1-3}. Although an intraurban variability of the morphological slum features is present in all datasets, the slums of each city are grouped into these three categories according to the most prominent morphologic slum features of all the slums in each city. Caracas, Mumbai, and Nairobi represent the first category of slum morphologies C_1, where high building densities, nonuniform building orientation, high heterogeneity of the slum buildings themself, very small building sizes, and low-rise building heights can be found. Delhi, Lagos, Medellin, and Shenzhen represent typical slum features of type C_2. In these cities, slum settlements can deviate to a minor extent from the aforementioned features. In Cape Town, Rio de Janeiro, and São Paulo, slums deviate more significantly from the slum morphologies of type C_1, forming a third type of slum category, C_3. Additionally, the dataset of these 10 cities can be described by four components seen in Table I: number of slums, mean size of slums, the number of image patches, and the slum sample proportion. In Mumbai, Rio de Janeiro, São Paulo, and Shenzhen, more than 400 slums are present in their dataset, but with a smaller mean slum size in Rio de Janeiro, São

TABLE II
DATA AUGMENTATION FOR THE TRAINING DATA

Augmentation	Crop [px]	Translation	Dropout	Gamma
1	$(10, 20)$	$(0.8, 1.2)$	None	$(0.7, 1.3)$
2	$(20, 10)$	$(1.0, 1.5)$	Salt&Pepper	$(0.7, 1.3)$
3	$(10, 5)$	$(1.5, 1.2)$	None	$(0.7, 1.3)$
4	$(5, 10)$	$(1.1, 1.5)$	Salt&Pepper	$(0.7, 1.3)$

Dropout and Gamma augmentations are only used on the images and not their annotations. All augmentations are rotated three times by 90°.

Paulo, and Shenzhen, only the Mumbai dataset surpasses a slum sample percentage of 25%. In contrast, Cape Town, Caracas, Lagos, Medellin, and Nairobi feature fewer slums, but a larger mean slum size in Caracas, Lagos, and Nairobi also shows a substantial slum sample proportion. Delhi exhibits the lowest slum sample proportion, with only 7% of all pixels labeled as slums. Although there are in total more than 200 slums in the dataset, the very small mean size of slums indicates a challenging dataset. Grouping the 10 cities by these 4 features can indicate where the XFCN is confronted with an easier or more challenging task. But regardless of a large slum sample proportion or a vast number of slums in the dataset, the decisive challenge is the combination of the slum morphology types C_{1-3} in combination with the training dataset components of Table I.

B. Experiments

The XFCN was trained on an augmented dataset for each single city as a benchmark to test transfer learning capabilities. The models that trained in one city and tested on unseen image patches of the same city are labeled as $\text{XFCN}_{\text{city}}$. A global poverty training dataset was created where all training patches were combined into one big dataset, whereas all images of the tested city were excluded. The XFCN trained on the global dataset, which was tested for each city in a leave-one-out manner, was named XFCN_{LSP} [large-scale poverty (LSP) dataset]. In addition, the XFCN_{LSP} was transfer learned to a training dataset of each city $\text{XFCN}_{\text{LSP}}^{\text{TF}}$; thus 30 experiments for each, the five- and six-dimensional input dataset were conducted. Throughout all experiments, the complete dataset and the dataset of each city were split into training (70%), validation (15%), and testing (15%), where the testing and validation image patches were selected manually for each city to create a coherent and spatially separated dataset and to compare results in a meaningful manner.

1) Transfer Learning: The XFCN was trained using an inductive transfer learning approach. Given a source domain dataset D^S and a learning task T^S, a target domain dataset D^T and learning task T^T, we aim to improve the learning of the target predictive function $f^T(\cdot)$ using the knowledge in D^S and T^S, where $T^S \neq T^T$ [54]. In this context, the $\text{XFCN}_{\text{LSP}}^{\text{TF}}$ is trained on the source domain dataset D_{LSP}^S to target dataset D_{city}^T of each city excluded from the D_{LSP}^S dataset. All variables of the XFCN were available for training during the transfer learning process.

2) Experimental Setup: The XFCN was implemented in TensorFlow and adapted from the works in[55] and [56]. To prevent overfitting, multiple constraints were employed. Batch normalization with a batch size of 16 was used to improve the learning procedure, including a weight decay of 0.99 for L2-regularization to reduce overfitting. After each convolution, a dropout layer followed. By randomly dropping nodes with a 20% probability during each weight update cycle, the model had to adapt to learn independent representations. The XFCN was trained using a softmax cross entropy-loss function and using the Adam optimizer [57]. All models used an exponential decaying learning rate. The initial starting learning rate was quite high at 0.1, which is possible due to using batch normalization, since no activation can be either too high or too low [58]. When the XFCN was transfer learned, a lower learning rate of 0.01 was used to start training. The XFCNs were trained depending on the size of their dataset. The total number of steps of each experiment can be seen in Table I, and for each experiment, early stopping was used to end the training process as soon as the validation accuracy did not substantially improve.

V. RESULTS

Unseen image patches from the test dataset with an image size of $299 \times 299 \times n_{\text{dim}}$ were used for testing and were predicted with an overlap of 199 pixels in both x and y directions. Thus, nine image patches can be used to create an area of 100×100 pixels of the same observation. The most probable result can be derived using a majority operator. This method not only ensures that uncertainties in the model variance are dealt with more robustly, but also reduces the difficulties of predicting in the edge region of the image patches.

Accuracies are reported in the $F1$-score and the Intersection over Union (IoU). The $F1$-score takes both error of omission and error of commission into consideration to compute its score. Thus, the $F1$-score can be recognized as the harmonic mean of precision and recall, as seen in (1)

$$F1 = 2 \times \frac{\text{TP}/(\text{TP}+\text{FP}) \times \text{TP}/(\text{TP}+\text{FN})}{\text{TP}/(\text{TP}+\text{FP}) + \text{TP}/(\text{TP}+\text{FN})} \tag{1}$$

$$\begin{aligned} \text{IoU} &= \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \\ \text{where TP} &= \text{True positives} \\ \text{FP} &= \text{False positives} \\ \text{FN} &= \text{False negatives.} \end{aligned} \tag{2}$$

The IoU in (2), also referred to as the Jaccard index, is defined as the size of the intersection between the ground truth and the classified map, divided by the size of the union of the sample sets. The IoU is a very penalizing metric and values above 50% can be considered an adequate match of the similarity between ground truth and the predicted map [59], since in real-world applications, it is not likely that the x and y coordinates of the predicted poverty area are going to exactly match the x and y coordinates of the ground truth. Results for all 60 experiments are reported in Table III. The results are grouped according the

TABLE III
RESULTS FOR ALL EXPERIMENTS USING THE IOU AND THE $F1$-SCORE ACCURACY MEASURES

Dataset		CA	MU	NA	DE	LA	ME	SH	CT	RI	SP
Slums category		C_1	C_1	C_1	C_2	C_2	C_2	C_2	C_3	C_3	C_3
	n_{dim}	Intersection over Union (IoU)									
$XFCN_{city}$	5	59.13	71.16	**73.13**	56.23	61.04	51.47	63.24	66.56	58.89	68.42
$XFCN_{LSP}$		58.16	50.58	49.05	57.97	**67.48**	61.27	63.01	57.48	56.74	56.43
$XFCN_{LSP}^{TF}$		80.70	**80.86**	75.63	58.10	70.44	68.35	70.11	77.98	73.37	70.49
$XFCN_{city}$	6	76.73	78.49	78.09	60.22	71.91	48.95	**85.98**	72.28	54.48	61.19
$XFCN_{LSP}$		78.14	66.32	64.54	67.18	80.77	70.51	**80.99**	78.63	65.25	64.08
$XFCN_{LSP}^{TF}$		81.62	81.80	79.73	64.65	74.72	69.83	**86.29**	81.54	60.95	64.20
		F1-score (F1)									
$XFCN_{city}$	5	63.66	76.29	56.63	61.22	51.76	**77.23**	71.19	71.66	64.12	74.84
$XFCN_{LSP}$		63.87	54.81	56.15	63.66	70.78	66.30	**77.31**	59.12	63.50	63.37
$XFCN_{LSP}^{TF}$		85.93	**86.98**	79.76	59.20	74.56	75.73	73.62	83.89	80.03	77.25
$XFCN_{city}$	6	81.48	83.98	82.67	68.47	71.99	60.15	**89.49**	73.81	57.27	64.16
$XFCN_{LSP}$		82.68	70.33	66.94	71.81	76.78	76.58	**83.59**	82.61	71.52	70.20
$XFCN_{LSP}^{TF}$		86.17	86.63	83.24	67.44	79.52	72.72	**89.48**	83.76	64.46	67.53

The top part of the table shows the experiments for the five-dimensional remote sensing data, whereas the bottom part includes the proximity to the road network as an additional sixth input dimension. The highest accuracies for each experiment are presented in bold; the highest overall accuracy for each accuracy score is highlighted in gray.

morphologic slum categories C_{1-3}. The highest accuracies for each row are presented in bold and the highest overall accuracy for each $F1$-score and IoU is highlighted in gray. The following paragraphs report the results based on the three experiments XFCN$_{\text{city}}$, XFCN$_{\text{LSP}}$, and XFCN$_{\text{LSP}}^{\text{TF}}$.

A. XFCN$_{city}$

The first set of experiments was trained on a single city's datasets and tested in a spatially separated area of the same city. The XFCN trained on five-dimensional input data, including the channels B, G, R, NIR, and the NDVI, achieved a mean IoU for all cities of 62.93% and a mean $F1$-score of 66.86%. Training the XFCN on six-dimensional data, including the proximity to the road network, achieved a mean IoU for all cities of 67.98% and a mean $F1$-score of 73.35%. Including the Open Street Map road network in the dataset could increase the mean IoU by 5.05% and the $F1$-score by 6.49%. The best results on the five-dimensional data were achieved in Mumbai and Nairobi (C_1) and São Paulo (C_3), with an IoU of up to 73%. When training on six-dimensional data, high IoU accuracies of over 70% could be reached in Caracas and Mumbai (C_1), Lagos and Shenzhen (C_2), and Cape Town (C_3).

B. XFCN$_{LSP}$

The second set of experiments was trained on a large-scale poverty dataset in a leave one out manner, training on a combined dataset of nine cities and testing the results on the remaining city. Thus, the XFCNs ability to map slums from features learned on a global slum repository was tested. The XFCN trained on five-dimensional input data achieved a mean IoU for all cities of 57.81% and a mean $F1$-score of 63.87%. Training the XFCN on six-dimensional data achieved a mean IoU for all cities of 71.64% and a mean $F1$-score of 75.30%. Including the Open Street Map road network in the dataset could increase the mean IoU by 13.82% and the $F1$-score by 11.41%. An IoU of over 60% could be reported in Lagos, Medellin, and Shenzhen (C_2) for the five-dimensional data. Best IoU accuracies of around 80% for six-dimensional inputs could be reached in Caracas (C_1), Lagos and Shenzhen (C_2), and Cape Town (C_3).

C. XFCN$_{LSP}^{TF}$

The third set of experiments was set up as an inductive transfer learning experiment, where the XFCN is first trained on a large-scale poverty dataset in a leave one out manner; afterward, the XFCN was transfer learned to the remaining city's training dataset and tested in a spatially separated area of the same city. The XFCN trained on five-dimensional input data, achieved a mean IoU for all cities of 72.60% and a mean $F1$-score of 77.69%. Training the XFCN on six-dimensional data achieved a mean IoU for all cities of 74.53% and a mean $F1$-score of 78.10%. Including the Open Street Map road network in the dataset could increase the mean IoU by 1.93% and the $F1$-score by 0.41%. In this experiment, the overall highest accuracies could be reached for the five-dimensional remote sensing data in Mumbai (C_1) with an IoU of 80.86% and for the six-dimensional data in Shenzhen (C_2) with an IoU of 86.29%. In general, the transfer learning approach is able to reach IoU scores of over 80% for the five-dimensional data in Caracas and Mumbai (C_1) and for the six-dimensional data in Caracas and Mumbai (C_1), Shenzhen (C_2), and in Cape Town (C_3).

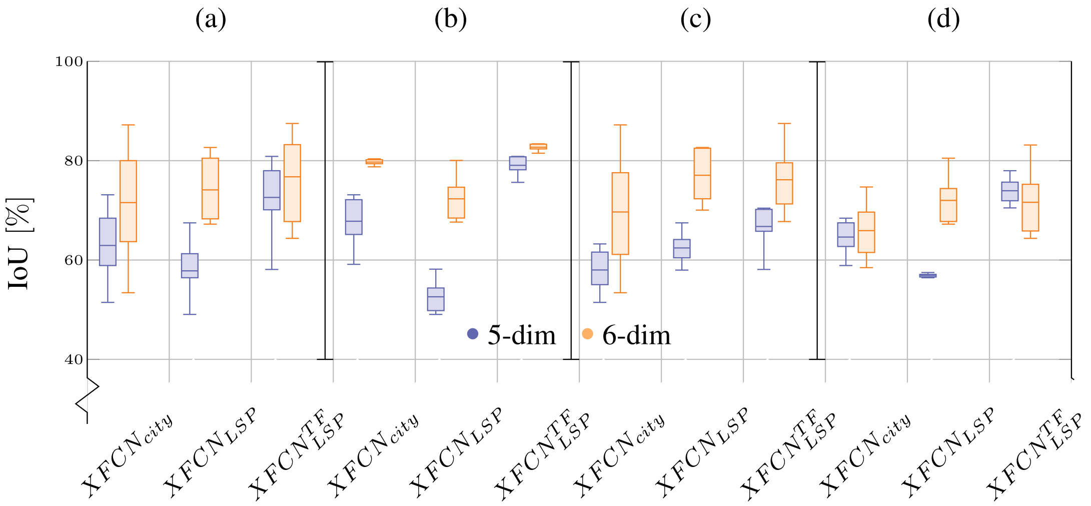

Fig. 4. IoU accuracies represented in a boxplot for (a) all 10 cities and (b)–(d) each slum category for the three experiments $\text{XFCN}_{\text{city}}$, XFCN_{LSP}, and $\text{XFCN}_{\text{LSP}}^{\text{TF}}$ on the five-dimensional remote sensing data and the six-dimensional data where the proximity to the road network is included. (a) C_{1-3}. (b) C_1. (c) C_2. (d) C_3.

VI. Discussion

Comparing the results of the XFCN from the five-dimensional input data, which solely consisted of remote sensing data, to the results of the six-dimensional input data where the proximity to the road network is added as an additional input layer, the accuracies of the model tended to increase. Rio de Janeiro and São Paulo are the only datasets where the IoU decreased when comparing the five- and six-dimensional input data. This can be attributed to slums featuring morphologic types of category C_3 in these cities. In both Rio de Janeiro and São Paulo, an orderly structured road network in slum settlements deviated significantly from typical complex slum morphologies, where often nonpaved roads define an irregular mosaic of settlement patterns. In general, the mean IoU for all five-dimensional experiments is 63.42% and can be increased to 75.94% when using the six-dimensional input data to train the XFCN. Thus, the proximity to the road network, used as an additional input dimension, is found to help the model to better differentiate between formal settlements and slum settlements.

In our tests, we defined the set of experiments where the XFCN was trained and tested within the same city ($\text{XFCN}_{\text{city}}$) as a baseline for comparison with the other experiments. In Fig. 4(a), all experiments can be compared to each other. The mean IoU decreases from 59.83% to 57.81% when comparing the $\text{XFCN}_{\text{city}}$ and the XFCN_{LSP} trained on five-dimensional input data, but increases from 68.83% to 71.64% when comparing the six-dimensional input data. These results show that including auxiliary information about the road network can help improving segmentation results when the XFCN is trained on a generalized large-scale dataset including various categories of slum morphologies. The results for the transfer learned XFCN ($\text{XFCN}_{\text{LSP}}^{\text{TF}}$) achieved the highest overall mean IoU accuracies with 72.60% for the five-dimensional data and 74.53% for the six-dimensional data. Table I shows the setup for all training datasets: We identify some challenging datasets when there are few training samples, a small slum sample proportion in the respective city, small-sized areas of urban poverty, or a combination of these issues. In these cases, we find the learning task can be difficult for the $\text{XFCN}_{\text{city}}$. These attributes can be seen in some variation throughout all datasets and slum categories; e.g., Nairobi (C_1), Delhi and Medellin (C_2), and Cape Town, Rio de Janeiro, and São Paulo (C_3). Accuracy measures confirmed this analysis in Delhi and Medellin (C_2), and Rio de Janeiro (C_3), with IoU accuracy scores lower than 58.98% for five-dimensional data and 60.22% for six-dimensional data. For Nairobi (C_1) and Cape Town (C_3), this is not the case, which can be attributed to stark differences in formal and informal settlement morphologies, even in Cape Town (C_3), where slum morphologies deviate significantly from the slum features found in category C_1.

In Fig. 4(b)–(d), the achieved accuracies are split into each morphologic slum type. For the first category of morphologic slum types C_1, the $\text{XFCN}_{\text{city}}$ and the transfer learned $\text{XFCN}_{\text{LSP}}^{\text{TF}}$ are able to achieve high mean IoU accuracies, between 67.8% and 81.1% for both the five- and six-dimensional input data. Mapping slums of category C_1 from features learned from the dataset of the nine other cities (XFCN_{LSP}) result in lower mean IoU accuracies, of 52.6% for the five-dimensional data and 69.7% for the six-dimensional input data. The XFCN_{LSP} cannot generalize well to slums of category C_1 on the five-dimensional remote sensing data. The results from XFCN_{LSP} show a 17.1% improvement of the mean IoU when comparing five- and six-dimensional input data. This increase of the IoU score can be explained by the inclusion of the Open Street Map road network; training the XFCN on a variety of different slum categories, the road network offers a feature set that is found in all slum categories of C_{1-3}. The accuracies for the datasets of the slum category C_2 suffer from the highest variance throughout all three experiments in both the five- and six-dimensional input data. While the mean IoU accuracies for the $\text{XFCN}_{\text{city}}$ and

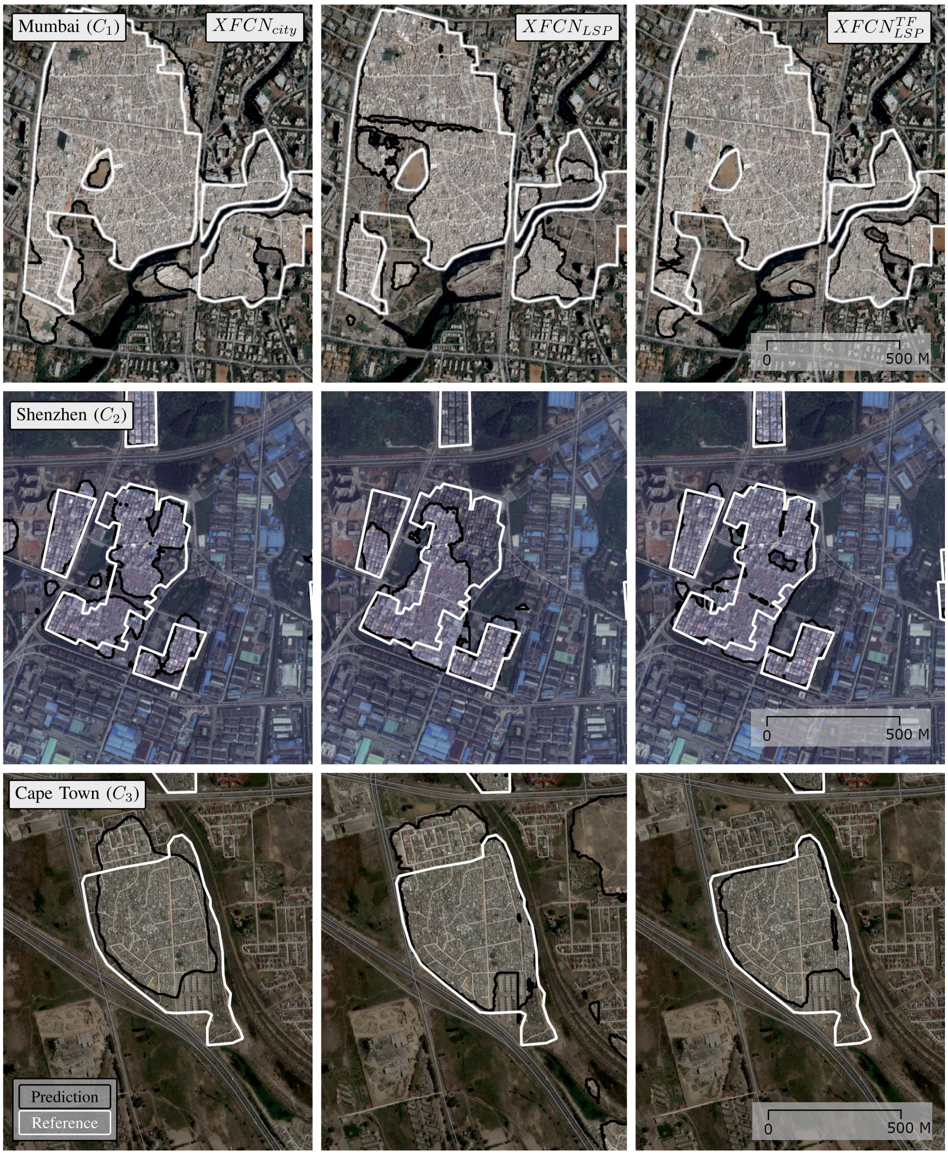

Fig. 5. Comparative alignment for three cities of each slum category (C_{1-3}). All results were trained on the six-dimensional input data. The left column shows the results for the XFCN$_{\text{city}}$, the middle column shows results from the XFCN$_{\text{LSP}}$ model, and finally, the right column shows the transfer learned XFCN$^{\text{TF}}_{\text{LSP}}$ results.

XFCN$^{\text{TF}}_{\text{LSP}}$ are the lowest of the three slum categories for the five-dimensional data with 57.9% and 66.8%, respectively, the highest overall IoU accuracies can be seen in XFCN$_{\text{LSP}}$ for the five-dimensional data and for the six-dimensional data. The XFCN$_{\text{LSP}}$ also achieves highest mean IoU accuracy, 74.86% for the six-dimensional input data, when comparing the three slum categories C_{1-3}. Consequently, the XFCN is able to robustly map slums of the category C_2 when it is previously trained on a large variety of slum morphologies. Although the slums of category C_3 deviate more significantly from the morphologic slum features found in C_{1-2}, it does not necessarily mean that the XFCN suffers from low mapping accuracies.

Based on the results in Table III, we can confirm that the XFCN is able to learn more robust representations of morphological slum features when it was previously trained on a large morphologic variety of slum morphologies and then transfer learned to a local domain dataset D^{city}_{T}. This is shown in a general increase of accuracies for the XFCN$^{\text{TF}}_{\text{LSP}}$ experiments. Slums are highly heterogeneous in nature, especially when comparing slum settlements on a global scale. While Table I can explain some differences of the general slum features, some are more complex to describe. Different morphologic slum types (C_{1-3}) can be seen in Fig. 5. Here, the mapped results for all three models, trained on the six-dimensional input data, can be depicted. The results in Mumbai (C_1) show that all three models XFCN$_{\text{city}}$, XFCN$_{\text{LSP}}$, and XFCN$^{\text{TF}}_{\text{LSP}}$ achieve an IoU score of over 66.32%. With 452 total slums, a mean slum size of $9.1[\text{ha}]$, and slum features of category C_1, slums can be mapped using all three XFCN models and only the XFCN$_{\text{LSP}}$ model suffers from some mild under classification. Results in Shenzhen (C_2) show similar effects as seen in Mumbai. With a dataset consisting of a large amount of slums, 1872, and a slum sample proportion of 22.7%, high IoU scores can be achieved in both $\text{XFCN}_{\text{city}}$ and $\text{XFCN}_{\text{LSP}}^{\text{TF}}$ with over 85.98%. The strength of transfer learning slum features form a large-scale poverty dataset to a small local dataset can be seen in the mapping results of Cape Town in Fig. 5. This dataset has a low amount of slums, 70, and only 2117 training patches. Thus, the $\text{XFCN}_{\text{city}}$ only achieves an IoU score of 72.28% and suffers from over and under classification. Only the transfer learned $\text{XFCN}_{\text{LSP}}^{\text{TF}}$ is able to differentiate better between the slums of category C_3 and the formal settlements.

In the cities with a lower IoU accuracy score of 65% (Delhi, Rio de Janeiro, and São Paulo), the XFCN struggles for various reasons. Slums of the morphologic category C_2 in Delhi and C_3 in Rio de Janeiro and São Paulo, in combination with the training datasets components (see Table I), indicate that these cities not only suffer from a small mean slum size of less than 6.5 ha and a slum sample proportion of less than 20%, but the slum settlements also share a certain similarity to formal settlements. This effect is also represented by a more regular road network in the slum settlements of in Rio de Janeiro and São Paulo. The accuracy scores for both cities are higher when the road network is not included in the training dataset. The highest accuracies could be reached in Mumbai and Shenzhen, where the training dataset in Table I provides a high number of slum patches and a large slum sample proportion, and the slum type morphologies of category C_1 and C_2 offer a stark difference between formal settlements and slums, as seen in Fig. 5. The big advantages of transfer learning to map slums could be observed in Caracas and Medellin (C_2), where the initial training dataset is quite small and, thus, training the XFCN from scratch is insufficient. Transferring poverty features learned from the large-scale poverty dataset to these cities could elevate to IoU from just under 48.9% to 69.8% in Medellin and from 59.1% to 80.7% in Caracas.

VII. Conclusion

Detecting urban poverty from remote sensing data is still a major challenge. It must deal with fuzzy feature spaces between formal and informal settlements, often with a significant imbalance of slum occurrences within the urban landscape and an inter- and intraurban variability of morphological slum features between different geographical regions. In this article, we propose a transfer-learned XFCN, which is trained on three experiments, testing whether it is possible to learn slum features in geographically separated regions. We have found that the success of transfer learning is not only dependent on the training dataset components, e.g., high slum sample percentage and a higher number of training patches, but also on the different slum morphologies. The combination of both the dataset and distinct slum morphology features are of importance to reach high mapping accuracies [Caracas, Mumbai, and Nairobi (C_1), Shenzhen (C_2), and Cape Town (C_3)]. In cases where the training dataset components are not ideal, the XFCN trained on various slum morphologies is able to match or surpass accuracies compared to training the XFCN within the same city. The best overall results were achieved when the XFCN was transfer learned from a large-scale poverty dataset to a smaller local dataset. Comparing the results from the five-dimensional input data, which consisted of only remote sensing data, and the six-dimensional data, where the proximity to the road network was added as an additional input dimension, accuracies improved segmentation outcomes in most cases. This shows that additional data can be of major importance to detecting urban poverty. Using more auxiliary data to accompany remote sensing data for mapping slums and novel deep learning architectures could potentially further increase accuracies; thus, data sources outside of remote sensing data could be used to make the decision process more robust during training to map slum settlements on a global scale.

Acknowledgment

The authors would like to thank J. Mast, who kindly provided the reference data for the city of Shenzhen, China.

References

[1] United Nations, *The Sustainable Development Goals Report*, 2019. [Online]. Available: https://unstats.un.org/sdgs/files/report/2017/thesustainabledevelopment goalsreport2017.pdf

[2] C. Tacoli, G. McGranahan, and D. Satterthwaite, "Urbanisation, rural-urban migration and urban poverty," IIED Working Paper, 2015.

[3] M. Kuffer, K. Pfeffer, and R. Sliuzas, "Slums from space—15 years of slum mapping using remote sensing," *Remote Sens.*, vol. 8, no. 6, May 2016, Art. no. 455.

[4] H. Taubenböck *et al.*, "A new ranking of the worlds largest cities—Do administrative units obscure morphological realities?" *Remote Sens. Environ.*, vol. 232, Oct. 2019, Art. no. 111353.

[5] UN-Habitat,"The challenge of slums: Global report on human settlements 2003," *Manage. Environ. Qual.: Int. J.*, vol. 15, no. 3, pp. 337–338, 2004.

[6] H. Taubenböck and N. J. Kraff, "The physical face of slums: A structural comparison of slums in Mumbai, India, based on remotely sensed data," *J. Housing Built Environ.*, vol. 29, no. 1, pp. 15–38, Feb. 2013.

[7] C. M. Gevaert, D. Kohli, and M. Kuffer, "Challenges of mapping the missing spaces," in *Proc. Joint Urban Remote Sens. Event*, May 2019, pp. 1–4.

[8] H. Taubenböck, N. Kraff, and M. Wurm, "The morphology of the arrival city—A global categorization based on literature surveys and remotely sensed data," *Appl. Geography*, vol. 92, pp. 150–167, Mar. 2018.

[9] M. Wurm and H. Taubenböck, "Detecting social groups from space—Assessment of remote sensing-based mapped morphological slums using income data," *Remote Sens. Lett.*, vol. 9, no. 1, pp. 41–50, Oct. 2017.

[10] M. Kuffer, F. Orina, R. Sliuzas, and H. Taubenbock, "Spatial patterns of slums: Comparing African and Asian cities," in *Proc. Joint Urban Remote Sens. Event*, Mar. 2017, pp. 1–4.

[11] M. Wurm, J. Goebel, G. G. Wagner, M. Weigand, S. Dech, and H. Taubenböck, "Inferring floor area ratio thresholds for the delineation of city centers based on cognitive perception," *Environ. Planning B: Urban Analytics City Sci.*, vol. 1067, pp. 1–19, Aug. 2019, doi: 10.1177/2399808319869341.

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jul. 2017, pp. 1800–1807.

[13] J. Hui, M. Du, X. Ye, Q. Qin, and J. Sui, "Effective building extraction from high-resolution remote sensing images with multitask driven deep neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 786–790, May 2019.

[14] K. Xu *et al.*, "Segmentation of building footprints with Xception and IoUloss," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2019, pp. 420–425.

[15] J. Friesen, H. Taubenböck, M. Wurm, and P. F. Pelz, "The similar size of slums," *Habitat Int.*, vol. 73, pp. 79–88, Mar. 2018.

[16] N. J. Kraff, H. Taubenbock, and M. Wurm, "How dynamic are slums? EO-based assessment of Kibera's morphologic transformation," in *Proc. Joint Urban Remote Sens. Event*, May 2019, pp. 1–4.

[17] M. Wurm, M. Weigand, A. Schmitt, C. Geiss, and H. Taubenbock, “Exploitation of textural and morphological image features in Sentinel-2A data for slum mapping,” in *Proc. Joint Urban Remote Sens. Event*, Mar. 2017, pp. 1–4.

[18] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, “Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 59–69, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271619300383

[19] M. Wurm, H. Taubenböck, M. Weigand, and A. Schmitt, “Slum mapping in polarimetric SAR data using spatial features,” *Remote Sens. Environ.*, vol. 194, pp. 190–204, Jun. 2017.

[20] R. Engstrom, D. Newhouse, V. Haldavanekar, A. Copenhaver, and J. Hersh, “Evaluating the relationship between spatial and spectral features derived from high spatial resolution satellite data and urban poverty in Colombo, Sri Lanka,” in *Proc. Joint Urban Remote Sens. Event*, Mar. 2017, pp. 1–4.

[21] R. Engstrom, J. Hersh, and D. Newhouse, “Poverty from space: Using high resolution satellite imagery for estimating economic well-being,” World Bank Policy Research, Working Paper 8284, 2016.

[22] M. R. Ibrahim, H. Titheridge, T. Cheng, and J. Haworth, “predictSLUMS: A new model for identifying and predicting informal settlements and slums in cities from street intersections using machine learning,” *Comput., Environ., Urban Syst.*, vol. 76, pp. 31–56, 2019.

[23] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, “Combining satellite imagery and machine learning to predict poverty,” *Science*, vol. 353, no. 6301, pp. 790–794, Aug. 2016.

[24] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, “Transfer learning from deep features for remote sensing and poverty mapping,” in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 3929–3935. [Online]. Available: http://dl.acm.org/citation.cfm?id=3016387.3016457

[25] B. J. Gram-Hansen *et al.*, “Mapping informal settlements in developing countries using machine learning and low resolution multi-spectral data,” in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2019, pp. 361–368.

[26] C. Persello and A. Stein, “Deep fully convolutional networks for the detection of informal settlements in VHR images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2325–2329, Dec. 2017.

[27] J. Mast, C. Wei, and M. Wurm, “Mapping urban villages using fully convolutional neural networks,” *Remote Sens. Lett.*, vol. 11, no. 7, pp. 630–639, May 2020.

[28] T. Stark, M. Wurm, H. Taubenböck, and X. X. Zhu, “Slum mapping in imbalanced remote sensing datasets using transfer learned deep features,” in *Proc. Joint Urban Remote Sens. Event*, May 2019, pp. 1–4.

[29] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[30] O. Ronneberger, P. Fischer, and T. Brox, “U-NET: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[31] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[32] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.

[33] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, 2015, pp. 1–9.

[34] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-ResNet and the impact of residual connections on learning,” in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–12.

[36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 770–778.

[37] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[38] L. Mou, Y. Hua, and X. X. Zhu, “A relation-augmented fully convolutional network for semantic segmentation in aerial scenes,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2019, pp. 12408–12417.

[39] Y. Shi, Q. Li, and X. X. Zhu, “Building segmentation through a gated graph convolutional neural network with deep structured feature embedding,” *J. Photogrammetry Remote Sens.*, vol. 159, pp. 184–197, Jan. 2020.

[40] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, and X. X. Zhu, “Buildings detection in VHR SAR images using fully convolution neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1100–1116, Feb. 2019.

[41] X. X. Zhu *et al.*, “Deep learning in remote sensing: A comprehensive review and list of resources,” *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

[43] T.-Y. Lin *et al.*, “Microsoft COCO: Common objects in context,” in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 740–755.

[44] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[45] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, “Semantic segmentation of aerial images with an ensemble of CNNs,” *ISPRS Ann. Photogrammetry, Remote Sens., Spatial Inf. Sci.*, vol. 3, pp. 473–480, 2016.

[46] N. Audebert, B. L. Saux, and S. Lefèvre, “Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks,” *J. Photogrammetry Remote Sens.*, vol. 140, pp. 20–32, Jun. 2018.

[47] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, “Classification with an edge: Improving semantic image segmentation with boundary detection,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 135, pp. 158–172, 2018.

[48] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, “BigearthNet: A large-scale benchmark archive for remote sensing image understanding,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5901–5904.

[49] C. Qiu, M. Schmitt, H. Taubenböck, and X. X. Zhu, “Mapping human settlements with multi-seasonal Sentinel-2 imagery and attention-based ResNeXt,” in *Proc. Joint Urban Remote Sens. Event*, May 2019, pp. 1–4.

[50] Planet Team, “Planet application program interface: In space for life on earth,” San Francisco, CA, USA, 2017. [Online]. Available: https://api.planet.com

[51] A. Jung, “imgaug,” GitHub Repository, 2017. [Online]. Available: https://github.com/aleju/imgaug

[52] D. Stiller, T. Stark, M. Wurm, S. Dech, and H. Taubenböck, “Large-scale building extraction in very high-resolution aerial imagery using mask r-CNN,” in *Proc. Joint Urban Remote Sens. Event*, May 2019, pp. 1–4.

[53] C. Robinson *et al.*, “Large scale high-resolution land cover mapping with multi-resolution data,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 12726–12735.

[54] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[55] K. S. Lee, “Tensorflow-xception,” GitHub Repository, 2017. [Online]. Available: https://github.com/kwotsin/TensorFlow-Xception

[56] S. Shekkizhar, “Fcn.tensorflow,” GitHub Repository, 2016. [Online]. Available: https://github.com/shekkizh/FCN.tensorflow

[57] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*.

[58] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” 2015, *arXiv:1502.03167*.

[59] A. Rosebrock, “Intersection over Union (IoU) for object detection,” pyimagesearch, 2016. [Online]. Available: https://www.pyimagesearch.com/

Thomas Stark received the M.Sc. degree in geodesy and geoinformation, in 2018 from the Technical University of Munich, Munich, Germany, where he is currently working toward the Ph.D. degree with the Chair of Signal Processing in Earth Observation, Department of Aerospace and Geodesy.

In 2017, he joined the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Weßling, Germany, as a Research Associate. His current research interests include urban remote sensing topics, with a focus on detecting urban poverty using machine learning methods.

Michael Wurm received the Diploma degree (Mag. rer. nat.) in geography with a specialization in remote sensing, GIS, and spatial research from the University of Graz, Graz, Austria, in 2007, and the Ph.D. degree (Dr. rer. nat.) in surveying and geoinformation from the Graz University of Technology, Graz, in 2013.

He was with the Institute of Digital Image Processing, Joanneum Research, Graz, in 2007. In 2008, he joined the University of Wurzburg, Germany, where he was engaged in interdisciplinary research between earth observation data and social sciences. Since 2011, he has been with the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Weßling, Germany, where he is involved in topics on urban geography, urban remote sensing, urban morphology, and slum mapping research. Since 2013, he has been a Lecturer with the University of Graz.

Xiao Xiang Zhu (Senior Member, IEEE) received the M.Sc. degree, the Dr.-Ing. degree, and the Habilitation degree in the field of signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is currently a Professor of Signal Processing in Earth Observation with the Technical University of Munich (TUM) and German Aerospace Center (DLR), Weßling, Germany; the Head of the Department EO Data Science with DLR's Earth Observation Center; and the Head of the Helmholtz Young Investigator Group SiPEO with DLR and TUM. Since 2019, she has been coordinating the Munich Data Science Research School. She is also leading the Helmholtz Artificial Intelligence Cooperation Unit (HAICU)—Research Field Aeronautics, Space and Transport. She was a Guest Scientist or Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, in 2009; Fudan University, Shanghai, China, in 2014; the University of Tokyo, Tokyo, Japan, in 2015; and the University of California, Los Angeles, CA, USA, in 2016. Her current research interests include remote sensing and earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of Young Academy (Junge Akademie/Junges Kolleg) with the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She is also an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.

Hannes Taubenböeck received the Diploma in geography from the Ludwig-Maximilians Universitt München, Munich, Germany, in 2004, and the Ph.D. degree (Dr.rer.nat.) in geography from the Julius Maximilian's University of Würzburg, Würzburg, Germany, in 2008.

In 2005, he joined the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Weßling, Germany. After a postdoctoral research phase with the University of Würzburg from 2007 to 2010, he returned in 2010 to DLR–DFD as a Scientific Employee. In 2013, he became the Head of the "City and Society" team. In 2019, he habilitated at the University of Würzburg in Geography. His current research interests include urban remote sensing topics, from the development of algorithms for information extraction to value adding to classification products for findings in urban geography.

A.3. Quantifying Uncertainty in Slum Detection: Advancing Transfer-Learning With Limited Data in Noisy Urban Environments

Reference: Stark, T., Wurm, M., Zhu, X. X., & Taubenbock, H. (2024). Quantifying uncertainty in slum detection: advancing transfer-learning with limited data in noisy urban environments. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Quantifying Uncertainty in Slum Detection: Advancing Transfer Learning With Limited Data in Noisy Urban Environments

Thomas Stark, Michael Wurm, Xiao Xiang Zhu, *Fellow, IEEE*, and Hannes Taubenböck

***Abstract*—In the intricate landscape of mapping urban slum dynamics, the significance of robust and efficient techniques is often underestimated and remains absent in many studies. This not only hampers the comprehensiveness of research but also undermines potential solutions that could be pivotal for addressing the complex challenges faced by these settlements. With this ethos in mind, we prioritize efficient methods to detect the complex urban morphologies of slum settlements. Leveraging transfer learning with minimal samples and estimating the probability of predictions for slum settlements, we uncover previously obscured patterns in urban structures. By using Monte Carlo dropout, we not only enhance classification performance in noisy datasets and ambiguous feature spaces but also gauge the uncertainty of our predictions. This offers deeper insights into the model's confidence in distinguishing slums, especially in scenarios where slums share characteristics with formal areas. Despite the inherent complexities, our custom CNN STnet stands out, delivering performance on par with renowned models like ResNet50 and Xception but with notably superior efficiency—faster training and inference, particularly with limited training samples. Combining Monte Carlo dropout, class-weighted loss function, and class-balanced transfer learning, we offer an efficient method to tackle the challenging task of classifying intricate urban patterns amidst noisy datasets. Our approach not only enhances artificial intelligence model training in noisy datasets but also advances our comprehension of slum dynamics, especially as these uncertainties shed light on the intricate intraurban variabilities of slum settlements.**

***Index Terms*—Imbalanced dataset, learning from few samples, noisy dataset, slum mapping, transfer learning, uncertainty estimation.**

I. Introduction

The criticality of data in artificial intelligence (AI), particularly in deep learning model development, are well-documented [1], [2], [3]. Quality datasets, free from biases and errors, are essential for creating algorithms that are generalizable and trustworthy for decision-making processes [4], [5], [6]. However, the prevalence of biases and inaccuracies in training datasets necessitates either thorough curation or specialized methods to handle these challenges. The advancement of AI architectures has been significant in addressing issues like imbalanced and noisy datasets, or classifying within fuzzy feature spaces [7], [8], [9]. These improvements are pivotal for handling the complexities and unpredictability of real-world scenarios, underscoring the importance of data quality in AI workflows for accurate and meaningful outcomes.

By leveraging AI, researchers can uncover hidden connections and gain a deeper understanding of complex phenomena, leading to more insightful studies and breakthrough discoveries. The constant evolution and improvement of AI architectures, especially in dealing with challenging datasets marked by imbalanced and noisy datasets [7], [8], or classifying within fuzzy feature spaces [9], has empowered researchers to handle diverse and unpredictable real-world scenarios effectively. As AI continues to progress, it brings the promise of more comprehensive and accurate solutions for the complexities of our dynamic world.

One area where AI has shown promising results is in remote sensing, particularly when it comes to understanding urban environments [10], [11]. This technology has been used to gather vast amounts of insightful data on cities, including information about population density [12], land use [13], or transportation patterns [14]. This also includes detecting urban poverty, where researchers and policymakers can gain valuable insights on locations of slum settlements. The utilization of high-resolution remote sensing imagery played a pivotal role in the comprehensive mapping of slums within the dynamic cityscape of Mumbai, as highlighted in [15]. Similarly, the city of Accra witnessed the integration of remote sensing data in conjunction with income data, facilitating an insightful mapping of poverty patterns, as seen in [16]. Furthermore, Kuffer et al. [17] conducted an intricate examination of the multifaceted factors that contribute to the enduring presence of slums, shedding light on their persistence within urban landscapes. Satellite imagery, population data, and economic indicators can help to recognize poverty patterns and map poverty levels to identify needy areas, enabling more focused and effective poverty reduction activities [18], [19], [20], [21].

Manuscript received 15 October 2023; revised 11 December 2023 and 18 January 2024; accepted 25 January 2024. Date of publication 29 January 2024; date of current version 15 February 2024. The work of X. Zhu is supported by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" under Grant 01DD20001. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under Grant 714087-So2Sat. *(Corresponding author: Hannes Taubenböck.)*

Thomas Stark and Xiao Xiang Zhu are with the Chair of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany (e-mail: t.stark@tum.de; xiaoxiang.zhu@tum.de).

Michael Wurm and Hannes Taubenböck are with German Remote Sensing Data Center, German Aerospace Center, 82234 Weßling, Germany (e-mail: michael.wurm@dlr.de; hannes.taubenboeck@dlr.de).

Digital Object Identifier 10.1109/JSTARS.2024.3359636

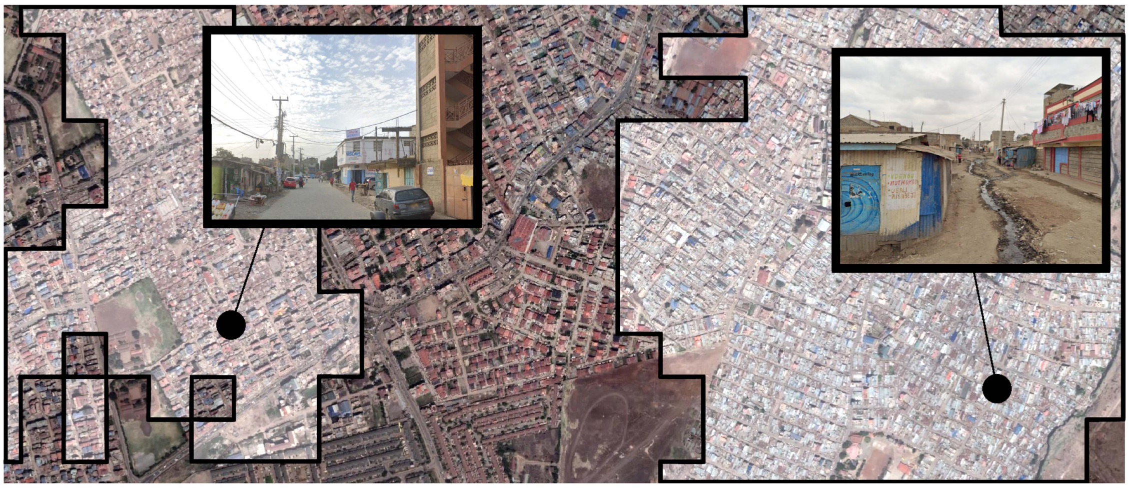

Fig. 1. Dense and low-rise areas shown with a black outline for the city of Nairobi [26]. Google Street View imagery is used to show that only some parts of the dense areas can also be considered a slum settlement highlighting the challenge of slum mapping.

Detecting urban poverty from remote sensing data is very challenging, due to data availability and the many different morphological features that can occur in slum settlements [22]. The issue of data availability is twofold: While some data exist for large and often studied areas [23], [24], for many cities of the Global South there are still very few data on slum settlements following a coherent and reproducible approach. The data that exist on slum settlements is often outdated, incomplete, and based on heterogeneous approaches on its definition regarding the morphology of slum settlements [24]. The second major challenge to detect slum settlements is the nature of its noisy feature space. Despite the fact that a typical morphological slum can be characterized by its high building density, small and complex street layouts, low-rise and small building structures, and use of a wide variety of construction materials, in reality slum settlements sometimes share just parts of these features [22]. Moreover, as indicated in [25], the delineation of slums is subject to variability owing to differing opinions on what constitutes a slum. Recognizing this, the data used for training an AI to classify slum settlements needs to diligently harmonized into a unified dataset to enhance a study's reliability, given that such variability in slum definitions could markedly affect the results. This subjectivity poses a challenge in classifying urban poverty. This impact can make it difficult to distinguish between a slum settlement and a formal built-up region. This challenge is depicted in Fig. 1 where the results from [26] show predictions of the local climate zone class seven, which is described as dense-low-rise buildings and shows two areas within the city of Nairobi, Kenya. While both highlighted areas display dense and low-rise building structures, only some parts of one highlighted area can be described as a slum upon having a closer look using Google Street View imagery. Thus, classifying a settlement as a slum cannot be solely determined by the previously mentioned features. Conversely, just because a settlement has a low-rise and dense structure does not automatically make it a slum. Similarly, the absence of density in a settlement does not guarantee that it cannot be classified as a slum. In other words, the combination of multiple morphological characteristics is a detrimental criterion for determining whether a settlement is a slum or not. While other factors, like plumbing and access to basic services need to be considered in evaluating the status of a settlement as well, these are not derivable from high-resolution remote sensing data. Thus, with the described noisiness of the dataset in mind, for the purpose of this research and considering the limitations in acquiring actual real ground-truth data, we rely here on the typical morphological appearance of slum settlements.

In our study, we focus on addressing two primary challenges: limited data availability and noisy datasets in the context of slum mapping using remote sensing data. Our main goal is to develop an efficient method for detecting slums with limited training samples, and to estimate the uncertainty in these predictions. To this end, we employ a transfer-learning approach, leveraging a large, imbalanced dataset to effectively train toward a smaller, balanced dataset. This method ensures that only a few samples are needed for successful slum detection. To tackle the issue of noisy datasets, we utilize Monte Carlo dropout. This technique allows us to approximate the uncertainty associated with predicting slum settlements, providing a more robust and reliable analysis. In addition, we introduce a custom convolutional neural network (CNN), the slum transfer network (STnet), specifically designed for high-resolution remote sensing data. STnet is engineered not only to enhance the training efficiency with a limited number of samples but also to offer significant improvements in processing time compared to standard CNN models. Our research aims to demonstrate the effectiveness of STnet in accurately detecting slums in various urban environments, thereby contributing to the broader field of urban studies and remote sensing.

II. Related Work

A. Detecting Urban Poverty Using Remote Sensing

Traditional machine learning approaches have already made significant contributions to the detection of urban poverty by enabling the analysis of large datasets [24]. These approaches have proven to be invaluable in providing researchers and policymakers with the necessary tools to gain a deeper understanding of poverty patterns within urban areas. By employing various machine learning algorithms, such as classification and regression models, researchers can process and analyze extensive datasets containing socioeconomic and spatial information [19].

A specific area where traditional machine learning has shown promise is in the application of remote sensing data for larger scale urban poverty detection [27], [28]. In the context of poverty detection, remote sensing data provide valuable information about the morphological patterns of slum settlements. This data can include features such as building density, land cover classification, and infrastructure characteristics [29], [30].

While traditional machine learning approaches have been effective in urban poverty detection, recent advancements in AI have further enhanced our ability to identify poverty using innovative techniques. AI, including deep learning models, has demonstrated remarkable capabilities in analyzing satellite imagery for poverty detection [15], [27], [31], [32], [33]. Deep learning algorithms, characterized by their ability to learn hierarchical representations of data, can automatically extract intricate visual features from satellite images, capturing subtle patterns that may indicate poverty.

However, despite these advancements, there is still a need for larger scale applications of poverty detection using AI. Most existing studies in this field are often limited to specific areas of interest within the same geographical region. To fully harness the potential of AI in urban poverty detection, it is essential to expand research efforts to encompass a broader range of urban environments worldwide. By doing so, it is intended to unlock the true power of AI in addressing the complex challenges associated with urban poverty on a global scale.

B. Training on Imbalanced Datasets

Studies revealed that slum morphologies in general consist of a small share of the built-up environment in cities, and in particular, mapping information is only scarcely if at all available [34]. Dealing with imbalanced datasets in deep learning involves several approaches that can help mitigate the issue of class imbalance. Some common methods include cost-sensitive learning, as seen in [35], which adjusts misclassification costs, favoring the minority class and improving overall performance on imbalanced datasets. Synthetic data generation increases the minority class representation by creating artificial samples, achieving a more balanced dataset and enhancing predictive accuracy [36], [37]. Using curriculum learning gradually exposes the algorithm to challenging examples, minimizing biases toward the majority class [38], [39].

Another simple approach is resampling the dataset by either oversampling the minority class or undersampling the majority class. Oversampling involves replicating or generating new instances from the minority class to balance the dataset. Undersampling reduces the majority class to match the minority class. Both approaches help achieve a more balanced class distribution and improve AI model performance [33], [40], [41]. The choice between them depends on the dataset and learning algorithm used.

Furthermore, class weight adjustment is a technique in AI used to tackle imbalanced datasets. By assigning higher weights to the minority class during training, the model places greater emphasis on learning from the minority class. This helps to address the issue of class imbalance and ensures that the model pays more attention to the minority class, improving its ability to correctly classify instances from that class. By adjusting the class weights, the model becomes more sensitive to the minority class and achieves a better balance in handling imbalanced datasets [42], [43].

It is important to carefully evaluate the performance of the model after implementing these methods to ensure that the imbalance has been effectively addressed without negatively impacting the overall performance. In this work, we direct our attention toward a class-weighted loss function for pretraining and for transfer learning an undersampling method. Both present a straightforward and efficient workflow that can be effortlessly replicated. By choosing to focus on these specific methods, we aim to harness their advantage and capitalize on their ease of implementation.

C. Transfer Learning From Few Samples

Transfer learning a CNN involves adjusting the weights of an already trained model to fit the specific task or dataset in the target domain. This is achieved by pretraining a model and retraining it with a smaller learning rate on a related classification task for the target domain [44]. The benefits of transfer learning a CNN include: faster training times as the model has already learned useful features from the pretraining data [45], [46], improved performance on the target task as compared to training a model from scratch, and the ability to leverage the knowledge gained by the pretrained model on a large dataset to improve the performance on a smaller dataset [47], [48].

When it comes to transfer learning with few samples, the situation is similar to few-shot learning techniques. However, in transfer learning, the focus is not solely on handling a few labeled examples of a new task. Instead, transfer learning aims to exploit the knowledge learned from a source task with sufficient labeled data and apply that knowledge to a target task with limited labeled data. Whereas in few-shot learning, the model is trained to learn from none or very few labeled samples. In [9], a few-shot learning technique from [49] was used in order detect complex morphologies representing poor areas within the urban environment, the authors found out that the technique works very well when only a hand-full of samples are available. Other approaches have been using self-supervised embedding optimization for adaptive generalization in urban settings [50] or using prototypical networks for urban damage detection after natural hazards [51].

D. Bayesian Uncertainty Estimation

In deep learning, Bayesian uncertainty refers to the incorporation of probabilistic inference into neural networks and can be categorized into two domains: epistemic and aleatoric.

The former, epistemic uncertainty, pertains to the uncertainty associated with the model parameters or weights, while the latter, aleatoric uncertainty, is commonly associated with data uncertainty.

Variational inference, which models the network's weights as probability distributions and employs optimization techniques to approximate them [52], [53], and Bayesian neural networks, which treat the network's parameters as random variables and infer posterior distributions [54], [55], offer valuable insights into model uncertainty. These approaches can significantly assist in improving the trustworthiness of deep learning methods in remote sensing tasks [56], [57], [58].

Monte Carlo dropout is another method used for uncertainty estimation in predictive models. It leverages dropout to approximate Bayesian inference for deep neural networks by performing multiple forward passes with dropout during inference [59]. Each pass generates different predictions, allowing for the calculation of prediction variance and capturing the inherent epistemic uncertainty in the model's output. It can also be used to prevent overfitting [60]. This technique has been applied successfully in various domains, such as computer vision [61], [62], natural language processing [63], and healthcare [64].

One of the key benefits of Monte Carlo dropout is its potential to enhance prediction interpretability [65]. By generating multiple predictions with dropout, the method provides a probabilistic distribution of possible outcomes, enabling a more comprehensive understanding of the model's uncertainty. This distribution can be visualized and analyzed to gain insights into the factors influencing the model's decisions. Monte Carlo dropout has found applications in a wide range of tasks. Uncertainty estimation helps to identify ambiguous regions in image classification tasks, or it can guide the system to seek clarification or avoid providing incorrect or misleading information. Moreover, Monte Carlo dropout has been utilized to understand the level of confidence in the model's predictions and assisting in making informed decisions [66].

III. Methodology

A. Convolutional Neural Networks

ResNet-50 [67] and Xception [68] are two widely acclaimed and standard CNNs that find extensive usage in various scientific domains, including remote sensing image classification tasks. ResNet-50, short for residual network with 50 layers, revolutionized the field of deep learning by introducing residual connections that mitigate the vanishing gradient problem and enable the training of extremely deep networks. This architecture facilitates the construction of deeper models, leading to improved accuracy in image classification tasks. On the other hand, Xception, an extension of the Inception architecture, takes the concept of depth-wise separable convolutions to an extreme level. It separates the spatial and channel-wise convolutions, reducing the computational cost significantly while maintaining high performance.

In our study we use both, ResNet-50 and Xception in order to introduce our Slum Transfer network (STnet), a custom CNN specifically designed to excel in processing high-resolution remote sensing imagery. The STnet is a heavily customized Xception network [68] and a simplified schematic can be seen in Fig. 2. The entry flow consists of five convolution combinations using residual skip connections. In order to capture a larger area when using high-resolution remote sensing imagery, the first two 2-D convolutions use large 9x9 kernels. In the middle flow, feature pyramid pooling is used to provide a unified framework to extract features at different scales. Finally, the classification flow is composed of two linear functions. Throughout the whole STnet, a combination of batch normalization and dropout layers afterwards are used. In total, STnet has 22 layers and 3.3 million trainable parameters.

B. Transfer Learning

The learning strategy employed in this procedure can be divided into two distinct phases. In the initial phase, the STnet undergoes pretraining on a class-imbalanced dataset denoted as $\mathcal{D}_{\text{base}}$. To address the class imbalance during this stage, we employ a weighted loss, as illustrated in (1), to give due importance to underrepresented classes and make the most of the available data. Subsequently, the STnet is transfer learned using an additional dataset, referred to as $\mathcal{D}^{\text{bal}}_{\text{loocv}}$. However, one of the classes in $\mathcal{D}_{\text{base}}$ is significantly imbalanced compared to the others, while in $\mathcal{D}^{\text{bal}}_{\text{loocv}}$, a class balanced dataset is created using undersampling. $\mathcal{D}^{\text{bal}}_{\text{loocv}}$ is designed to be class balanced, meaning it contains an equal number of images from all classes. By ensuring that each class is represented equally in $\mathcal{D}^{\text{bal}}_{\text{loocv}}$, we mitigate the bias toward the imbalanced class from $\mathcal{D}_{\text{base}}$. This balanced dataset allows for a fair and unbiased transfer-learning process, as each class contributes equally to the training of the new classifier.

During pretraining and transfer learning, we use a class weighted cross entropy loss L as seen in (1) where w_i is the weight for each class, scaled by the inverted count of the class occurrence

$$L(x, c, w) = -\sum_i w_i \cdot y'_i \cdot \log\left(\frac{\exp(x_i)}{\sum_j \exp(x_j)}\right)$$

$$\begin{aligned} &\text{where} \\ &\quad w_i : \text{weight for class } i; \\ &\quad y'_i : \text{target distribution after} \\ &\quad\quad \text{label smoothing for class } i; \\ &\quad x_i : \text{logit for class } i. \end{aligned} \tag{1}$$

During transfer learning, the complete CNN remains trainable, and no layers are frozen. This means that all the layers of the pretrained CNN, are trained using the $\mathcal{D}_{\text{loocv}}$ dataset. By keeping all layers trainable, the CNN can adapt its learned features to the new dataset while still benefiting from the knowledge gained on the base dataset. This approach allows the CNN to capture task-specific features from $\mathcal{D}^{\text{bal}}_{\text{loocv}}$ while retaining the general knowledge acquired from $\mathcal{D}_{\text{base}}$.

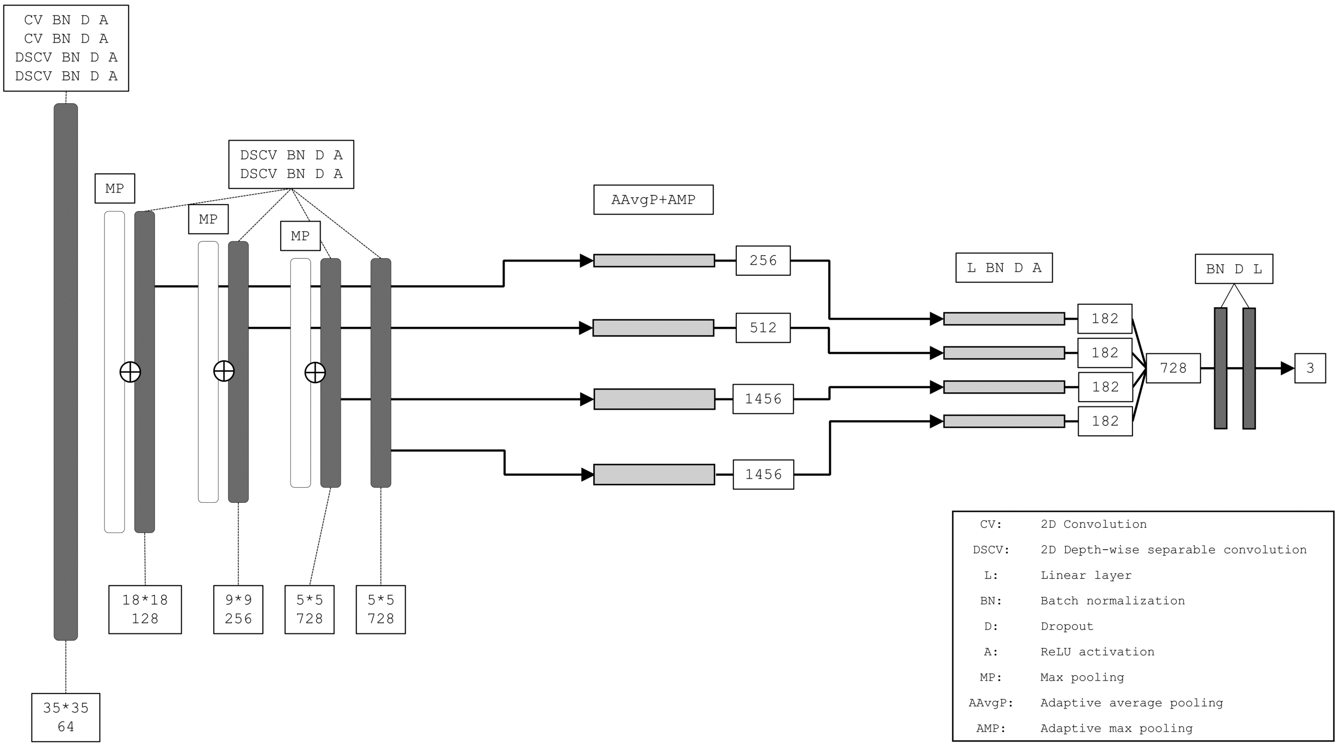

Fig. 2. Simplified schematic of the STnet architecture, comprising five convolutional variants in the entry -flow, succeeded by feature pyramid pooling layers and a classification- flow in the end. This light-weight architecture encompasses 3.3 million trainable parameters.

C. Monte Carlo Dropout for Uncertainty Estimation

In our classification setting, we have a dataset $D(X, Y)$, where $X = x_1, x_2, \ldots, x_n$ represents the records of input images and $Y = y_1, y_2, \ldots, y_n$ denotes the corresponding reference labels. We employ our STnet model to predict new outputs $\bar{y}$ from new data $\hat{x}$. The model's predictions rely on a set of weights, and the task at hand involves finding the optimal set of these weights through an optimization problem

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y^{(t)}$$

where

$$\bar{y} : \text{Averaged prediction over Monte Carlo runs;}$$

$$T : \text{Total number of Monte Carlo runs;}$$

$$y^{(t)} : \text{Prediction for the } t\text{th forward pass.} \qquad (2)$$

To incorporate the Monte Carlo dropout technique as seen in (2), we use a probability $p = 0.3$ for each dropout layer, and for each model in all our experiments. This decision was informed by the preliminary test with $p = 0.1$, $p = 0.3$, and $p = 0.5$, where $p = 0.3$ offered the most effective balance between the Monte Carlo probabilities and the accuracies of the models.

During the forward pass, a unit is dropped and set to zero if its corresponding binary variable is zero. By utilizing Monte Carlo dropout, we aim to model the distribution, and subsequently, the predictive posterior distribution of $\bar{y}$. Notably, we can achieve this by training the neural network as if it were a typical network, with the inclusion of dropout layers after each layer with weight parameters and performing T predictions.

In summary, unlike the conventional classification setting where a single prediction $y^{(t)}$ is obtained, the Monte Carlo dropout technique allows us to model a predictive distribution. This approach entails training the network with dropout layers and making multiple predictions, resembling the training process of a standard neural network with slight modifications.

IV. Data and Experimental Setup

A. Dataset

The remote sensing data used in this study were acquired using PlanetScope satellites during 2021. In total, 8-bit RGB data were used and all scenes were resampled to 3-m resolution per pixel. Data from eight cities of the Global South were collected including Cape Town, Caracas, Lagos, Medellin, Mumbai, Nairobi, Rio de Janeiro, and Sao Paulo. The division of the remote sensing data into $88 * 88$ pixel patches (equivalent to 264 m $*$ 264˜m) was methodically chosen based on empirical evidence from previous studies in the domain of learning with few samples, which demonstrated the efficacy of this specific patch size [9], [49].

Our dataset consists of three target classes: zero background,one formal built-up areas, and two slums. The formal built-up areas were derived by using data from the LCZ42 dataset [26]. Reference data for the slum settlements were created by mapping polygons from experts in the field of remote

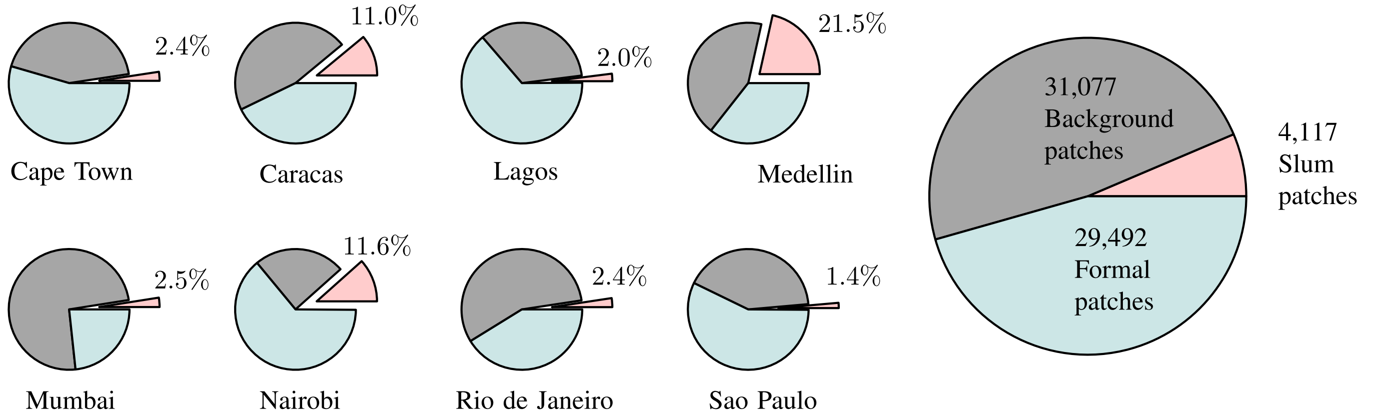

Fig. 3. Class distribution in eight cities and combined distribution. The figure displays nine pie charts depicting the class distribution in eight cities, with the slum sample proportion highlighted for each city. The final pie chart showcases the combined distribution, illustrating the overall class proportions across all cities.

sensing and urban poverty on the basis of up to date aerial imagery using Google Earth. To ensure data consistency for the reference data gathered from all sources, all polygons were checked and if necessary adjusted by the authors.

Each image patch used for training and testing the AI model has a dimension of $88 * 88 * 3$, with each label patch is $88 * 88 * n_{\text{cl}}$, where $n_{\text{cl}} = 3$ for the three classes used. If a reference patch contains at least 25% pixels of slum settlements, it is considered toward the slum class, patches with less than 25% but containing at least one slum pixel are discarded during training the model. For all other samples, the class with the highest pixel tally is considered as the main class. In total, 64 686 samples are available in the $\mathcal{D}_{\text{original}}$ dataset used for training and testing our approach as seen in Fig. 3.

B. Data Sampling

We define $\mathcal{D}_{\text{original}}$ in (3) as the set of ordered pairs, where each pair consists of an image X and CL as its corresponding city's location. X_n is the nth image in the set and CL_i as the location of the nth image, where i ranges between the city's location ID from 1 to 8. This dataset contains a wide range of diverse samples, encompassing various morphologies of urban patterns relevant to our topic.

For all experiments, we use a leave-one-out cross-validation approach, which is instrumental in ensuring comprehensive model evaluation and robustness across diverse urban environments, reflecting the variability in slum morphologies. This method also effectively mitigates the risk of overfitting, ensuring the model's adaptability and generalizability to different geographical contexts, crucial for the real-world application of urban poverty analysis and slum mapping. The image patches from seven of the eight cities are used for training and validation, while the remaining city's dataset is used for testing and transfer-learning. This process is repeated for all eight cities creating eight pretrained models to use for the test datasets. We partitioned $\mathcal{D}_{\text{original}}$ into two distinct datasets as seen in (4). The first subset, named $\mathcal{D}_{\text{base}}$, was employed for pretraining the STnet. $\mathcal{D}_{\text{base}}$ served as the foundation for training the initial weights and learning representations, the dataset always consist of seven cities of the dataset as seen in (5). The second subset, called $\mathcal{D}_{\text{loocv}}$ in (6), was dedicated to the transfer-learning phase. By using a leave-one-out cross-validation dataset $\mathcal{D}_{\text{loocv}}$, we were able to refine and optimize the STnet's performance, ensuring its adaptability and robustness. Overall, the division of $\mathcal{D}_{\text{original}}$ into $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\text{loocv}}$ played a crucial role in our research, enabling us to achieve accurate and reliable results.

During the transfer-learning phase, the dataset $\mathcal{D}_{\text{loocv}}$ is turned into a class balanced dataset $\mathcal{D}^{\text{bal}}_{\text{loocv}}$, using undersampling of the majority class. In (7) X_n is the nth image in the dataset with its corresponding label Y_n. We count the occurrence of all classes c and randomly sample j patches used for transfer learning.

$$\mathcal{D}_{\text{original}} = \{(X_1, \text{CL}_1), (X_2, \text{CL}_2), \ldots, (X_n, \text{CL}_i)\} \tag{3}$$

$$\mathcal{D}_{\text{original}} = \mathcal{D}_{\text{base}} \cup \mathcal{D}_{\text{loocv}} \tag{4}$$

$$\mathcal{D}_{\text{base}} = \{(X_1, \text{CL}_i), \ldots, (X_n, \text{CL}_i)\} \in \mathcal{D}_{\text{original}} \tag{5}$$
$$\text{CL}_i \neq \text{loocv}$$

$$\mathcal{D}_{\text{loocv}} = \{(X_1, \text{CL}_i), \ldots, (X_n, \text{CL}_i)\} \in \mathcal{D}_{\text{original}}$$
$$\text{CL}_i = \text{loocv} \tag{6}$$

$$\mathcal{D}^{\text{bal}}_{\text{loocv}} = \Big\{ (X_n, Y_n) \;\Big|\; \text{count}(Y_n \in \mathcal{D}_{\text{loocv}}, Y_n = c_1) = j,$$
$$\text{count}(Y_n \in \mathcal{D}_{\text{loocv}}, Y_n = c_2) = j,$$
$$\text{count}(Y_n \in \mathcal{D}_{\text{loocv}}, Y_n = c_3) = j \Big\}. \tag{7}$$

To evaluate the number of image patches required for transfer learning, we examine 1, 5, 10, 25, 50, and 100 image samples per class. For each experiment, we randomly select these samples per class from $\mathcal{D}_{\text{loocv}}$, and use the remaining city, not included in the training dataset, for transfer learning. The samples chosen for transfer learning are subsequently eliminated from the test dataset. This process ensures that our experiments avoid bias and accurately reflect the model's capability to generalize from limited data. In order the address the effects of randomly choosing image samples, we averaged the outcomes of five differently seeded experiments and report the standard deviations in our results, highlighting the impact of sample selections on the model's performance. In order to guarantee that there are sufficient samples of each class, particularly the slum class, 100 samples were the maximum number of samples required to verify our transfer-learning strategy.

C. Experimental Setup

To examine the impact of transfer learning on noisy datasets, we follow the setup outlined as follows. In all experiments, we warm up the optimizer for three epochs with a learning rate of $1e-8$. For pretraining, we use a learning rate of $1e-3$ and for transfer learning $1e-4$. All experiments use an Adam optimizer and weighted soft cross entropy loss. In addition, a batch size of 16 is used for training. To tackle both, dataset noise and model prediction uncertainty, we employ Monte Carlo dropout. This technique involves obtaining an average of 25 outputs from the model's predictions. We compute the average of the raw logits produced by the model and calculate the corresponding entropy value in order to compare the level of uncertainty of the prediction.

In our evaluation framework, it is important to note that while our models were trained on three classes to effectively manage class (im-)balance, the accuracy metrics reported specifically pertain to the slum class. This focused approach is due to our primary interest in slum mapping. Classes representing background and urban/formal built-up areas are not included in the accuracy assessment. Therefore, in assessing performance, we use three commonly used metrics for image classification problems, namely the F1-score, precision, and recall, as our primary metrics to gauge the effectiveness of our models in accurately identifying slum areas. To further compare the efficiency of different models, we analyze the training time required for each. In addition, we assess the influence of Monte Carlo steps on our results, examining how variations in this parameter impact the models' stability and inference time. By integrating both performance metrics and computational efficiency measures, we ensure a thorough evaluation that guides our decision making process and optimizes the overall quality of our outcomes.

A fundamental challenge in the context of transfer learning is the variability in model performance when using a limited number of samples. This variability arises due to the random selection of training samples, leading to potential sample selection bias. To obtain a comprehensive understanding of model performance and address the issues arising from outlier data training, it is imperative to employ a rigorous approach. Specifically, we conduct five seeded runs to effectively assess the models' capabilities. By averaging the results obtained from these diverse seeded runs, we obtain a robust estimation of model performance, which allows for a more accurate representation of sample selection bias. This approach aids in reducing the impact of random fluctuations, providing a clearer picture of the model's general performance across varying training data subsets.

TABLE I

Results for Eight Cities Comparing Different Number of Samples Used for Transfer Learning the STnet, Including the Standard Deviation for Five Seeded Runs

Samples	F1	Precision	Recall
Inference	0.7201 ± .11	0.6941 ± .10	0.7788 ± .19
1	0.7324 ± .09	0.7312 ± .11	0.7706 ± .17
5	0.7806 ± .05	0.7626 ± .06	0.8091 ± .10
10	0.8082 ± .05	0.7830 ± .06	0.8500 ± .09
25	0.8358 ± .04	0.8250 ± .05	0.8541 ± .08
50	0.8432 ± .05	0.8558 ± .07	0.8447 ± .10
100	**0.8624** ± .05	**0.8718** ± .06	**0.8638** ± .10

V. Results

A. Transfer-Learning Results

The results of the transfer-learned STnet reveal an empirical relationship between the number of samples per class used for transfer-learning and the corresponding F1-score as seen in Table I. Notably, an increase in the number of samples yielded improved F1-scores. However, it is noteworthy that even with just a single sample per class, the model achieved commendable F1-scores of 73.24%. Nevertheless, after 50 samples, the F1-score seems to plateau, suggesting an upper limit of high 80% F1-score for this classification task. These findings indicate the potential for achieving favorable F1-score with STnet, even when training data are scarce. The highest F1-score of 86.24% was achieved when using 100 samples per class for transfer-learning.

In addition, when examining the precision and recall values in Table I of the transfer-learned STnet, notable patterns emerge. While the precision values increases more drastically as the number of samples for transfer learning increases, the recall values, however, only steadily increases. These results underscore the effectiveness of the transfer-learning approach in refining the model's precision and recall, leading to improved overall performance and indicating the potential of STnet in applications with limited training data.

Fig. 4 depicts the F1-scores for eight cities and the corresponding number of samples used to transfer learn our STnet model. The general trend observed in the figure indicates that as the number of samples per class used for transfer learning increases, the F1-scores also increase. The experiment was conducted five times, with each transfer-learning approach utilizing different random samples. The error band in Fig. 4 from the five runs uses confidence intervals of 95% to draw around estimated values.

In Cape Town (93.60%), Caracas (90.62%), and Medellin (91.09%), we achieve the highest F1-scores, when using 100 samples for transfer learning. But it needs to be noted that in Medellin and Caracas, we already achieve high accuracies using simple inference of over 82.10%. We also observe a decrease in F1-score when only one sample per class is used for transfer learning, in Caracas, Medellin, Lagos, and Mumbai, indicating a more challenging setting for transfer learning. But even in

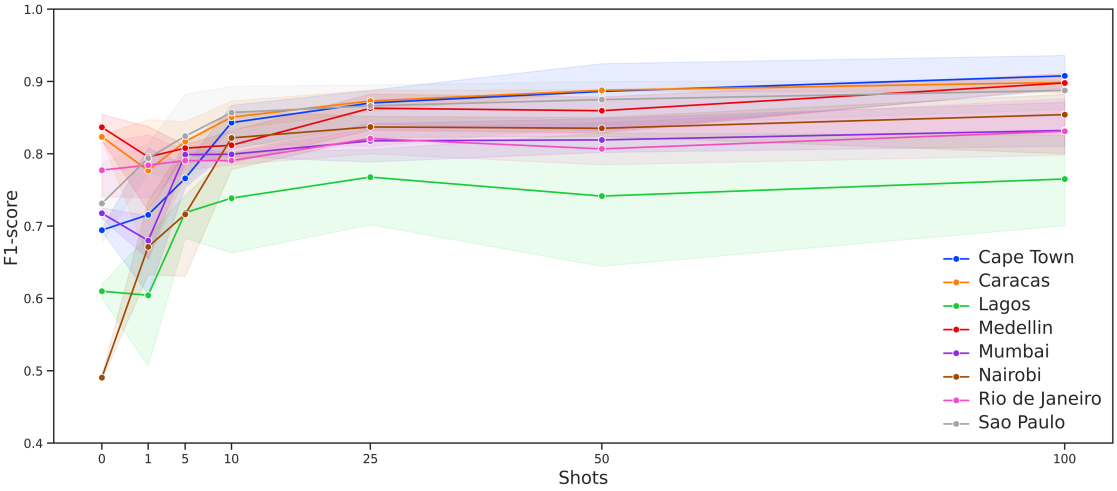

Fig. 4. F1-score for eight cities using a variety of number of samples to transfer learn the STnet pretrained using a transfer-learning approach.

TABLE II
COMPARISON OF F1-SCORES FOR STNET, XCEPTION, AND RESNET50, AVERAGED OVER FIVE DIFFERENTLY SEEDED RUNS SHOWN WITH ITS STANDARD DEVIATIONS

Samples	F1-score		
	STnet	Xcpetion	ResNet50
Inference	**0.7201** ± .11	0.6724 ± .16	0.7150 ± .09
1	**0.7324** ± .09	0.7309 ± .08	0.7010 ± .12
5	**0.7806** ± .05	0.7804 ± .05	0.7811 ± .05
10	**0.8082** ± .05	0.8031 ± .06	0.7941 ± .04
25	0.8358 ± .04	0.8380 ± .06	**0.8553** ± .04
50	0.8432 ± .05	0.8502 ± .04	**0.8709** ± .04
100	0.8624 ± .05	0.8775 ± .05	**0.8957** ± .02

All models employed 25 Monte Carlo iterations.

TABLE III
COMPARING DIFFERENT CNN ARCHITECTURES TO EACH OTHER BASED ON THEIR SIZE AND TRAINING TIME

	Parameters	Training time	
		per step	Total time
STnet	3.3m	33.25it/sec.	56:34 36 epochs
Xception	20.8m	21.28it/sec.	1:22:45 24 epochs
ResNet50	23.5m	17.85it/sec.	1:15:37 20 epochs

All models employed 25 Monte Carlo iterations.

Lagos and Mumbai, using only five samples per class results in a major improvement of roughly 10% in F1-score compared to simple inference.

B. *Comparing Various CNNs*

In Table II, we conduct a comprehensive comparative analysis of the F1-scores for three distinct CNNs: STnet, Xception, and ResNet50. This comparison spans a range of scenarios in transfer learning, starting from simple inference results to the use of 1–100 image samples in transfer-learning processes. Each model's F1-score is calculated as an average across five independently seeded runs, and we provide the standard deviations to illustrate the variability in performance. In addition, all models were subjected to 25 Monte Carlo iterations to ensure consistency in our evaluation methodology. Our analysis reveals that STnet, despite having a considerably lower parameter count of only 3.3 million, achieves performance metrics that are comparable to those of Xception and ResNet-50, which are significantly more parameter intensive. Notably, in scenarios where only a limited number of samples are employed for transfer learning, STnet demonstrates superior performance, outscoring both Xception and ResNet-50.

Table III provides a detailed comparison of the training times for each step and the total time required to achieve the best validation metric for the three models. Although the overall F1-scores of these CNNs are relatively similar, a significant difference is observed in their training durations. This discrepancy is largely attributed to STnet's more streamlined architecture, which makes it considerably lighter and faster in processing compared to Xception and ResNet-50, as evidenced in Table III. Notably, STnet not only demonstrates faster processing times but also requires less total time to attain the optimal validation metric. However, it is important to note that despite its shorter overall training duration, STnet demands more epochs to reach the best model fitness, in contrast to Xception and ResNet-50. This aspect highlights the efficiency of STnet in terms of time management.

C. *Comparing Monte Carlo Dropout Rates*

In Table IV, we investigate the impact of varying the number of Monte Carlo dropout test runs on our STnet model. The performance of the model is evaluated using inference time, F1-score, and finally, the entropy value, which is a measure of uncertainty or randomness within the predicted distributions

TABLE IV
COMPARISON OF STNET'S INFERENCE TIME, F1-SCORE, AND ENTROPY ACROSS DIFFERENT MONTE CARLO DROPOUT ITERATIONS, TRAINED ON 100 SAMPLES PER CLASS

Monte Carlo iterations	Inference time	F1-score	Entropy
1	31.59 it/sec. 2:11 min	0.8423	–
5	20.43 it/sec. 3:20 min	0.8515	0.7845
25	5.68 it/sec. 12:02 min	0.8624	0.7832
50	3.10 it/sec. 22:17 min	0.8679	0.7810

generated by the Monte Carlo dropout technique. Specifically, we compare the results obtained from using 1, 5, 25, and 50 Monte Carlo dropout test runs. In [69], 50 iterations are mentioned when using Monte Carlo dropout, but they only used a dropout layer in the last layer of their CNN. Since STnet uses dropout throughout its complete architecture, we test iterations of up to 50.

Our findings reveal that increasing the number of Monte Carlo dropout test runs leads to a slight improvement in the F1-score. Furthermore, we observe a slight decrease in the entropy values as the number of Monte Carlo dropout test runs increases, which implies that the predictions become more focused and certain as more test runs are performed. Significant observations were made regarding the inference time when increasing the number of Monte Carlo dropout iterations. The results indicate a substantial increase in inference time, with a 275% rise observed when transitioning from five Monte Carlo dropout iterations to 25, followed by an additional 84% increase when reaching 50 iterations. Despite the availability of insightful uncertainty measurements with just five iterations, the experiments conducted in this study employed 25 iterations as the preferred configuration for analysis.

VI. DISCUSSION

A. *Uncertainty of Slums*

Fig. 5 shows the results obtained for all cities using the transfer-learned STnet with 100 samples per class. The incorporation of Monte Carlo dropout as a method for uncertainty estimation unveils a significant advantage. It allows us to discern the STnet's level of certainty in predicting the location of slum settlements and identifies cases where its predictions are inconclusive. This not only provides crucial insights into the decision-making process of the STnet but also sheds light on the inherent challenges associated with the classification of slum areas.

The analysis demonstrates that the STnet exhibits elevated confidence in predicting the presence of typical slum settlements, characterized by typical morphologic slum features, including high density, heterogeneous building patterns, and irregular road shapes. This pattern is evident in cities that achieve high F1-scores, namely Cape Town, Caracas, Medellin, and Mumbai.

In addressing the challenges faced in slum classification within specific cities, it is observed that Lagos presents notable difficulties with underclassification of slums. Conversely, in Nairobi, Rio de Janeiro, and Sao Paulo, the primary challenge lies in overclassification. These issues are largely due to two key factors. The first factor is the absence of distinct morphological features typically found in slum settlements, which are otherwise noticeable in cities like Caracas and Medellin. The second factor contributing to these classification challenges is the presence of formal settlement structures that share similarities with slum areas in terms of density and low-rise characteristics. This overlap in physical attributes complicates the task of clearly differentiating between formal and slum classes in these urban environments.

This highlights the complexity of slum classification due to local morphologic specifics in relation to the surrounding built-up morphologies as well as it emphasizes the importance of taking into account differences in morphological characteristics present within slums. Moreover, in fringe regions, where slum settlements are intertwined with urban formal settlements, vegetation areas, or both, higher uncertainties are observed.

In regards to assessing the uncertainty of slums and their prediction, evaluating the chosen dropout value during training and Monte Carlo inference becomes a crucial aspect of our methodology. The decision to implement a 30% dropout rate was a strategic one, aimed at striking an optimal balance in our models. This rate was selected after observing effects of different dropout rates in preliminary tests. At a lower 10% dropout rate, we noticed less variability in uncertainties, but this did not significantly enhance the models' accuracies. On the other hand, a higher dropout rate of 50% adversely impacted the models' accuracies, suggesting a potential overadjustment in the learning process.

This understanding of the impacts of varying dropout rates was important in optimizing the performance of our models. By settling on a 30% dropout rate, we managed to maintain a balance where the accuracy of the models was not overly compromised, nor was the effectiveness of the Monte Carlo estimation diluted. This decision was crucial in ensuring that our models remained robust and efficient in predicting slum areas.

It is crucial to acknowledge the influence of the inherently noisy dataset on our results. Although we have unified the dataset into a coherent representation of slums, as detailed earlier in this article, the intra- and interurban variability inherently introduces a significant level of noise. This variability means there is a wide range of slum characteristics to learn and predict. However, this diversity also serves as a key advantage of employing the Monte Carlo Dropout method. By using this technique, we can observe the effects of this variability in the probabilities, which is also evident in the maps presented in Fig. 6.

Furthermore, it is important to consider how the application of our models to different definitions of morphological slums

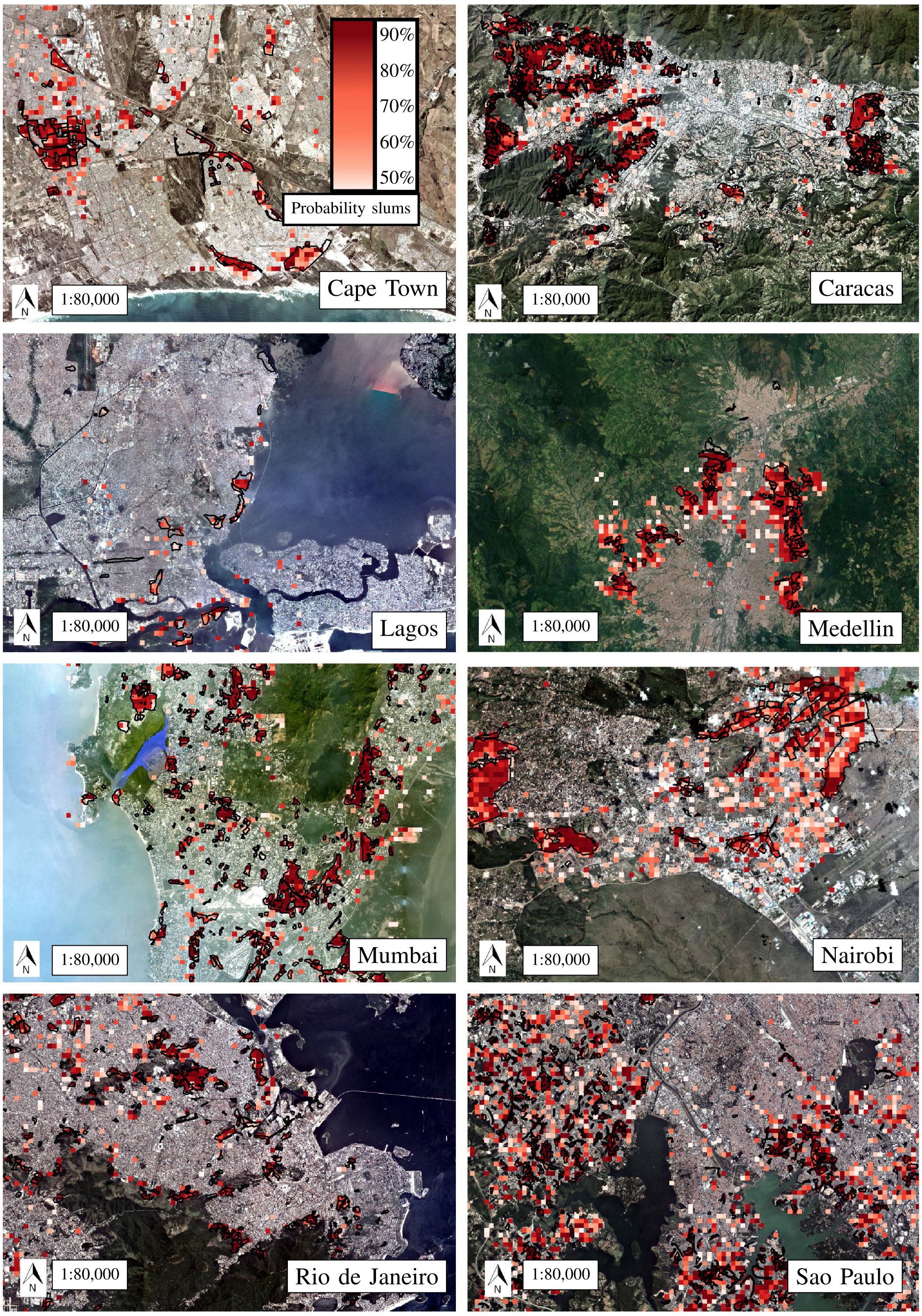

Fig. 5. Results for all eight cities using the transfer-learned STnet trained on 100 samples per class. All results are in the same scale of 1:80 000 and use the same color bar for the probability value of the slum class. Black outlines are used for the reference slum polygons.

could impact the results. Slums can vary greatly in their physical characteristics, spatial distributions, and overall appearances from one urban area to another. If our models were applied to slum areas with different morphological characteristics than those on which they were trained, this could potentially lead to variations in predictive accuracy and uncertainty estimations. Such a transfer would require careful consideration and possibly adjustments to the model to account for these differences. This aspect underscores the importance of context and adaptability in model application, especially in diverse urban environments.

B. Transfer Learning With Few Samples

In Fig. 6, we present the results obtained for the STnet within a similar area of interest as depicted in Fig. 1. To provide additional clarity, we have outlined the slum reference polygons with a black border. Furthermore, we present the slum probability results obtained from the five different training techniques using the same red colorbar. These results shed light on the model's performance in identifying slum settlements. All images (a)–(f) within this figure are consistently displayed at a scale

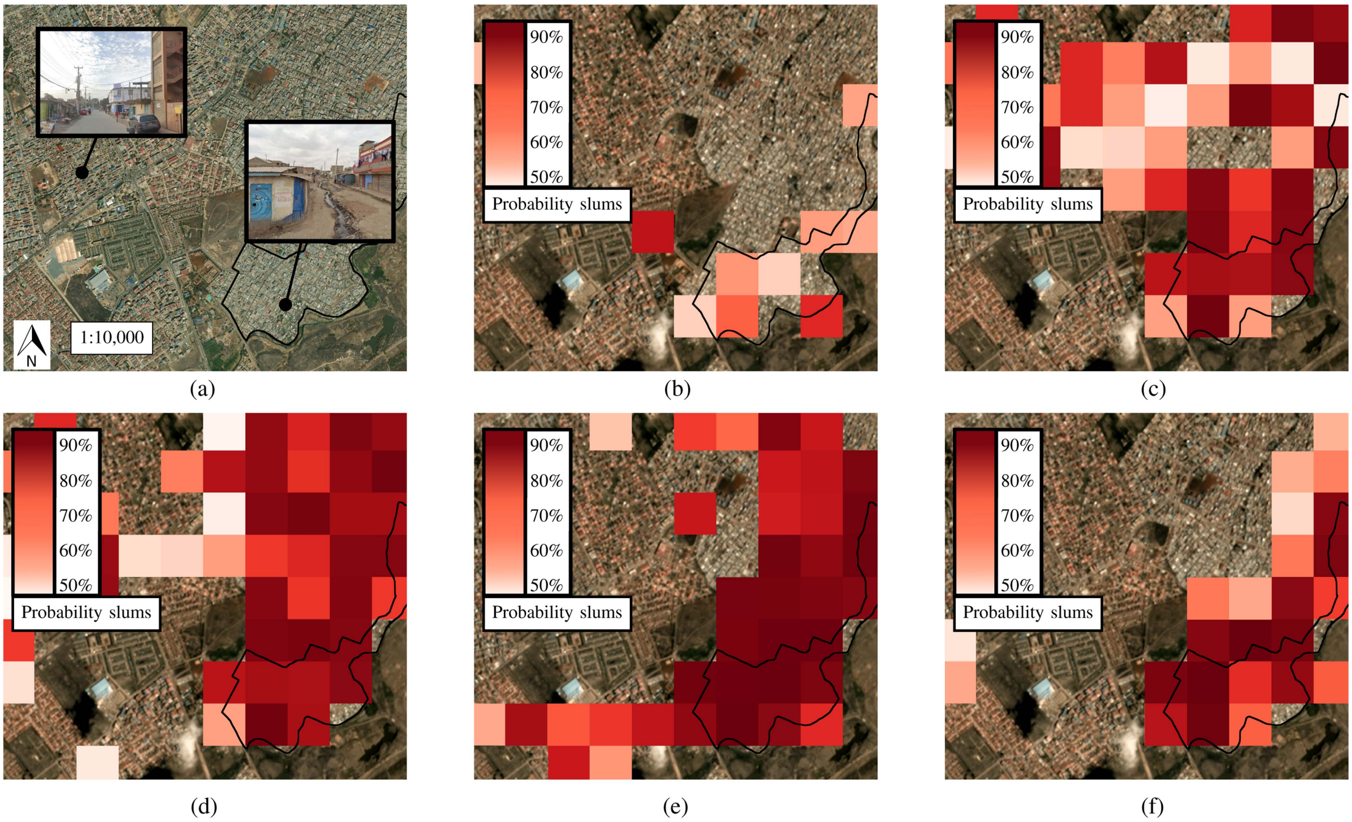

Fig. 6. Results for the STNet in a comparable area of interest, as depicted in Fig. 1. All images (a)–(f) are presented in a consistent scale of 1:10 000. Image (a) showcases a very high-resolution Google satellite imagery of the identical point of interest shown in Fig. 1. Images (b)–(f) exhibit the outcomes obtained using the STNet, with variations from no transfer learning (b) to transfer learning from 1 to 50 samples per class (c)–(f). (a) Google Satellite basemap. (b) No transfer learning. (c) One sample. (d) Five samples. (e) Ten samples. (f) Fifty samples.

of 1:10 000. Fig. 6(a) offers a detailed view, featuring very high-resolution Google satellite imagery of the exact point of interest showcased in Fig. 1. The subsequent images, Fig. 6(b) through 6(f), illustrate the diverse outcomes achieved through the utilization of the STnet. Fig. 6(b) presents results obtained without the application of transfer learning, while images Fig. 6(c) through 6(f) demonstrate the progressive impact of transfer learning with 1 to 50 samples, highlighting the evolution of performance and insights gained through this process. The variation in results for Nairobi, transitioning from utilizing 50 samples per class to 100 samples per class for transfer learning, exhibits negligible differences in both accuracy metrics and visual outcomes. Consequently, we conclude the figure at the 50 sample mark, as further iterations do not yield significant improvements in performance or visual representation. By leveraging transfer learning, we aim to improve the model's ability to recognize and understand the unique features of Nairobi's urban landscape.

From Fig. 4, we find a comprehensive overview of the STnet's performance metrics for the entire city of Nairobi. Specifically, we evaluate the model's F1-score. When employing simple inference without transfer learning, the F1-score achieved was as low as 49.06%, indicative of an initial struggle to map the slums of Nairobi. While Fig. 6(b), initially presents promising results with minimal overclassification tendencies, it is essential to consider the broader context. The depicted area represents only a small portion of the dataset. What is particularly noteworthy is the relatively low confidence values associated with these predictions. This underscores the significance of considering local context, which becomes evident that the models using transfer learning displays higher confidence in its classifications, emphasizing the value of leveraging transfer learning to enhance the classification accuracy and contextual understanding.

However, as we incorporate one sample per class for transfer learning, we observe a notable improvement, with the F1-score rising to 66.78%. This demonstrates the efficiency of using a limited number of labeled samples to enhance the model's understanding of Nairobi's unique morphologic characteristics. In Fig. 6(c), we observe a significant increase in the F1-score for the entire city of Nairobi. However, this improvement is accompanied by a notable issue of overclassification in the area of interest. In addition, there is an evident rise in the overconfidence levels of the predictions, highlighting a disparity between quantitative scores and visual accuracy. From Fig. 6(c) to 6(e), there is a noticeable progression in the F1-score. However, it is not until Fig. 6(f), when a sufficient number of samples are utilized for transfer learning, that the visual outcomes demonstrate considerable improvement. In this instance, the results are promising, exhibiting only minor instances of over- and underclassification.

As discussed in Section I, low-rise and dense settlement structures do not necessarily equate to slum settlements. The region depicted in Fig. 6 exemplifies this challenge, containing only one slum settlement amidst several dense formal settlements. This blend of characteristics intensifies the difficulty of accurate classification. Furthermore, these findings hold broader implications, suggesting that our results are highly generalizable to other cities with similar fuzzy feature spaces between formal, low-rise dense settlements and slum settlements. Cities like Lagos, Rio de Janeiro, and Sao Paulo, known for their similar morphological appearances of slums, can especially benefit from these insights, as they present comparable classification challenges.

VII. Conclusion

Through the integration of Monte Carlo dropout, we gained valuable insights into the uncertainties in our predictions, allowing us to identify areas where our AI model is more or less certain in its slum classification. The presence of multiple typical slum morphologies led to higher certainty in the model's predictions. However, challenges arose when slums shared features with formal areas, which made the classification task more complex. Despite this, the application of Monte Carlo dropout proved to be effective, especially when dealing with noisy datasets and fuzzy feature spaces, which typically pose significant challenges for any classification tasks.

Moreover, we introduced our custom CNN STnet, which demonstrated comparable results to renowned models like ResNet50 and Xception while offering significantly reduced processing time. We have successfully attained an elevated F1-score of 86.24%, a performance that can be deemed remarkable in the context of slum mapping, where we address intricate urban patterns and challenges. Particularly noteworthy was its performance when trained on limited samples, making it an ideal choice for scenarios with fewer available training data. We were able to outscore both Xception and ResNet50 when using ten or fewer samples per class for transfer learning. By combining Monte Carlo dropout, a class-weighted loss function for pretraining, and class-balanced transfer learning, we presented a simple yet efficient approach for accurately classifying challenging urban patterns in noisy and imbalanced datasets. Our approach not only addressed the uncertainties in slum classification but also tackled the inherent complexities of working with real-world data, which often lacks perfect labels and may exhibit imbalances across classes. In summary, our research provides a valuable contribution to the field of urban pattern classification and demonstrates the importance of considering uncertainties in AI models for more accurate and robust predictions. The proposed framework opens avenues for future research in improving the understanding of slum settlements and urban planning, ultimately leading to more effective and targeted interventions in urban development and poverty alleviation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Data Availability

The implementation of all steps in order to reproduce and recreate these results are available through GitHub starkt/UncertaintySlumDetection. The slum data, given its sensitive ethical nature, will be shared exclusively upon reasonable request. Regrettably, the PlanetScope remote sensing data cannot be shared due to copyright restrictions. However, GitHub repository offers resampled Sentinel-2 RGB imagery, serving as a representative example for many of our results.

Authors' contributions

T. Stark, M. Wurm, H. Taubenböck, and X. X. Zhu designed the scope of the study. T. Stark was responsible curating the dataset; X. X. Zhu and team prepared the PlanetScope remote sensing data; T. Stark was responsible for methods and experiments. T. Stark took the lead in writing the manuscript, and M. Wurm, H.Taubenböck, and X. X. Zhu reviewed the manuscript.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] F. Hutter, L. Kotthoff, and J. Vanschoren, *Automated Machine Learning: Methods, Systems, Challenges*. Berlin, Germany: Springer, 2019.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[4] T. Gebru et al., "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[5] C. M. Gevaert, M. Carman, B. Rosman, Y. Georgiadou, and R. Soden, "Fairness and accountability of AI in disaster risk management: Opportunities and challenges," *Patterns*, vol. 2, no. 11, 2021, Art. no. 100363.

[6] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1521–1528.

[7] B. Kellenberger, D. Marcos, and D. Tuia, "Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning," *Remote Sens. Environ.*, vol. 216, pp. 139–153, 2018.

[8] Q. Li, Y. Chen, and P. Ghamisi, "Complementary learning-based scene classification of remote sensing images with noisy labels," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 8021105.

[9] T. Stark, M. Wurm, X. X. Zhu, and H. Taubenböck, "Detecting challenging urban environments using a few-shot meta-learning approach," in *Proc. Joint Urban Remote Sens. Event*, 2023, pp. 1–4.

[10] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[11] M. Schmitt, S. A. Ahmadi, and R. Hänsch, "There is no data like more data—Current status of machine learning datasets in remote sensing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 1206–1209.

[12] C. Geiß, J. Maier, E. So, and Y. Zhu, "LSTM models for spatiotemporal extrapolation of population data," in *Proc. IEEE Joint Urban Remote Sens. Event*, 2023, pp. 1–4.

[13] C. Qiu, X. Tong, M. Schmitt, B. Bechtel, and X. X. Zhu, "Multilevel feature fusion-based CNN for local climate zone classification from Sentinel-2 images: Benchmark results on the So2Sat LCZ42 dataset," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2793–2806, 2020.

[14] D. Stiller et al., "Spatial parameters for transportation," *J. Transport Land Use*, vol. 14, no. 1, pp. 777–803, 2021.

[15] T. Fisher et al., "Uncertainty-aware interpretable deep learning for slum mapping and monitoring," *Remote Sens.*, vol. 14, no. 13, 2022, Art. no. 3072.

[16] R. Engstrom, D. Pavelesku, T. Tanaka, and A. Wambile, "Mapping poverty and slums using multiple methodologies in Accra, Ghana," in *Proc. Joint Urban Remote Sens. Event*, 2019, pp. 1–4.

[17] M. Kuffer et al., "The scope of Earth-observation to improve the consistency of the SDG slum indicator," *ISPRS Int. J. Geo- Inf.*, vol. 7, no. 11, 2018, Art. no. 428.

[18] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, “Combining satellite imagery and machine learning to predict poverty,” *Science*, vol. 353, no. 6301, pp. 790–794, 2016.

[19] M. Wurm and H. Taubenböck, “Detecting social groups from space—Assessment of remote sensing-based mapped morphological slums using income data,” *Remote Sens. Lett.*, vol. 9, no. 1, pp. 41–50, 2018.

[20] H. Taubenböck and N. Kraff, “The physical face of slums: A structural comparison of slums in Mumbai, India, based on remotely sensed data,” *J. Housing Built Environ.*, vol. 29, no. 1, pp. 15–38, 2014.

[21] C. Mood and J. O. Jonsson, “The social consequences of poverty: An empirical test on longitudinal data,” *Social Indicators Res.*, vol. 127, pp. 633–652, 2016.

[22] H. Taubenböck, N. J. Kraff, and M. Wurm, “The morphology of the arrival city—A global categorization based on literature surveys and remotely sensed data,” *Appl. Geogr.*, vol. 92, pp. 150–167, 2018.

[23] O. Gruebner et al., “Mapping the slums of Dhaka from 2006 to 2010,” *Dataset Papers Sci.*, vol. 2014, p. 172182, 2014, doi: 10.1155/2014/172182.

[24] M. Kuffer, K. Pfeffer, and R. Sliuzas, “Slums from space—15 years of slum mapping using remote sensing,” *Remote Sens.*, vol. 8, no. 6, 2016, Art. no. 455.

[25] N. J. Kraff, M. Wurm, and H. Taubenböck, “Uncertainties of human perception in visual image interpretation in complex urban environments,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4229–4241, 2020.

[26] X. Zhu, C. Qiu, J. Hu, Y. Shi, Y. Wang, and M. Schmitt, “NEW: So2Sat LCZ42,” 2019.

[27] M. Wurm, H. Taubenböck, M. Weigand, and A. Schmitt, “Slum mapping in polarimetric SAR data using spatial features,” *Remote Sens. Environ.*, vol. 194, pp. 190–204, 2017.

[28] R. Engstrom, J. S. Hersh, and D. L. Newhouse, “Poverty from space: Using high-resolution satellite imagery for estimating economic well-being,” World Bank Policy Research Working Paper 8284, 2017.

[29] R. Prabhu and B. Parvathavarthini, “An enhanced approach for informal settlement extraction from optical data using morphological profile-guided filters: A case study of Madurai city,” *Int. J. Remote Sens.*, vol. 42, no. 17, pp. 6688–6705, 2021.

[30] N. Mudau and P. Mhangara, “Investigation of informal settlement indicators in a densely populated area using very high spatial resolution satellite imagery,” *Sustainability*, vol. 13, no. 9, 2021, Art. no. 4735.

[31] C. Persello and A. Stein, “Deep fully convolutional networks for the detection of informal settlements in VHR images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2325–2329, Dec. 2017.

[32] D. Verma, A. Jana, and K. Ramamritham, “Transfer learning approach to map urban slums using high and medium resolution satellite imagery,” *Habitat Int.*, vol. 88, 2019, Art. no. 101981.

[33] T. Stark, M. Wurm, X. X. Zhu, and H. Taubenböck, “Satellite-based mapping of urban poverty with transfer-learned slum morphologies,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5251–5263, 2020.

[34] J. Friesen, H. Taubenböck, M. Wurm, and P. F. Pelz, “The similar size of slums,” *Habitat Int.*, vol. 73, pp. 79–88, 2018.

[35] S. Park, J. Lim, Y. Jeon, and J. Y. Choi, “Influence-balanced loss for imbalanced visual classification,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 735–744.

[36] M. S. Reza and J. Ma, “Imbalanced histopathological breast cancer image classification with convolutional neural network,” in *Proc. IEEE 14th Int. Conf. Signal Process.*, 2018, pp. 619–624.

[37] V. H. Barella, L. P. Garcia, M. P. de Souto, A. C. Lorena, and A. de Carvalho, “Data complexity measures for imbalanced classification tasks,” in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2018, pp. 1–8.

[38] M. Burduja and R. T. Ionescu, “Unsupervised medical image alignment with curriculum learning,” in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 3787–3791.

[39] Y. Huang et al., “Curricularface: Adaptive curriculum learning loss for deep face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 5901–5910.

[40] E. Elyan, C. F. Moreno-Garcia, and C. Jayne, “Cdsmote: Class decomposition and synthetic minority class oversampling technique for imbalanced-data classification,” *Neural Comput. Appl.*, vol. 33, pp. 2839–2851, 2021.

[41] A. Ali-Gombe, E. Elyan, and C. Jayne, “Multiple fake classes GaN for data augmentation in face image dataset,” in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–8.

[42] J. Kim, J. Jeong, and J. Shin, “M2m: Imbalanced classification via major-to-minor translation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13896–13905.

[43] S. García, Z.-L. Zhang, A. Altalhi, S. Alshomrani, and F. Herrera, “Dynamic ensemble selection for multi-class imbalanced datasets,” *Inf. Sci.*, vol. 445–446, 2018, pp. 22–37.

[44] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, “Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 59–69, 2019.

[45] N. Karim, U. Khalid, N. Meeker, and S. Samarasinghe, “Adversarial training for face recognition systems using contrastive adversarial learning and triplet loss fine-tuning,” 2021, *arXiv:2110.04459*.

[46] S. O. Ngesthi, I. Setyawan, and I. K. Timotius, “The effect of partial fine tuning on Alexnet for skin lesions classification,” in *Proc. 13th Int. Conf. Inf. Technol. Elect. Eng.*, 2021, pp. 147–152.

[47] S. Hershey et al., “The benefit of temporally-strong labels in audio event classification,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 366–370.

[48] M. Saini and S. Susan, “Vggin-Net: Deep transfer network for imbalanced breast cancer dataset,” *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 20, no. 1, pp. 752–762, Jan./Feb. 2022.

[49] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, “Meta-baseline: Exploring simple meta-learning for few-shot learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9062–9071.

[50] Y. Li, Z. Shao, X. Huang, B. Cai, and S. Peng, “Meta-FSEO: A meta-learning fast adaptation with self-supervised embedding optimization for few-shot remote sensing scene classification,” *Remote Sens.*, vol. 13, no. 14, 2021, Art. no. 2776.

[51] E. Koukouraki, L. Vanneschi, and M. Painho, “Few-shot learning for post-earthquake urban damage detection,” *Remote Sens.*, vol. 14, no. 1, 2022, Art. no. 40.

[52] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.

[53] G. Yang, H.-C. Li, W. Yang, K. Fu, T. Celik, and W. J. Emery, “Variational Bayesian change detection of remote sensing images based on spatially variant Gaussian mixture model and separability criterion,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 3, pp. 849–861, Mar. 2019.

[54] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight uncertainty in neural network,” in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, vol. 37, pp. 1613–1622.

[55] M. Li, A. Stein, and K. M. De Beurs, “A Bayesian characterization of urban land use configurations from VHR remote sensing images,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 92, 2020, Art. no. 102175.

[56] M. Rußwurm, M. Ali, X. X. Zhu, Y. Gal, and M. Körner, “Model and data uncertainty for satellite time series forecasting with deep recurrent models,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 7025–7028.

[57] J. Gawlikowski et al., “A survey of uncertainty in deep neural networks,” 2022, *arXiv:2107.03342*.

[58] P. Ebel, V. S. F. Garnot, M. Schmitt, J. D. Wegner, and X. X. Zhu, “UnCRtainTS: Uncertainty quantification for cloud removal in optical satellite time series,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2023, pp. 2085–2095.

[59] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, vol. 48, pp. 1050–1059.

[60] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra, “Reducing overfitting in deep networks by decorrelating representations,” 2015, *arXiv:1511.06068*.

[61] Z. Li, T. Zhang, S. Cheng, J. Zhu, and J. Li, “Stochastic gradient Hamiltonian Monte Carlo with variance reduction for Bayesian inference,” *Mach. Learn.*, vol. 108, pp. 1701–1727, Sep. 2019.

[62] A. Lemay et al., “Improving the repeatability of deep learning models with Monte Carlo dropout,” *NPJ Digit. Med.*, vol. 5, Nov. 2022, Art. no. 174.

[63] Y. Xiao and W. Y. Wang, “Quantifying uncertainties in natural language processing tasks,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 7322–7329.

[64] T. J. Loftus et al., “Uncertainty-aware deep learning in healthcare: A scoping review,” *PLoS Digit. Health*, vol. 1, pp. 1–15, 2022.

[65] S. Ma, J. Huang, Y. Xie, and N. Yi, “Identification of breast cancer prognosis markers using integrative sparse boosting,” *Methods Inf. Med.*, vol. 51, no. 2, pp. 152–161, 2012.

[66] A. Jungo et al., “On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation,” in *Proc. 21st Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, Granada, Spain, Sep. 16–20, 2018, pp. 682–690.

[67] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[68] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251–1258.

[69] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

Thomas Stark received the M.Sc. degree in geodesy and geoinformation in 2018 from the Technical University of Munich, Munich, Germany, where he is currently working toward the Ph.D. degree on the topic of “Towards detecting global urban poverty” with the Department of Aerospace and Geodesy, Data Science in Earth Observation.

In 2017, he joined the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Weßling, Germany, as a Research Associate. His current research interests include encompass a robust and dynamic exploration of the field of computer vision, and delving into its multifaceted intricacies. A particular area of profound fascination lies in his keen interest in harnessing the potential of remote sensing data, which adds an extra layer of dimensionality to his investigations. This interest stems from a genuine passion for unraveling the hidden insights concealed within vast datasets, a pursuit driven by a profound commitment to the UN Sustainable Development Goals. With an unwavering dedication to these global objectives, he aspires to pave the way for a more sustainable and equitable future by ingeniously applying the principles of computer vision and adept data analysis techniques.

Michael Wurm received the diploma degree (Mag. rer. nat.) in geography with a specialization in remote sensing, GIS, and spatial research from the University of Graz, Graz, Austria, in 2007, and the Ph.D. degree (Dr. rer. nat.) in surveying and geoinformation from the Graz University of Technology, Graz, in 2013.

He was with the Institute of Digital Image Processing, Joanneum Research, Graz, in 2007. In 2008, he joined the University of Wurzburg, Germany, where he was involved in interdisciplinary research between Earth observation data and social sciences. Since 2011, he has been with the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Weßling, Germany. In 2022, he became the Head of the “City and Society” team where he is involved in topics on urban geography, urban remote sensing, and urban morphology, and slum mapping research. Since 2013, he has been a Lecturer with the University of Graz.

Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc., Dr.Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She was the Founding Head with the Department “EO Data Science,” Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany. She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan, and the University of California at Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. Since 2019, she has been a Co-Coordinator with the Munich Data Science Research School, Munich. Since 2019, she has been heading the research field “Aeronautics, Space, and Transport” with Helmholtz Artificial Intelligence. Since May 2020, she has been the Principal Investigator (PI) and the Director with the International Future AI lab “AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond,” Munich. Since October 2020, she has been serving as the Director with the Munich Data Science Institute (MDSI), TUM, where she is currently the Chair Professor of data science in Earth observation. Her main research interests include remote sensing and Earth observation, signal processing, machine learning, and data science, with their applications in tackling societal grand challenges, e.g., global urbanization, United Nations (UN’s) Sustainable Development Goals (SDGs), and climate change.

Dr. Zhu is a member of the Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She serves on the scientific advisory board of several research organizations, including the German Research Center for Geosciences (GFZ) and the Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She serves as an Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.

Hannes Taubenböck received the diploma degree in geography from the Ludwig-Maximilians Universitat Muenchen, Munich, Germany, in 2004, the Ph.D. (Dr.rer.nat.) degree in geography from the Julius Maximilian’s University of Wuerzburg, Wuerzburg, Germany, in 2008, and the Habilitation degree in geography from the University of Wuerzburg, Wuerzburg, in 2019.

In 2005, he joined the German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Weßling, Germany. After a postdoctoral research phase with the University of Wuerzburg (2007–2010), he returned in 2010 to DLR–DFD as a Scientific Employee. Since 2022, he has been a Professor with the Institute of Geography and Geology, University of Wuerzburg, for the working group Earth observation. Since 2022, he has also been the Head of the Geo-Risks and Civil Security Department, DFD. His current research interests include urban remote sensing topics, from the development of algorithms for information extraction to value adding to classification products for findings in urban geography.

A.4. Uncertainty Aware Slum Mapping in 55 Heterogeneous Cities

Reference: Stark, T., Wurm, M., Debray, H., Zhu, X. X., & Taubenböck, H. Uncertainty Aware Slum Mapping in 55 Heterogeneous Cities. Submitted to Int. J. Appl. Earth Obs. Geoinf. (2024).

Highlights

Uncertainty Aware Slum Mapping in 55 Heterogeneous Cities

Thomas Stark,Michael Wurm,Henri Debray,Xiao Xiang Zhu,Hannes Taubenböck

- Using advanced machine learning to map slums in 55 cities with limited data.
- Uncertainty Mapping: Dropout, augmentation, and ensembles for robust detections.
- Large Scale Slum Dataset: First to use coherent methods and probability estimates on data.
- Urban Variability: Maps reveal slum details, offering crucial insights for urban planning.

Uncertainty Aware Slum Mapping in 55 Heterogeneous Cities

Thomas Stark[a,b,*], Michael Wurm[b], Henri Debray[a,b,c], Xiao Xiang Zhu[a,d] and Hannes Taubenböck[b,c]

[a]*Data Science in Earth Observation, Technical University of Munich (TUM), 80333 Munich, Germany,*
[b]*German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), 82234 Oberpfaffenhofen, Germany,*
[c]*Institute of Geography and Geology, University of Würzburg, 97074 Würzburg, Germany,*
[d]*Munich Center for Machine Learning, 80333 Munich, Germany,*

ARTICLE INFO

Keywords:
Slum Mapping
Transfer Learning
Ensemble Classifications
Uncertainty Quantification
Remote Sensing

ABSTRACT

Slums are densely populated urban areas characterized by substandard housing and squalor. These areas often lack basic infrastructure and services, making them challenging to manage and improve. Mapping slums on a large scale is particularly difficult due to their complex, dynamic, and non-uniform nature and the scarcity of available data. This research leverages advanced machine learning techniques and uncertainty aware methodologies to map slum areas across 55 heterogeneous cities. We effectively address the challenges posed by limited labeled data and achieve robust slum probability maps. Our coherent methodology, applied to a large slum dataset across the Global South, includes probability estimates for each prediction, offering a detailed understanding of slums within each city. These insights offer a spatially detailed map of slums and the multiple facets that come with slum settlements and their probabilities. By providing a nuanced view of slum distributions, our work highlights the diversity and complexity of slum settlements, contributing to a more comprehensive understanding of these areas. A significant achievement of our research is detecting various slum categories, depending on their set of slum morphology features, and their different probabilities, especially in cases where slum settlements gradually transition into formal settlements or display atypical characteristics. This approach offers a substantial improvement over traditional binary slum classification methods that focus solely on typical slum morphologies.

1. Introduction

Slums remain a prominent feature of urban landscapes in the Global South, underscoring their socio-economic significance despite ongoing urban development efforts (UN-Habitat et al., 2020). While substantial research has focused on the detection and characterization of slums (Kuffer et al., 2016; Mahabir et al., 2018; Taubenböck et al., 2018) and advancements in methodological approaches (Persello and Stein, 2017; Wurm et al., 2019; Verma et al., 2019a), the limitations of large-scale mapping highlights the necessity of moving beyond existing methods. The challenge is particularly evident in less-studied cities, unlike well-documented ones such as Lagos, Mumbai, and Nairobi, where comprehensive urban poverty data are more readily accessible (Taubenböck and Kraff, 2014; Badmos et al., 2018; Kraff et al., 2019; Mahabir et al., 2020; Stark et al., 2020).

The identification and assessment of slums across diverse urban areas, from small to large, present unique complexities (Friesen et al., 2018; Kuffer et al., 2020; Kraff et al., 2020). Unlike well-known mega-cities, smaller and mid-sized cities often lack visibility and recognition in academic and policy circles, resulting in limited resources and attention towards their urban dynamics, including slum prevalence and characteristics. Traditional methods face significant obstacles, including data scarcity and limited scientific studies on slum locations in these cities (Kuffer et al., 2016; Stark et al., 2023). Therefore, there is an urgent need for a novel approach capable of large-scale analysis, as existing methodologies prove insufficient for comprehensive city-wide and global contexts.

Addressing the persistent proliferation of slums, especially in mid-sized cities, is critical for sustainable urban development and social equity, aligning with the Sustainable Development Goals (Sachs et al., 2022). These cities are vital economic and cultural hubs, yet their marginalized populations face vulnerabilities and deprivation similar to those in larger metropolitan areas (Abascal et al., 2022). Rapid urban growth exacerbates inequality and pressures resources and services, necessitating targeted actions to reduce urban poverty (United Nations, 2024a) and improve living conditions. Innovative approaches such as the utilization of remote sensing for developing a global slum repository, as highlighted by (Kuffer et al., 2021), play a crucial role in these efforts. In line with these strategies, the United Nations has emphasized the importance of better data collection and utilization starting in 2024 to support these initiatives and ensure inclusive and effective urban development (United Nations, 2024b).

Urban slums, as defined by the United Nations, are areas characterized by a lack of access to improved water and sanitation, insufficient living space, poor structural quality of housing, and insecure tenure (UN-Habitat, 2003). Remote sensing imagery can identify slum settlements by examining physical characteristics such as varied housing structures, poor-quality building materials, and limited road access. These characteristics also infer indirectly measurable factors like high population density and low income levels to aid in identification (Taubenböck and Kraff, 2014; Wurm and Taubenböck, 2018; Wang et al., 2019b; Engstrom et al.,

*Corresponding author
thomas.stark@dlr.de (T. Stark)
ORCID(s): 0000-0002-6166-7541 (T. Stark); 0000-0001-5967-1894 (M. Wurm); 0000-0002-4329-0541 (H. Debray); 0000-0001-5530-3613 (X.X. Zhu); 0000-0003-4360-9126 (H. Taubenböck)

2021) However, numerous challenges complicate accurate assessment:

Inter-urban variability: Slum morphologies differ between cities, with some displaying all hallmark indicators while others exhibit only a subset. **Intra-urban variability**: Within a single city, slum characteristics vary, leading to diverse manifestations (Taubenböck et al., 2018; Debray et al., 2023). For instance, Lagos hosts slums on water bodies with high density and areas blending into formal urban structures. **Fuzzy feature space**: High-resolution remote sensing struggles to distinguish between informal settlements and dense formal urban areas, especially when clear borders are absent, resulting in gradual transitions (Stark et al., 2020; Li et al., 2022). **Data scarcity**: Despite high population density, the actual land area occupied by slums is limited, leading to insufficient data availability, posing challenges for deep learning algorithms that rely on abundant data for accurate analysis (Wurm et al., 2019).

Recognizing the gradual change from informal to formal settlements and the inherent uncertainties in this spectrum is essential for advancing the field. Uncertainty aware methods provides valuable insights into the diverse morphologies of slum settlements and aids in developing more robust methodologies. Addressing these challenges requires innovative approaches, integrating remote sensing data, advanced machine learning algorithms, and efficient labelling techniques to enhance understanding and characterization of urban slums and their interactions with various city types (Wurm et al., 2017; Taubenböck et al., 2020; Thomson et al., 2020; Debray et al., 2021; Stark et al., 2024). Collaborative efforts between researchers, policymakers, and local communities are essential for developing effective strategies to improve living conditions and alleviate challenges faced by slum residents globally (Kuffer et al., 2020).

Previous methodologies in slum mapping through deep learning have made significant strides in accuracy but highlight the need for more recent advancements in machine learning methodologies (Persello and Stein, 2017; Wurm et al., 2019). Techniques such as few-shot learning (Stark et al., 2023) and efficient transfer-learning approaches (Stark et al., 2020; Verma et al., 2019b; Zhao et al., 2023) have been introduced to identify slum settlements with minimal training data. Recent developments focus on uncertainty-aware predictions in slum classification, enhancing our understanding of slum morphologies (Fisher et al., 2022; Stark et al., 2024). Interpretable deep learning techniques for consistent large-scale urban population estimation using earth observation data are also being utilized (Doda et al., 2024). Despite promising trends, widespread application on a large scale across various city types remains limited, underscoring the need for further research and implementation to scale up slum mapping initiatives.

Our study aims to significantly improve slum mapping by integrating various methodologies at large scale. We focus on enhancing slum classification using robust methods to quantify uncertainty, such as test-time dropout, test-time augmentation, and model ensembles, to estimate both aleatoric and epistemic uncertainty. Efficient transfer-learning strategies make labeling efforts manageable. Our research applies these methods to 55 entire cities across the Global South, providing a city-wide slum probability maps for each city, which is unprecedented. This extensive slum dataset allows us to compare and identify similarities and differences in slum settlement morphologies.

2. Materials

To examine the effectiveness of our transfer-learning approach, we carefully selected a diverse set of 55 cities for our study. Our criteria for selection focused on cities located in the Global South and characterized by a significant presence of densely built-up urban areas, as mapped by the local climate zones classes 3, 6, and 7, as identified in the dataset provided by (Zhu et al., 2020), based on the classification scheme introduced in (Stewart and Oke, 2012). These specific classes were chosen because they typically exhibit dense settlement patterns, which are not only indicative of potential slum areas but also present a challenge in distinguishing between formal settlements and informal slum areas. As illustrated in Figure 1, the selected cities are geographically dispersed across the Global South, representing a wide spectrum of population sizes. For example, Ilorin, with a population of 842 thousand, is much smaller than Delhi, which has over 17 million inhabitants. This diverse selection includes cities ranging from smaller urban centers to large mega-cities.

2.1. Remote Sensing Data

To analyze the selected 55 cities comprehensively, we acquired PlanetScope data from the year 2022. This dataset consists of three channels: Red, Green, and Blue, with a radiometric resolution of 8 bits and a geometric resolution of 4.77 meters. Although most of the scenes exhibit minimal cloud coverage, there are occasional instances where clouds may obstruct the view of the cities.

In order to ensure a consistent comparison of city scales, we utilize a bounding box around the morphological urban areas as defined by (Taubenböck et al., 2019) to crop the PlanetScope data. We use the bounding box to ensure an adequate representation of vegetation and water bodies for our classification schema. This approach allows us to capture a broader range of environmental features while maintaining a standardized method for analyzing urban areas across different cities.

2.2. Reference Dataset

The reference dataset was crafted through the integration of our own slum dataset (Stark et al., 2020, 2024) and the local climate zones (LCZ) dataset (Zhu et al., 2020). The slum dataset was generated through manual mapping of slum polygons, delineated by a team of remote sensing experts with extensive knowledge in urban poverty. To ensure consistency and accuracy, all labels were standardized and updated, if necessary, to align with the PlanetScope imagery from 2022. The mapping process utilized information from

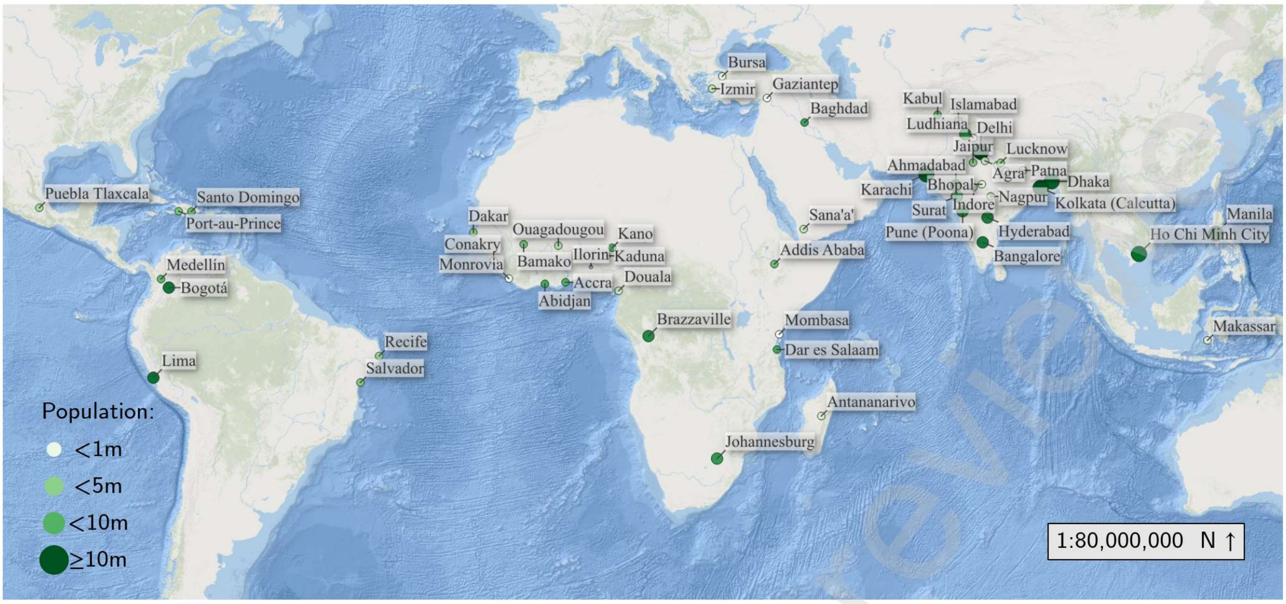

Fig. 1: Location of the 55 cities across the Global South. The cities are scaled and colored by their population size.

various sources, including the original PlanetScope data, Google Satellite imagery, and, where accessible, Google Street View data. This comprehensive approach allowed for a robust and detailed representation of slum areas within the dataset.

In our slum dataset, we labeled a maximum of five slum settlements per city or the equivalent of approximately 4.6 square kilometers. This strategy was employed to balance the need for detailed, high-quality data with the practical constraints of manual labeling efforts. By focusing on a limited number of slum areas per city, we aimed to ensure the accuracy and consistency of our labels, which is crucial for reliable analysis. This selective approach was applied across all 55 cities included in the study. In many of the 55 cities, we expect the number of slum settlements to be higher than what we have mapped. Recognizing the potential presence of additional slums is essential for ensuring that our findings are robust and generalizable across different urban contexts.

To enhance this dataset, the LCZ data was reclassified into four distinct classes. This reclassification involved merging urban classes derived from LCZ classes 1 through 10, consolidating all non-built-up and vegetation classes from LCZ classes A(11) through F(16), and incorporating a water class. Subsequently, the slum data was amalgamated into this reclassified dataset, resulting in a comprehensive representation of the urban landscape.

2.3. Dataset Preparation

To facilitate our analysis, we divide our remote sensing data into smaller image tiles measuring 224 × 224 pixels, corresponding to an area of 1052 square meters per tile. Each tile overlaps with its neighboring tiles by 45 pixels in both the x and y directions. To assign the appropriate label to each image tile, we examine the frequency of pixel values from the reference dataset within the tile. During the training and validation phases, we exclusively utilize image tiles where the occurrence of labels comprises at least 70% of urban, vegetation, or water classes, or if the occurrence of slum class pixels is at least 10%.

From the pixel occurrences within each image tile, we determine the corresponding label. For urban, vegetation, and water classes, the label is determined by the majority class within the tile. However, if the tile contains at least 10% of slum pixels, it is classified as a slum label, recognizing the significance of informal settlements within the urban landscape. This approach ensures that our dataset accurately reflects the diverse characteristics of urban areas, including both formal and informal settlements, enabling robust training and validation of our models.

3. Methods

In our study, we employ a robust methodology that leverages transfer-learning principles to enhance predictive performance. A simplified schematic overview of the complete process can be seen in Figure 2. Initially, we pretrain Convolutional Neural Networks (CNN) utilizing data from four representative cities, enabling the network to grasp foundational patterns and features.

Subsequently, we seamlessly transfer this knowledge to a target city selected from a pool of 55 diverse urban environments. To enhance the reliability of our predictions and account for uncertainty, we integrate approximation methods, incorporating test-time augmentations and dropout techniques. Moreover, to bolster the robustness of our model, we deploy an ensemble of four CNNs, enabling us to aggregate predictions effectively and enhance overall accuracy. This multifaceted approach underscores our commitment

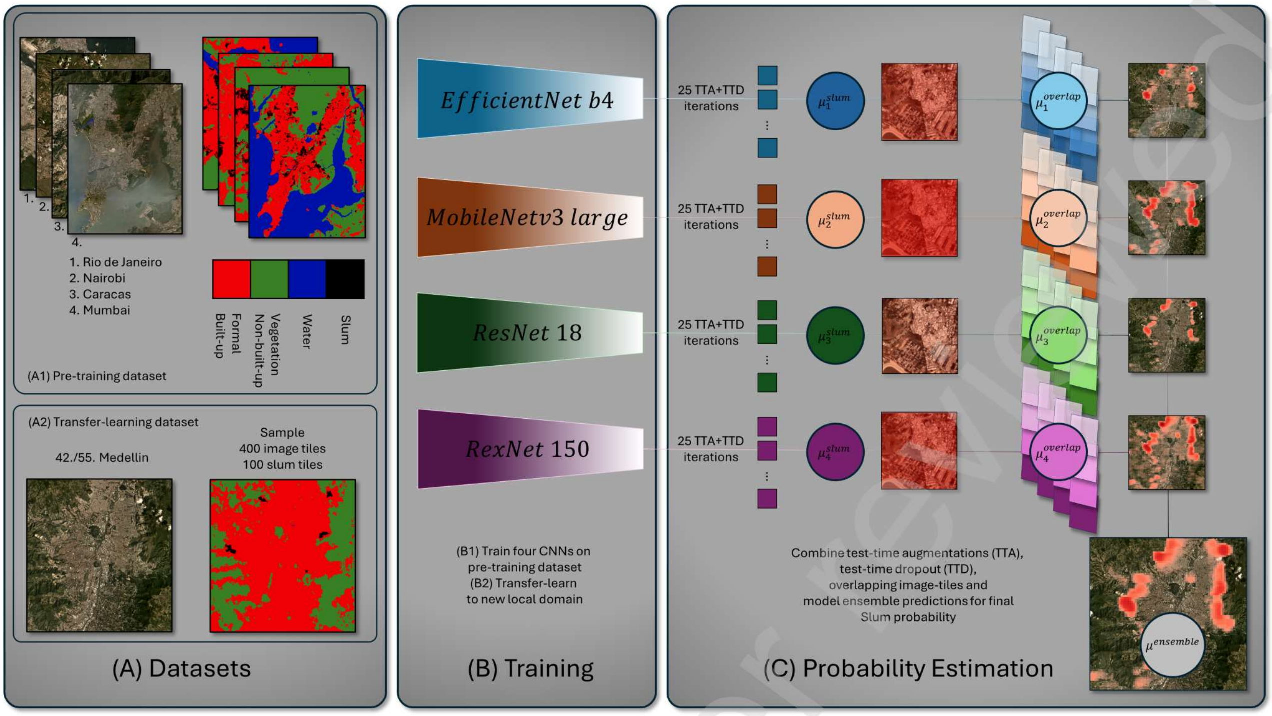

Fig. 2: A simplified schematic overview of our approach to estimate slum probabilities on a large scale using Medellin, Colombia, as an example for one (42^{th}) of the 55 cities in our slum probability dataset. First, we pre-train four CNNs on an initial imbalanced dataset and then transfer-learn these representations to a balanced city's dataset. We approximate the final slum probability using multiple methods, including test-time augmentations, test-time dropout, overlapping image tiles, and model ensembles.

to achieving comprehensive and reliable results in urban prediction tasks.

In our study, we employ four CNNs, where each CNN possesses unique characteristics and advantages, contributing to a comprehensive evaluation framework. Firstly, ResNet-18, as proposed by (He et al., 2015), is a cornerstone in deep learning with 11.7 million parameters, balancing complexity and efficiency. Secondly, ReXNet-150, introduced by (Han et al., 2021), comprises 9.8 million parameters, optimizing performance with effective channel dimension configuration. Thirdly, EfficientNet-B4, proposed by (Tan and Le, 2019), features 19.5 million parameters and scales depth, width, and resolution dimensions, offering substantial capacity for capturing intricate patterns. Lastly, MobileNetV3 Large, introduced by (Howard et al., 2019), has 5.5 million parameters, excelling in resource-constrained environments with its compact yet powerful design. By incorporating these diverse CNN architectures, we aim to benefit from their capabilities across a spectrum of tasks.

3.1. Transfer-learning

Introducing our tranfer-learning methodology, illustrated in Figure 2, we start by pre-training CNN models initialized with ImageNet weights on a sizable but imbalanced remote sensing dataset. Seen in Figure 2 (A1) and (B1) we first pre-train our CNNs using data from four well-known cities: Caracas, Mumbai, Nairobi, and Rio de Janeiro. We selected these cities with well documented spatial information (Stark et al., 2020, 2024). This allows us to create a large dataset of 143,188 image tiles, with 21,448 tiles classified as slum areas.

Subsequently, we employ transfer-learning to adapt these models to a specific city's dataset, which is carefully balanced with 100 image tiles per class as seen in Figure 2 (A2) and (B2). For each city depicted in Figure 1, we selected 100 samples per class, totaling 400 samples. The classes include urban, vegetation, and water, which are randomly sampled. The slum samples are specifically sampled from different slum areas to ensure geographic diversity for transfer-learning, transfer-validation, and transfer-testing. It is important to note that while urban, vegetation, and water samples are drawn from the entire city, slum samples are limited to a few slum settlements. This means that due to random sampling, slum image tiles might be present in the urban, vegetation, and water categories, resulting in a class-balanced transfer-learning dataset but with potentially noisy labels. During the transfer-learning process, the entire CNN architecture remains trainable, with no layers being frozen. This strategy allows for enhanced adaptability to the unique features present within the selected cities, optimizing model performance and accuracy in classification tasks.

3.2. Test-time Augmentation and Dropout

Each transfer-learned model estimates the uncertainty of its predictions by averaging over 25 iterations (Gal and

Ghahramani, 2016; Stark et al., 2024). To gauge uncertainty we use test-time augmentation to address epistemic uncertainty (model uncertainty) by applying various data augmentations to the test data, thereby reducing the model's lack of knowledge through consensus predictions. We use the same methods for data augmentations as seen in (Wang et al., 2019a). Test-time dropout addresses both epistemic uncertainty and aleatoric uncertainty (data uncertainty) capturing variability due to model uncertainty and intrinsic noise in the data (Wang et al., 2019a; Ebel et al., 2023). For the dropout method we use the same value of 0.3 as shown in (Stark et al., 2024).

For our experiments we combine several steps regarding uncertainty approximations within our methodology:

First, we combine the logits of our final layer L for $i = 1$ to 25 iterations into an array of size $n{\times}i$, where $n = 4$ classes:

$$L = \begin{bmatrix} L_{11} & L_{12} & \dots & L_{1i} \\ L_{21} & L_{22} & \dots & L_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \dots & L_{ni} \end{bmatrix} \tag{1}$$

Next, we apply a sigmoid function $\sigma(x)$ over the array to scale each network output to the range $[0, 1]$:

$$\sigma(L) = \begin{bmatrix} \sigma(L_{11}) & \sigma(L_{12}) & \dots & \sigma(L_{1i}) \\ \sigma(L_{21}) & \sigma(L_{22}) & \dots & \sigma(L_{2i}) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(L_{n1}) & \sigma(L_{n2}) & \dots & \sigma(L_{ni}) \end{bmatrix} \tag{2}$$

Finally, we calculate the mean for only the corresponding values to the 4^{th} class, which we define as our slum class probability μ_{slum}.

$$\mu_{slum} = \frac{1}{i}\sum_{k=1}^{i} \sigma(L)_{4k} \tag{3}$$

3.3. Overlapping image tile predictions

After calculating the approximated slum probability μ_{slum}, each image tile is georeferenced to its original remote sensing data source. To address the inherent characteristics of remote sensing datasets and mitigate edge-related issues, we employ a strategy of predicting on overlapping image tiles.

Due to the use of overlapping image tiles, each tile overlaps with its neighboring tiles by 45 pixels in both the x and y directions. This overlap results in five overlapping predicted image tiles for each location in the original image. To generate a final prediction for each location, we calculate the mean probability from these five overlapping tiles. This process involves averaging the predicted probabilities, which helps to smooth out noise and improve the robustness of the predictions.

As a result, the final output is a set of image tiles with a size of 45 × 45 pixels, where each pixel represents the mean probability derived from the overlapping predictions. This approach leverages the redundancy provided by overlapping tiles to enhance the accuracy and reliability of the final probabilistic outputs.

The overlapping mean probability $\mu_{overlap}$ in equation 4 for each pixel (i, j) can be expressed as follows:

$$\mu_{overlap} = \frac{1}{5}\sum_{k=1}^{5} p_{i,j,k} \tag{4}$$

where $p_{i,j,k}$ is the predicted probability at the pixel (i, j) from the k-th overlapping tile, and k ranges from 1 to 5.

3.4. Predicted model ensembles

Our research adopts an additional approach of averaging model predictions across multiple models. We are essentially trying to capture different sources of uncertainty stemming from variations in model architectures and initializations. By averaging these predictions as seen in equation 5, we are attempting to mitigate the uncertainty associated with individual models.

$$\mu_{ensemble} = \frac{1}{n}\sum_{j=1}^{n} \mu_4 \tag{5}$$

This strategy aims to enhance predictive stability and interpretability within ensemble learning frameworks, as shown in Figure 2 (C). By averaging predictions within each model, we reduce inherent variability and mitigate the influence of outliers or noisy predictions. Averaging within each model promotes greater stability, clearer interpretation of contributions, and stabilizes predictions despite high variability (Song et al., 2023).

4. Results

4.1. Classification accuracy metrics

Table 1 presents the accuracy metrics for the four different classes, including the standard deviation of results from the 55 cities. These results are based on the transfer testing dataset. It is important to note that the testing dataset is noisy, as the mapped slum settlements in most of the 55 cities are incomplete and not fully represented. Nevertheless, Table 1 provides an indicator of the class-based accuracies of our method.

Here is the text with percentage signs added in LaTeX style:

The urban, formal built-up class exhibits high scores, with an F1 score of 95.27% ± 5.68%, precision of 95.42% ± 2.17%, and recall of 95.63% ± 8.46%. These values indicate that the model is highly effective at correctly identifying urban areas with both high sensitivity and specificity. For the vegetation, non built-up class, the metrics are also robust, with an F1 score of 92.06% ± 6.09%, precision of 94.74% ± 4.70%, and recall of 89.84% ± 8.25%, suggesting that vegetation is generally well-detected, though slightly less so than urban areas. The water class shows a high

Table 1
Accuracy metrics for the four classes in our classification schema.

Class	F1 [%]	Precision [%]	Recall [%]
Built-up	95.27±5.68	95.42±2.17	95.63±8.46
Non built-up	92.06±6.09	94.74±4.70	89.84±8.25
Water	73.98±28.56	98.17±4.79	76.63±26.08
Slum	47.41±19.62	47.81±29.81	87.19±21.50

variability, with an F1 score of 73.98% ± 28.56%, precision of 98.17% ± 4.79%, and recall of 76.63% ± 26.08%. The high precision but lower recall for water indicates that while water bodies are accurately classified when detected, many instances may be missed, resulting in false negatives.

Our focus is on the slum class, which presents unique challenges in the context of slum scene classification. The slum class has an F1 score of 47.41% ± 19.62%, a precision of 47.81% ± 29.81%, and a recall of 87.19% ± 21.50%. The high recall value indicates that the model is able to correctly identify a significant portion of actual slum areas. However, the lower precision suggests that a considerable number of non-slum areas are incorrectly classified as slums. This discrepancy is expected due the aforementioned noisiness of our testing dataset.

4.2. Slum probability maps

In our study, we provide a visual representation of slum probabilities across all 55 cities, depicted in Figures 3 and 4. These visualizations showcase slum probabilities within the morphological urban areas, employing a consistent color scale ranging from 0% to 100%.

The visual results reveal diverse types of slum settlements. We observe a range of slum sizes, from small pockets in cities like Baghdad, Bangalore, Conakry, Johannesburg, Kaduna, Karachi, Calcutta, Lahore, Lucknow, Patna, Puebla, Pune, and Surat, to large, extensive settlements in Douala, Ilorin, Kano, Medellin, Mombasa, Ouagadougou, Port Au Prince, Recife and Salvador. The probability values exhibit significant variability. For instance, cities like Dakar, Ilorin, Indore, Kano, Lahore, Lucknow, Mombasa, Port Au Prince, Recife and Salvador show relatively low probability values, indicating less certain slum areas. In Delhi and Ho Chi Minh slum probability values are very low. In contrast, cities such as Johannesburg, Manila, Jaipur, Islamabad, Douala, and Ouagadougou exhibit high slum probabilities, suggesting more extensive or more certain slum areas.

The challenge of inter- and intra-city variability is noticeable. The results demonstrate a wide range of different morphological slum types, which affect the probabilities observed within each city. This variability includes differences in the size of slum settlements. Additionally, the morphological features present in these slums influence the probability values. This effect can be seen when comparing the results for cities with large areas of low slum probabilities to cities with fewer and smaller slums with high probability values.

Due to the wide range of slum probabilities even within single cities, it is problematic to describe any city as having uniformly large and certain slum areas. Thus extra care is needed to categorize the diverse nature of slum settlements and their varying probabilities.

4.3. Comparing slum probabilities for all cities

Figure 5 illustrates the probability of slums in each city, providing a comprehensive overview of the likelihood of slum presence for each city visually represented in Figures 3 and 4. For our analysis, we utilized all non-zero slum probabilities to calculate these statistics, ensuring that only significant data points were included. The cities are grouped into three distinct slum categories based on their probability distributions, which offer insights into the varying levels of slum probability across different urban areas.

In Figure 5, the bar length represents the range of slum probability pixel values, from the minimum to the maximum value. The lower and upper whiskers indicate the pixel values between the 25th and 75th percentiles, respectively. The median slum probability value is depicted by the 50th percentile.

In Figure 5, Category 1 (light green) includes cities with a large distribution of high slum probabilities, typically between the 50th and 75th percentiles. Category 2 (medium green) and Category 3 (dark green) represent cities with smaller distributions and lower slum probabilities. The error bars indicate the variability in the slum probability estimates for each city.

We observe high slum probabilities in cities such as Douala, Islamabad, and Johannesburg, indicated by the longer bars and higher percentile ranges. Conversely, cities like Conakry, Ho Chi Minh City, and Delhi exhibit the lowest slum probabilities, as evidenced by the shorter bars and lower percentile ranges. This categorization helps in understanding the distribution of slums across different urban areas.

5. Discussion

5.1. Impact of incomplete reference data

In our large-scale urban slum mapping analysis, we face significant challenges primarily due to the quality of the reference dataset. We suspect that the dataset underrepresents the actual number of slum settlements, as mapping 55 cities completely would not be feasible otherwise. Keeping this in mind, the performance metrics from Table 1 reveal a high recall rate of 87.19%, which confirms the model's strong ability to detect slum areas. However, the precision is considerably lower at 47.81%, indicating a substantial number of false positives where non-slum areas are mistakenly classified as slums.

The primary reason for this issue appears to be the incomplete mapping of slum areas in our reference dataset. When our model identifies an area as a slum that isn't listed as such in the reference data, it is incorrectly counted as a false positive, rather than being recognized as a true positive.

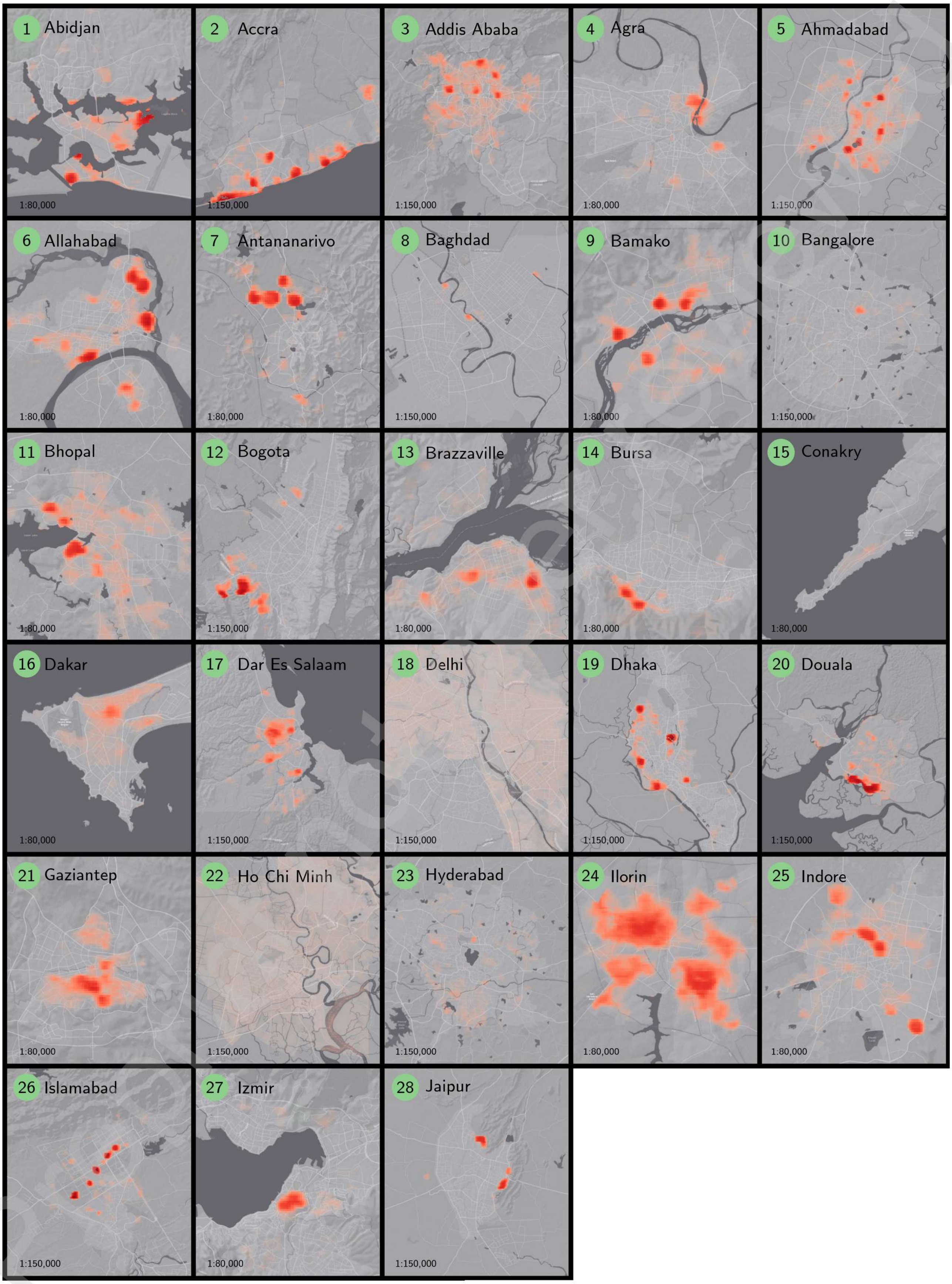

Fig. 3: Slum probability maps for the cities 1 to 28.

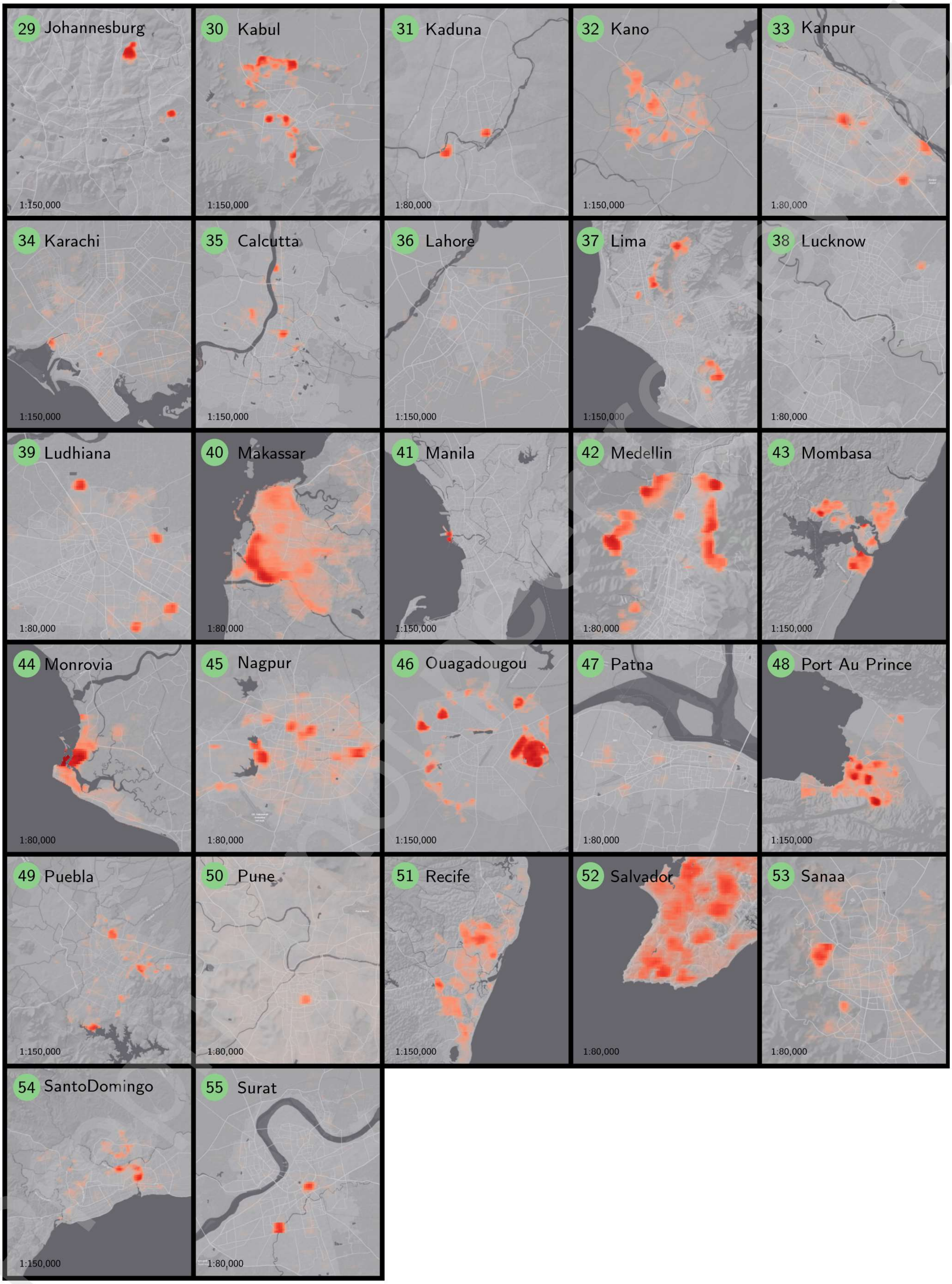

Fig. 4: Slum probability maps for the cities 29 to 55.

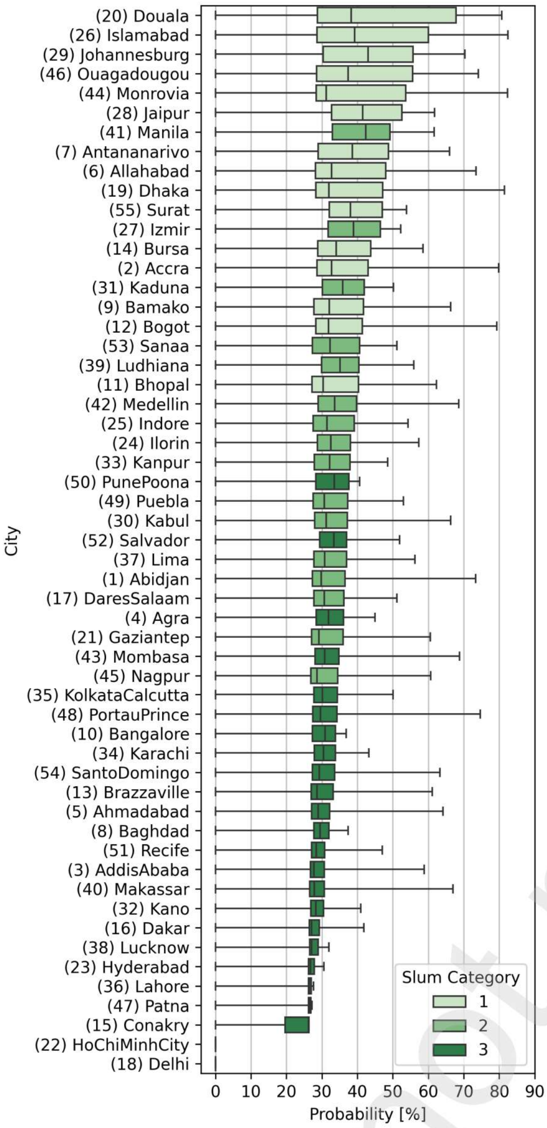

Fig. 5: The probability of slums in each city is categorized into three distinct groups, ranging from high to low slum probabilities. The city IDs correspond to the numbers in Figure 3 and Figure 4.

Addressing this requires a focused effort to enhance the dataset quality, ensuring that it more accurately represents the actual distribution of slum areas. Improving data collection and expanding the scope of mapped slum regions could help in aligning the model's output with real-world conditions, thereby increasing the precision while maintaining high recall levels.

5.2. Categorizing slum probabilities

As seen in Figure 5, we grouped the slum probabilities into three categories. This approach follows the methodology of (Taubenböck et al., 2018), where slums are categorized based on their morphological features. In (Stark et al., 2020), we adopted this approach and created three slum categories: typical slum morphologies, somewhat typical slum morphologies, and mostly unlike typical slum morphologies.

In this study, we use slum probability data to classify slum morphologies. We grouped cities by their slum probability pixel counts: Category 1 includes cities with many high-probability pixels, Category 2 has a moderate number, and Category 3 has few. This categorization helps better understand and address slum distributions in urban areas.

The challenge of identifying typical slum morphologies is compounded by the diversity of slum conditions across different regions. In some cities, slums may be characterized by informal, makeshift housing with little to no access to basic services. In others, more permanent structures may exist, but still lack adequate infrastructure and legal recognition. This variability necessitates a nuanced approach to slum classification, taking into account local contexts and the dynamic nature of urban growth and development.

Accurately identifying and categorizing slum areas is crucial. This analysis provides a detailed categorization of slum distributions and their probabilities, enhancing our ability to accurately identify and classify slum areas.

In Figure 6, three cities—Islamabad, Medellin, and Port Au Prince—are shown to highlight each slum category. A city-wide slum probability map for each city is provided on the left of Figure 6, all using the same scale of 1:60,000. The mean slum probability for each city is shown from Figure 5. For each city, two areas of interest (AOI) are provided, each with a scale of 1:5,000. These AOIs use very high-resolution Google satellite imagery instead of our own Planet data to improve the visual readability of the slums. It is important to note that these AOIs were specifically selected from the test dataset and were not used in training the transfer-learning process.

Slum category 1 Islamabad, the capital of Pakistan, has a population of approximately 1.2 million. According to (Rehman et al., 2022), Islamabad's slums are categorized into normal and temporary slums, and are described as typical morphological slums based on our slum categorization schema. In the city-wide slum probability map (Figure 6), many small slum settlements with high slum probability are evident. A closer look at AOIs 1 and 2 in Islamabad shows that these slum settlements have clear borders distinguishing them from formal settlements and non-built-up areas, resulting in high slum probability values.

Slum category 2 Medellin, with 2.6 million inhabitants, falls into the second category of slums, where most morphological traits of slums are present, but slum settlements often have concrete multi-story buildings. Nevertheless many slum settlements in Medellin are exposed to a high vulnerability of landslides (Kühnl et al., 2023).

In Figure 6, the city-wide slum probability map shows high probabilities across large areas, with a gradient indicating a nuanced view of slum settlements. This is particularly visible in AOI 1, where slum areas gradually transition into formal settlements, resulting in a gradual decrease in slum probabilities. In AOI 2, a slum settlement with a lower slum

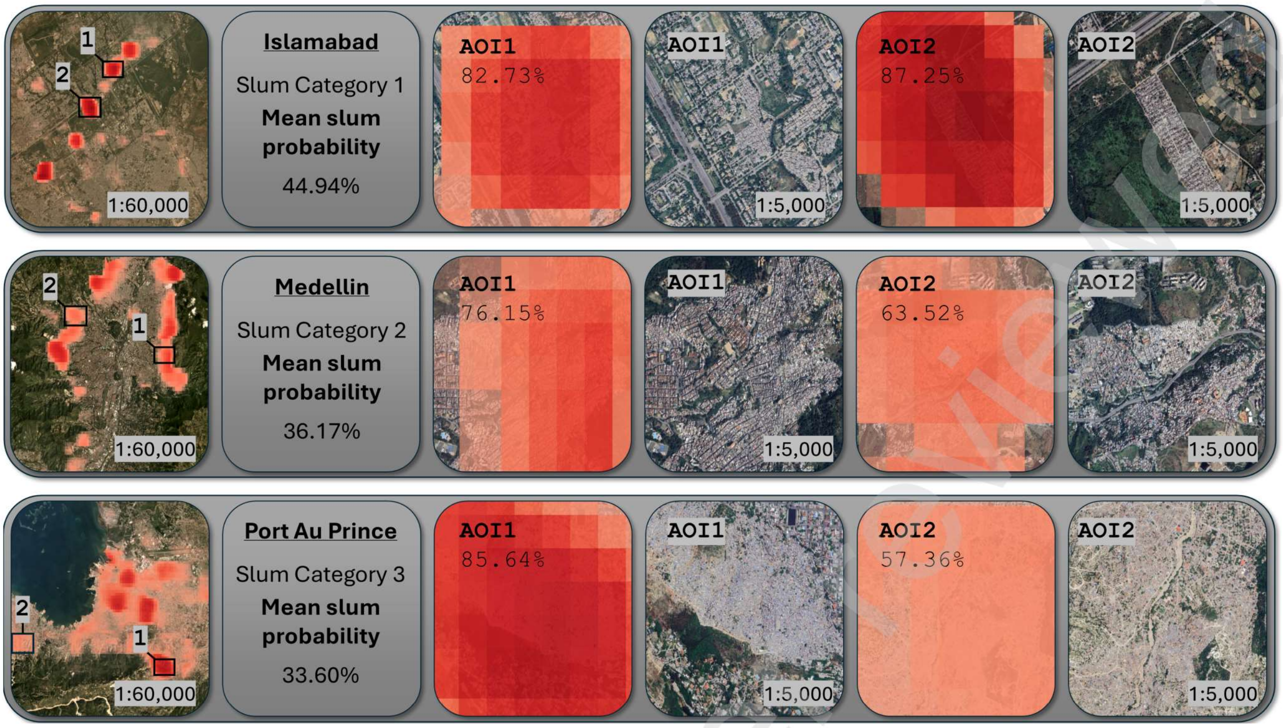

Fig. 6: Examples from three cities, each representing a different slum category. The figure displays the city-wide slum probability on the left and two areas of interest (AOI) on the right. For each AOI, Google satellite imagery is shown. Slum probabilities are provided for the entire city and for each AOI.

probability is observed, lacking the typical heterogeneous and unstructured morphology of typical slum areas.

Slum category 3 According to (Joseph et al., 2014), an overwhelming majority (over 60%) of residents in Port Au Prince, the capital of Haiti, live in very low or low environmental quality conditions in densely populated areas. This is reflected in our city-wide slum probability map, which shows large areas with low probability scores and some distinct high-probability slum settlements. This variability categorizes Port Au Prince into the third slum category due to its intra-urban variability. Only a few settlements exhibit typical slum morphologies, as seen in AOI 1, while many other neighborhoods display atypical slum characteristics, as shown in AOI 2.

Our research goes beyond traditional boundaries in slum mapping. In Figure 6, we highlight how our approach maps slum settlements in various geographical environments, identifying both typical and atypical slum morphologies. We detect different slum categories and their probabilities, including areas where slums transition into formal settlements. This method improves upon traditional binary classifications that only recognize typical slum characteristics.

5.3. Challenges of estimating slum sizes using scene classification

In our study, we did not prioritize the size of slums as a primary metric because we used a scene classification approach. This method focuses on identifying areas with its slum probabilities rather than measuring the physical size of slum settlements. Accurately estimating slum size would be challenging with this approach, as it requires precise boundary delineation, which can be complex and variable as there are just cases where there is no boundary between a slum settlement and a formal settlement. The scene classification approach allows us to understand slum distributions robustly without needing exact size measurements.

As seen in Figure 6, the mapped areas in Port Au Prince encompass most regions of the city. In contrast, the slums in Islamabad are much smaller but achieve high slum probabilities. This disparity suggests that comparing slum sizes directly between different cities is not appropriate and that slum size alone is not a sufficient indicator of slum conditions.

5.4. Data Limitations and Implications for Urban Planning

While our research has made significant strides in applying advanced machine learning techniques and uncertainty-aware methodologies to map slum areas, it is important to critically discuss the limitations and capabilities of the PlanetScope data with its 4.77m resolution. Accurately distinguishing between formal dense settlements and slum settlements at this resolution remains challenging.

Geometric resolution is crucial for mapping slums because it allows for detailed observation of morphological features characteristic of slum areas, such as irregular building patterns, narrow alleys, and high building density. These features are often obscured at lower resolutions, making it difficult to accurately identify slum areas. Despite these challenges, it is possible to map typical slum morphologies using Sentinel-2 data (Wurm et al., 2019; Gram-Hansen et al., 2019; Verma et al., 2019c; Owusu et al., 2024). However, very high-resolution data would greatly enhance our ability to detect and understand all categories of slum morphologies, leading to more precise and actionable insights. The United Nations has emphasized the need for improved data quality to better address urban challenges (United Nations, 2024b), and the call for a global slum repository by (Kuffer et al., 2021) aligns with this objective. Access to very high resolution data would significantly benefit our efforts to better understand and detect slums. This would ultimately improve the effectiveness of urban planning and interventions, allowing for more targeted and impactful strategies to address urban poverty and promote sustainable development.

Our findings have significant implications for enhancing available slum data repositories and for concrete urban planning in cities. The slum probability maps from our study can be integrated into other slum datasets, such as the global slum repository (Kuffer et al., 2021), enhancing their comprehensiveness and accuracy. Additionally, our maps can be compared with existing slum datasets to validate and refine these resources, supporting a centralized and reliable source of slum information crucial for global urban policy and research.

For concrete planning in cities, the slum probability maps help identify non-typical slums that might be overlooked by conventional methods. These maps can guide governments and NGOs to allocate resources effectively, ensuring interventions are more inclusive. By providing a detailed understanding of slum distributions and conditions, the maps can help urban planners develop targeted strategies to improve living conditions, infrastructure, and services, supporting long-term sustainable urban development.

6. Conclusion

Our research demonstrates the effective application of advanced machine learning and uncertainty-aware methods to map slum areas across 55 diverse cities. Using transfer learning and ensemble predictions, we overcome the challenge of limited labeled data and achieve high accuracy in slum detection. The resulting slum probability maps provide valuable insights into urban poverty patterns, aiding policymakers and urban planners in addressing socio-economic disparities.

We applied a coherent methodology to a comprehensive slum dataset across the Global South, including probability estimates for each prediction. This approach offers a nuanced understanding of slum categories within each city, revealing intra- and inter-urban variability. These insights are crucial for tailoring interventions to specific urban needs, leading to more effective urban planning.

By integrating transfer learning with large-scale remote sensing data, our study enhances the understanding of urban environments and promotes sustainable development. The slum probability maps serve as valuable tools for addressing urban poverty and fostering equitable growth. This work highlights the potential of advanced machine learning in transforming urban analysis and addressing the complex challenges of cities in the Global South.

A. Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

B. Data availability

The implementation of all steps in order to reproduce and recreate these results are available through GitHub stark-t/UncertaintyAwareSlumMapping. The slum data, given its sensitive ethical nature, will be shared exclusively upon reasonable request. Regrettably, the PlanetScope remote sensing data cannot be shared due to copyright restrictions. However, our GitHub repository offers resampled Sentinel-2 RGB imagery, serving as a representative example for many of our results.

C. Acknowledgment

This research was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant number: 714087-So2Sat).

The work of X. Zhu is also supported by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001).

D. Author contributions statement

T.S., M.W., H.T., and X.Z. designed the scope of the study. T.S. and H.D. were responsible for curating the dataset, X.Z. and team prepared the PlanetScope remote sensing data, T.S., was responsible for methods and experiments. T.S. took the lead in writing the manuscript and M.W., H.T., H.D., and X.Z. reviewed the manuscript.

References

Abascal, A., Rothwell, N., Shonowo, A., Thomson, D.R., Elias, P., Elsey, H., Yeboah, G., Kuffer, M., 2022. “domains of deprivation framework” for mapping slums, informal settlements, and other deprived areas in lmics to improve urban planning and policy: A scoping review. Comput. Environ. Urban. Syst. 93, 101770. doi:https://doi.org/10.1016/j.compenvurbsys.2022.101770.

Badmos, O.S., Rienow, A., Callo-Concha, D., Greve, K., Jürgens, C., 2018. Urban development in west africa—monitoring and intensity analysis of slum growth in lagos: Linking pattern and process. Remote Sens. 10. URL: https://www.mdpi.com/2072-4292/10/7/1044, doi:10.3390/rs10071044.

Debray, H., Kraff, N.J., Zhu, X.X., Taubenböck, H., 2023. Planned, unplanned, or in-between? a concept of the intensity of plannedness and its empirical relation to the built urban landscape across the globe. Landsc. Urban Plan. 233, 104711. doi:https://doi.org/10.1016/j.landurbplan.2023.104711.

Debray, H., Qiu, C., Schmitt, M., Wang, Y., Zhu, X.X., Taubenböck, H., 2021. Types of morphological configurations of the city across the globe-a remote sensing based comparative approach, in: 26th International Conference on Urban Planning and Regional Development in the Information Society GeoMultimedia 2021, Compet. Cent. Urban Reg. Plan.. pp. 969–978.

Doda, S., Kahl, M., Ouan, K., Obadic, I., Wang, Y., Taubenböck, H., Zhu, X.X., 2024. Interpretable deep learning for consistent large-scale urban population estimation using earth observation data. Int. J. Appl. Earth. Obs. Geoinf. 128, 103731. doi:https://doi.org/10.1016/j.jag.2024.103731.

Ebel, P., Garnot, V.S.F., Schmitt, M., Wegner, J.D., Zhu, X.X., 2023. Uncrtaints: Uncertainty quantification for cloud removal in optical satellite time series, in: Proc. IEEE Int. Conf. Comput. Vis., pp. 2086–2096.

Engstrom, R., Hersh, J., Newhouse, D., 2021. Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-being. World Bank Econ. Rev. 36, 382–412. doi:10.1093/wber/lhab015.

Fisher, T., Gibson, H., Liu, Y., Abdar, M., Posa, M., Salimi-Khorshidi, G., Hassaine, A., Cai, Y., Rahimi, K., Mamouei, M., 2022. Uncertainty-aware interpretable deep learning for slum mapping and monitoring. Remote Sens. 14, 3072.

Friesen, J., Taubenböck, H., Wurm, M., Pelz, P.F., 2018. The similar size of slums. Habitat Int. 73, 79–88. doi:https://doi.org/10.1016/j.habitatint.2018.02.002.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: Balcan, M.F., Weinberger, K.Q. (Eds.), Proc. 33rd Int. Conf. Mach. Learn., PMLR, New York, New York, USA. pp. 1050–1059.

Gram-Hansen, B.J., Helber, P., Varatharajan, I., Azam, F., Coca-Castro, A., Kopackova, V., Bilinski, P., 2019. Mapping informal settlements in developing countries using machine learning and low resolution multi-spectral data, in: AAAI/ACM Conf. AI Ethics Soc., pp. 361–368.

Han, D., Yun, S., Heo, B., Yoo, Y., 2021. Rethinking channel dimensions for efficient model design. arXiv:2007.00992.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. CoRR abs/1512.03385. arXiv:1512.03385.

Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al., 2019. Searching for mobilenetv3, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., pp. 1314–1324.

Joseph, M., Wang, F., Wang, L., 2014. Gis-based assessment of urban environmental quality in port-au-prince, haiti. Habitat Int. 41, 33–40. doi:https://doi.org/10.1016/j.habitatint.2013.06.009.

Kraff, N.J., Taubenböck, H., Wurm, M., 2019. How dynamic are slums? eo-based assessment of kibera's morphologic transformation, in: J. Urban Remote Sens. Event, pp. 1–4. doi:10.1109/JURSE.2019.8808978.

Kraff, N.J., Wurm, M., Taubenböck, H., 2020. The dynamics of poor urban areas - analyzing morphologic transformations across the globe using earth observation data. Cities 107, 102905. doi:https://doi.org/10.1016/j.cities.2020.102905.

Kuffer, M., Grippa, T., Persello, C., Taubenböck, H., Pfeffer, K., Sliuzas, R., 2021. Mapping the Morphology of Urban Deprivation. John Wiley & Sons, Ltd. chapter 14. pp. 305–323. doi:https://doi.org/10.1002/9781119625865.ch14.

Kuffer, M., Pfeffer, K., Sliuzas, R., 2016. Slums from space—15 years of slum mapping using remote sensing. Remote Sens. 8, 455. doi:https://doi.org/10.3390/rs8060455.

Kuffer, M., Thomson, D.R., Boo, G., Mahabir, R., Grippa, T., Vanhuysse, S., Engstrom, R., Ndugwa, R., Makau, J., Darin, E., de Albuquerque, J.P., Kabaria, C., 2020. The role of earth observation in an integrated deprived area mapping "system" for low-to-middle income countries. Remote Sens. 12. doi:10.3390/rs12060982.

Kühnl, M., Sapena, M., Wurm, M., Geiß, C., Taubenböck, H., 2023. Multitemporal landslide exposure and vulnerability assessment in medellín, colombia. Nat. Hazards 119, 883–906. doi:https://doi.org/10.1007/s11069-022-05679-z.

Li, Q., Taubenböck, H., Shi, Y., Auer, S., Roschlaub, R., Glock, C., Kruspe, A., Zhu, X.X., 2022. Identification of undocumented buildings in cadastral data using remote sensing: Construction period, morphology, and landscape. Int. J. Appl. Earth. Obs. Geoinf. 112, 102909. doi:https://doi.org/10.1016/j.jag.2022.102909.

Mahabir, R., Agouris, P., Stefanidis, A., Croitoru, A., Crooks, A.T., 2020. Detecting and mapping slums using open data: a case study in kenya. Int. J. Digit. Earth 13, 683–707. doi:10.1080/17538947.2018.1554010.

Mahabir, R., Croitoru, A., Crooks, A.T., Agouris, P., Stefanidis, A., 2018. A critical review of high and very high-resolution remote sensing approaches for detecting and mapping slums: Trends, challenges and emerging opportunities. Urban Science 2. URL: https://www.mdpi.com/2413-8851/2/1/8, doi:10.3390/urbansci2010008.

Owusu, M., Nair, A., Jafari, A., Thomson, D., Kuffer, M., Engstrom, R., 2024. Towards a scalable and transferable approach to map deprived areas using sentinel-2 images and machine learning. Comput. Environ. Urban Syst. 109, 102075. doi:https://doi.org/10.1016/j.compenvurbsys.2024.102075.

Persello, C., Stein, A., 2017. Deep fully convolutional networks for the detection of informal settlements in vhr images. IEEE Geosci. Remote. Sens. Lett. 14, 2325–2329. doi:10.1109/LGRS.2017.2763738.

Rehman, M.F.U., Aftab, I., Sultani, W., Ali, M., 2022. Mapping temporary slums from satellite imagery using a semi-supervised approach. IEEE Geosci. Remote. Sens. Lett. 19, 1–5. doi:10.1109/LGRS.2022.3180162.

Sachs, J.D., Kroll, C., Lafortune, G., Fuller, G., Woelm, F., 2022. Sustainable development report 2022. Cambridge University Press.

Song, L., Estes, A.B., Estes, L.D., 2023. A super-ensemble approach to map land cover types with high resolution over data-sparse african savanna landscapes. Int. J. Appl. Earth. Obs. Geoinf. 116, 103152. doi:https://doi.org/10.1016/j.jag.2022.103152.

Stark, T., Wurm, M., Zhu, X.X., Taubenböck, H., 2020. Satellite-based mapping of urban poverty with transfer-learned slum morphologies. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 5251–5263. doi:10.1109/JSTARS.2020.3018862.

Stark, T., Wurm, M., Zhu, X.X., Taubenböck, H., 2023. Detecting challenging urban environments using a few-shot meta-learning approach, in: J. Urban Remote Sens. Event, pp. 1–4. doi:10.1109/JURSE57346.2023.10144170.

Stark, T., Wurm, M., Zhu, X.X., Taubenböck, H., 2024. Quantifying uncertainty in slum detection: Advancing transfer learning with limited data in noisy urban environments. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17, 4552–4565. doi:10.1109/JSTARS.2024.3359636.

Stewart, I.D., Oke, T.R., 2012. Local climate zones for urban temperature studies. Bull. Am. Meteorol. Soc. 93, 1879 – 1900. doi:10.1175/BAMS-D-11-00019.1.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: Int. Conf. Mach. Learn., PMLR. pp. 6105–6114.

Taubenböck, H., Kraff, N.J., Wurm, M., 2018. The morphology of the arrival city-a global categorization based on literature surveys and remotely sensed data. Appl. Geogr. 92, 150–167. doi:https://doi.org/10.1016/j.apgeog.2018.02.002.

Taubenböck, H., Debray, H., Qiu, C., Schmitt, M., Wang, Y., Zhu, X., 2020. Seven city types representing morphologic configurations of cities across the globe. Cities 105, 102814. doi:https://doi.org/10.1016/j.cities.2020.102814.

Taubenböck, H., Kraff, N.J., 2014. The physical face of slums: a structural comparison of slums in mumbai, india, based on remotely sensed data. J. Hous. Built Environ. 29, 15–38. doi:10.1007/S10901-013-9333-X.

Taubenböck, H., Weigand, M., Esch, T., Staab, J., Wurm, M., Mast, J., Dech, S., 2019. A new ranking of the world's largest cities—do administrative units obscure morphological realities? Remote Sens. Environ. 232, 111353. doi:https://doi.org/10.1016/j.rse.2019.111353.

Thomson, D.R., Kuffer, M., Boo, G., Hati, B., Grippa, T., Elsey, H., Linard, C., Mahabir, R., Kyobutungi, C., Maviti, J., Mwaniki, D., Ndugwa, R., Makau, J., Sliuzas, R., Cheruiyot, S., Nyambuga, K., Mboga, N., Kimani, N.W., de Albuquerque, J.P., Kabaria, C., 2020. Need for an integrated deprived area "slum" mapping system (ideamaps) in low- and middle-income countries (lmics). Social Sciences 9. doi:10.3390/socsci9050080.

UN-Habitat, 2003. The Challenge of Slums: Global Report on Human Settlements 2003. Earthscan Publications Ltd.

UN-Habitat, Thabit, S., Aguinaga, G., Maroso, R., Mohn, C., Edilbi, B., Donnelly, L., Tandon, A., Caglin, P., Smillie, M., Suqi, F., et al., 2020. United nations human settlements programme (un-habitat) .

United Nations, 2024a. Goal 1: End poverty in all its forms everywhere. https://sdgs.un.org/goals/goal1. Accessed: 2024-06-23.

United Nations, 2024b. Un world data forum: Promoting better data for sustainable development. https://unstats.un.org/unsd/undataforum/programme/. Accessed: 2024-06-23.

Verma, D., Jana, A., Ramamritham, K., 2019a. Transfer learning approach to map urban slums using high and medium resolution satellite imagery. Habitat Int. 88, 101981. doi:https://doi.org/10.1016/j.habitatint.2019.04.008.

Verma, D., Jana, A., Ramamritham, K., 2019b. Transfer learning approach to map urban slums using high and medium resolution satellite imagery. Habitat Int. 88, 101981. doi:https://doi.org/10.1016/j.habitatint.2019.04.008.

Verma, D., Jana, A., Ramamritham, K., 2019c. Transfer learning approach to map urban slums using high and medium resolution satellite imagery. Habitat Int. 88, 101981. doi:https://doi.org/10.1016/j.habitatint.2019.04.008.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019a. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34–45. doi:https://doi.org/10.1016/j.neucom.2019.01.103.

Wang, J., Kuffer, M., Pfeffer, K., 2019b. The role of spatial heterogeneity in detecting urban slums. Comput. Environ. Urban Syst. 73, 95–107. doi:https://doi.org/10.1016/j.compenvurbsys.2018.08.007.

Wurm, M., Stark, T., Zhu, X.X., Weigand, M., Taubenböck, H., 2019. Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 150, 59–69. doi:https://doi.org/10.1016/j.isprsjprs.2019.02.006.

Wurm, M., Taubenböck, H., 2018. Detecting social groups from space – assessment of remote sensing-based mapped morphological slums using income data. Remote Sens. Lett. 9, 41–50. doi:10.1080/2150704X.2017.1384586.

Wurm, M., Taubenböck, H., Weigand, M., Schmitt, A., 2017. Slum mapping in polarimetric sar data using spatial features. Remote Sens. Environ. 194, 190–204. doi:https://doi.org/10.1016/j.rse.2017.03.030.

Zhao, X., Hu, J., Mou, L., Xiong, Z., Zhu, X.X., 2023. Cross-city landuse classification of remote sensing images via deep transfer learning. Int. J. Appl. Earth. Obs. Geoinf. 122, 103358. doi:https://doi.org/10.1016/j.jag.2023.103358.

Zhu, X.X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Haberle, M., Hua, Y., Huang, R., Hughes, L., Li, H., Sun, Y., Zhang, G., Han, S., Schmitt, M., Wang, Y., 2020. So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]. IEEE Trans. Geosci. Remote Sens. 8, 76–89. doi:10.1109/MGRS.2020.2964708.