JULIUS-MAXIMILIANS-UNIVERSITÄT WÜRZBURG (JMU)
Institute of Geography and Geology
Department of Remote Sensing

DEUTSCHES ZENTRUM FÜR LUFT- UND RAUMFAHRT (DLR)
German Remote Sensing Data Centre
Geo-Risks and Civil Security

Deep Learning for Refugee Camps - Mapping Settlement Extents with Sentinel-2 Imagery and Semantic Segmentation

Master Thesis

submitted by

Katrin Wernicke

June 15th, 2023

Supervision:

Prof. Dr. Hannes Taubenböck (DLR, JMU)
Jakob Schwalb-Willmann (JMU)
Matthias Weigand (DLR)

Acknowledgements

Abstract

The number of people forced to flee their homes has more than doubled in the last decade, with over 103 million displaced people by mid-2022. Many seek shelter in refugee camps and informal settlements, which were originally built as temporary facilities. Remote sensing and Deep Learning serve as independent tools for monitoring camps in addition to localised in-situ data. However, research shows an underrepresentation of refugee camps in satellite-based settlement products. This work assessed the applicability of six Deep Learning (DL) models for mapping refugee settlement extents worldwide using semantic segmentation of Sentinel-2 satellite imagery. Two DL architectures, U-Net and FPN, were trained with the encoders EfficientNet-B0, MobileNet-V2, and ResNet-18, and their results were assessed in a comparative analysis. Furthermore, the model accuracies across space and among different morphological structures were evaluated. The results showed that all models faced significant challenges in accurately mapping the settlement extents, although to different degrees. However, some models were successful in localising the camps but overestimated the extents. The analysis revealed that the accuracies varied among camps, and regional clusters of similar accuracies were observed. It is discussed that refugee camps are heterogeneous and complex settlement types which are difficult to delineate in space based on their spatial appearance alone and inconsistent morphological structures. The work highlights the complexity of mapping refugee settlements in a large-scale approach and emphasizes the consideration of morphological differences among camps in image analysis tasks. The findings of this work serve as a foundation for future research on mapping refugee settlement extents with remote sensing for humanitarian aid.

Zusammenfassung

Bis Mitte 2022 waren über 103 Millionen Menschen weltweit auf der Flucht, diese Zahl hat sich in den letzten zehn Jahren mehr als verdoppelt. Viele der Geflüchtete suchen Schutz in Flüchtlingslagern und informellen Siedlungen, welche ursprünglich als temporäre Einrichtungen gebaut wurden. Fernerkundung und Deep Learning (DL) können, zusätzlich zu in situ Daten, als unabhängige Instrumente für das Monitoring von Lagern eingesetzt werden. Die Forschung zeigt jedoch, dass Flüchtlingslager in satellitengestützten Siedlungsprodukten deutlich unterrepräsentiert sind. Daher wurde in dieser Arbeit die Anwendbarkeit von sechs DL-Modellen zur Kartierung der Ausdehnung von Flüchtlingslagern weltweit bewertet, unter der Verwendung von semantischer Segmentierung und Sentinel-2 Satellitenbildern. Zwei DL-Architekturen, U-Net und FPN, wurden mit den Encodern EfficientNet-B0, MobileNet-V2 und ResNet-18 trainiert, und ihre Ergebnisse in einer vergleichenden Analyse bewertet. Darüber hinaus wurden die Modellgenauigkeiten zwischen Camp Standorten und zwischen verschiedenen morphologischen Strukturen verglichen und analysiert. Die Ergebnisse zeigten, dass alle Modelle erhebliche Schwierigkeiten hatten die Siedlungsausdehnungen genau zu Kartieren, wenn auch in unterschiedlichem Maße. Einige Modelle waren erfolgreich bei der Lokalisierung der Lager, überschätzten jedoch deren Ausdehnung. Die Analyse ergab, dass die Genauigkeit zwischen den Lagern variiert, und es wurden regionale Cluster mit ähnlicher Genauigkeit beobachtet. Es wurde erörtert, dass es sich bei Flüchtlingslagern um heterogene und komplexe Siedlungstypen handelt, die allein aufgrund ihres räumlichen Erscheinungsbildes und ihrer uneinheitlichen morphologischen Strukturen schwer räumlich abzugrenzen sind. Die Arbeit verdeutlicht die Komplexität der Kartierung von Flüchtlingssiedlungen in einem großräumigen Ansatz, weswegen bei Bildanalyseaufgaben die morphologischen Unterschiede zwischen den Lagern berücksichtig werden sollten. Die Ergebnisse dieser Arbeit dienen als Grundlage für künftige Forschungsarbeiten zur Kartierung von Flüchtlingssiedlungen mithilfe von Fernerkundung für die humanitäre Hilfe.

Table of Contents

List of Figures

List of Tables

Abbreviations

CE	Cross Entropy
CEw	Weighted Cross Entropy
CNN	Convolutional Neural Network
CV	Computer Vision
DL	Deep Learning
FPN	Feature Pyramid Network
GB	Gigabyte
HR	High resolution
IDP	Internally Displaced People, Internally Displaced Person
ML	Machine Learning
NN	Neural Network
OBIA	Object Based Image Analysis
SC	Structure Catalogue (by Weigand et al. 2023)
SMP	Semantic Segmentation Pytorch
UN	United Nations
UNHCR	United Nations High Commissioner for Refugees
VHR	Very high resolution
WSF	World Settlement Footprint

1 Introduction

1.1 Background

The United Nations High Commissioner for Refugees (UNHCR) estimates that by mid-2022, 103 million people worldwide had been forced to flee their homes because of "conflicts, violence, fear of persecution and human rights violations" (UNHCR 2022b). This is more than double the number of people displaced a decade ago, and the most people displaced since World War II (UNHCR 2022a). Figure 1 shows that the trend strongly intensified in the past decade, with 26.7 million new refugees in 2022 in contrast to 10.5 million in 2012 (UNHCR 2022a). Beyond that did the number of internally displaced people (IDPs) triple in the same time span to 62.5 million at the end of 2022 (iDMC 2023). With newly erupting conflicts and unresolved, ongoing emergencies, such as the current war in the Ukraine, more and more displacements are happening (iDMC 2023). The numbers by UNHCR and the Internal Displacement Monitoring Centre (iDMC) underestimate the total amount of displaced people worldwide, since displacements caused by natural disasters and unreported cases are not included in these figures. Notably, this trend is predicted to further increase with mounting impacts of climate change (iDMC 2022).
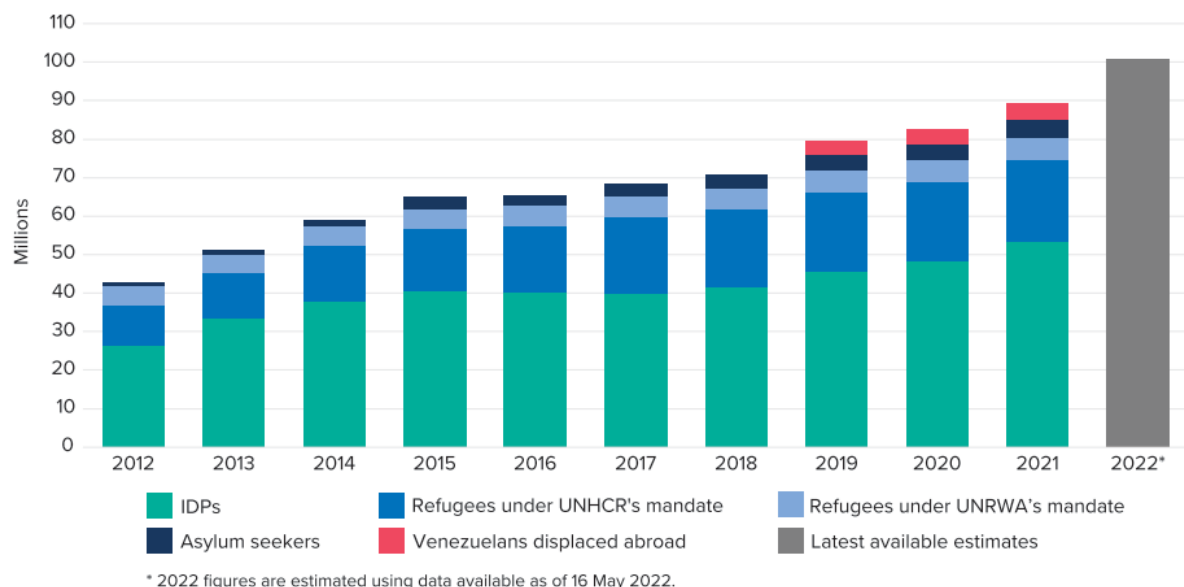
Figure 1: People forced to flee 2012 - 2022. Source: UNHCR (2022a)

The majority of the world refugee population lives in urban areas, whereas over 20 % seek shelter in camps and informal settlements. In 2021, around 4.5 million refugees resided in planned and managed camps, and approximately 2 million were sheltered in self-settled camps (USA for UNHCR 2021).

1.2 Contextualization of Refugee Camps

Camp settlements offer shelter for displaced people and embody a manifestation of their displacement (Weigand et al. 2023). Numerous camps are managed by the UNHCR or other humanitarian aid agencies, who offer formally built camps with containers and stable constructions. However, large migration waves often cause the emergence of inappropriate shelters with self-constructed dwellings and missing infrastructures, either in the proximity of existing camps or at new locations (Kraff et al. 2022).

Refugee camps are intentionally built as temporary facilities to provide immediate assistance and protection (Ramadan 2013). Despite their temporary character, people may end up living there for several years or even decades, if the underlying emergency in a refugee's home country or region remains unresolved (Turner 2016). Therefore, numerous refugee settlements evolved into fully established cities with adequate infrastructure, thriving economies, established structures of civic institutions and governance (USA for UNHCR 2023a).

Refugee camps are semi-formal and semi-legal spaces while legally and physically separated from their hosting country. Their liminal status confers semi-formality, while the sovereignty of humanitarian agencies and the resulting absence of state authority introduces uncertain legality (Ramadan 2013; Turner 2016). Physically, camps often possess fences and walls in order to prevent relationships between the displaced and the local community. However, in practice, those borders are crossed by people, goods and ideas (Turner 2016).

Globally, a noticeable trend of larger and more complex settlements can be observed, as more people are seeking life in urban areas (Taubenböck et al. 2019). This trend is evident in formal as well as in informal and refugee settlements. Given the growing number of displaced people, refugee settlements are becoming increasingly important as places of residence. Therein, one ongoing challenge in refugee settlements is to establish appropriate resource allocation and health care services for refugees (Benz et al. 2019; USA for UNHCR 2021). The United Nations (UN) also identified these challenges and anchored strategies to systematically improve the conditions within their Sustainable Development Goals 2, 3, 6 and 7 (Gao et al. 2022a).

To distribute living resources equally, reliable data on refugee settlements is demanded. Knowledge of the area can serve as a proxy for determining the number of affected people, necessary infrastructure, financial investments, necessary delivery of goods, or other humanitarian aid services (Lang et al. 2020). Furthermore, the land usage and the area are

important to estimate the impact of e.g., a new or growing refugee camp on the local community and the potentially resulting local political problems due to land consumption.

1.3 State-of-the-Art

The UNHCR distinguishes between refugee and IDP camps. Following the 1951 refugee convention, refugees are people who have been displaced over country borders, whereas IDP were displaced within their home country (UNHCR 2010). Based on the definition by the Oxfords Learners Dictionary, a refugee is a displaced person who forcingly had to leave their country or home due to violence, persecution or human rights violation (Hornby 2023). This work will refer to both groups, refugees and IDPs as refugees, and to their shelters as refugee camps or refugee settlements.

For humanitarian organizations like UNHCR, which are hosting several hundreds of refugee camps, information on camp population, structure, size, and expansion is essential in order to achieve informed and sustainable planning. Remote sensing serves as an independent data source, which offers area-wide and up-to-date information. Refugee migration is primarily triggered by conflicts and disasters. Therefore, remote sensing can be deployed as an independent and reliable tool for acquiring data that cannot be collected on the ground, such as precise land use or numbers of dwellings (Avtar et al. 2021). The regular acquisition allows for monitoring over time and historic analysis (Lang et al. 2020). Applications of remote sensing became more diverse with the open and free availability of Landsat imagery. Since the launch of the European Sentinel program, high-resolution imagery became openly available and has since then be used in myriad of applications (e.g. Wendt et al. 2017; Hassan et al. 2018; Braun et al. 2019; Weigand et al. 2020).

Satellite-based human settlement products like the World Settlement Footprint (Marconcini et al. 2020), the Global Urban Footprint (Esch et al. 2017) or the Global Human Settlement Layer (Pesaresi et al. 2013) globally describe human settlement in different spatial resolutions and for different time validity periods. One would expect them to also include information on refugee and IDP camps, as they are spaces of human settlement, despite their semi- to informality. Nevertheless, van den Hoek and Friedrich (2021) identified a strong underrepresentation of refugee camps in these products, regardless of the camps age or size. This underrepresentation highlights the need for approaches describing and analysing refugee camps, as well as other informal settlements.

Therefore, satellite imagery has been employed in diverse applications concerning refugee and IDP camps. In various studies, it is used to derive detailed spatial information on camps through the detection of single dwellings (e.g. Aravena Pelizari et al. 2018; Gella et al.

2022). These footprints can be used as a proxy for population estimations and resource management. The extent of refugee settlements can also be derived from satellite imagery (Hassan et al. 2018; Braun et al. 2019; Benz et al. 2019). As refugee camps can be highly dynamic settlements, their rapid establishment can be observed when using satellite imagery in combination with time-series approaches (Friedrich et al. 2020). Moreover, the impact of establishing and growing camps on the surrounding environment can be assessed, like Braun et al. (2019), as well as Hossain and Moniruzzaman (2021) analysed for Kutupalong camp in Bangladesh.

Refugee camps exhibit heterogeneity in their appearance, leading to variations across different geographical locations. Weigand et al. (2023) conducted an analysis among 285 refugee and IDP camps and identified different morphological structure types. They employed manual visual image interpretation (MVII) of very high-resolution (VHR) satellite imagery. The study revealed some regional clusters of camps with similar structural morphologies. Furthermore, the findings align with Quinn et al. (2018), emphasizing the complex structures of refugee camps. Therefore, it is important to consider morphological variations across camps for image analysis.

This complexity poses a challenge for Machine Learning-based image processing methods. Deep Learning (DL) methods have proven to be particularly effective for contextual object recognition in high-dimensional space (Höser and Künzer 2020). DL is also increasingly used in remote sensing applications, as, for instance, Convolutional Neural Networks (CNNs) can extract highly complex features from satellite imagery (Zhu et al. 2017; Ma et al. 2019). In the context of remote sensing, DL methods have found application for various tasks, including object detection (e.g., Wickert et al. 2021), scene classification (e.g., Chi et al. 2017), instance segmentation (e.g., Gella et al. 2022) and semantic segmentation (e.g., Wurm et al. 2019). The employment of DL demands computational power and large datasets for training. However, the approaches often outperform previous methods like object-based image analysis (OBIA) or common classifiers like Random Forest (Quinn et al. 2018).

Most DL applications for analysing refugee camps approach the issue of dwelling detection within camps. With the knowledge on number and size of structures, the refugee population within a camp can be estimated. This can be substantial for humanitarian planning and managing. DL approaches for dwelling detection achieved promising results accuracy values, by applying semantic segmentation with e.g., the U-Net architecture (Quinn et al. 2018; Gella et al. 2021; Chan et al. 2022; Gao et al. 2022a; 2022b; Ghorbanzadeh et al. 2022), LinkNet and Feature Pyramid Networks (Gao et al. 2022c), MaskR-CNN (Gella et al. 2022) or object detection with a Faster-Recurrent CNN (Wickert et al. 2021). VHR data is

necessary here to also detect small dwellings like tents and huts. However, large-scale applications are challenged by the limited availability of VHR data and their spatial coverage (Quinn et al. 2018).

Expanding the detection of camps beyond one region to another region or country, several studies tested a so-called "transfer learning". Therein, a pretrained model is transferred onto another application, dataset, or region, by only short retraining in the new domain. This approach was tested by several studies. Some compared the model performance of models pretrained with satellite imagery with non-pretrained models for dwelling detection in refugee camps (Gella et al. 2021; 2022; 2023; Quinn et al. 2018). Others also investigated the influence of non-remote sensing weights, i.e. COCO and ImageNet on the models performance to detect camp dwellings (Wurm et al. 2019; Gella et al. 2022). The analyses revealed an increased performance with pretrained weights and therefore show the potential of DL models for applications in refugee related contexts.

Despite the large number of studies on refugee dwelling detection, studies on the expansion of camps are scarce. Remote sensing can help marking out the informal settlement extents. Braun et al. (2019) monitored the extent of Kutupalong camp in Bangladesh over two timestamps, using the synthetic aperture radar (SAR) data of ALOS-2 and Sentinel-1. Other comparable methodologies can be found in the domain of mapping slums or informal settlements (Mboga et al. 2017; Persello and Stein 2017; Liu et al. 2019; Stark et al. 2019; 2020; Wurm et al. 2019). Slums as well as refugee camps can be described as informal settlements, and both structurally differ from the formally built-up settlement.

Fransen and Haas (2022) argue that in relative terms, the number of refugees as share of the world population has not increased over the last decade. However, it is crucial to highlight the magnitude of the absolute number of refugees and the upward trend observed during the past decade. Research addressing the consequences of the growing number of displacements holds great relevance.

1.4 Identifying the Research Gaps

The previous paragraphs showed that several studies have examined refugee and informal settlements using various approaches. However, these studies have mostly focused on analysing individual or a small number of camps. Aside from the Structure Catalogue (SC) of Weigand et al. (2023), no research is available on approaching a global image analysis of refugee settlements. According to van den Hoek and Friedrich (2021), existing global human settlement products fail to depict refugee and informal settlements partly or in some cases entirely. Furthermore, OpenStreetMap (OSM) as one of the largest source for

geographic data worldwide, does not cover refugee settlements to a satisfactory level (Friedrich and van den Hoek 2020). Moreover, those settlement products offer only a static snapshot of human settlement expansion at a certain point in time, whereas refugee camps can be highly dynamic settlements. Despite recent advances, such as the work of Cornebise et al. (2022) presents, data on refugee settlements remains scare. This scarcity highlights a significant gap of topical reference data for refugee camps worldwide. Weigand et al. (2023) and Quinn et al. (2018) emphasized in their works, that refugee camps are very heterogeneous in their structure itself and when comparing the structure of camps across the world. Therefore, analysing camp settlements is highly complex and needs a large amount of data to accurately identify the distinguishing features that set camps apart from formal settlements.

1.5 Objectives of this Work

This study aims to fill the presented research gaps by testing the applicability of DL models for mapping refugee settlement extents in a global approach, using Sentinel-2 imagery and the methodology of semantic segmentation.

The work will focus on the following research questions:

1) *Is it possible to map refugee settlement extents with Sentinel-2 data and Deep Learning? Will the classifier be able to distinguish between refugee camp and formal urban settlement?*
2) *Which DL model architecture performs better for refugee settlement extent mapping? Which model encoder is best suited for this task?*
3) *Are there patterns in the model's performance among regions or the structure types identified by Weigand et al. (2023)? Does the methodology work better for some structure types than others?*

The following chapter of this thesis will introduce an overview on the studied areas and the theoretical background on DL, neural networks, and specific model architectures. The required data for the applied methodology, and the implemented DL model training with the preceding training data generation are presented in chapter 3. Also included is the presentation of the subsequent analysis of the models' results by applying an accuracy assessment. Chapter 4 presents the findings of this study, which are critically discussed in chapter 5 and concluded in chapter 6.

2 Theoretical Background

2.1 Study Areas
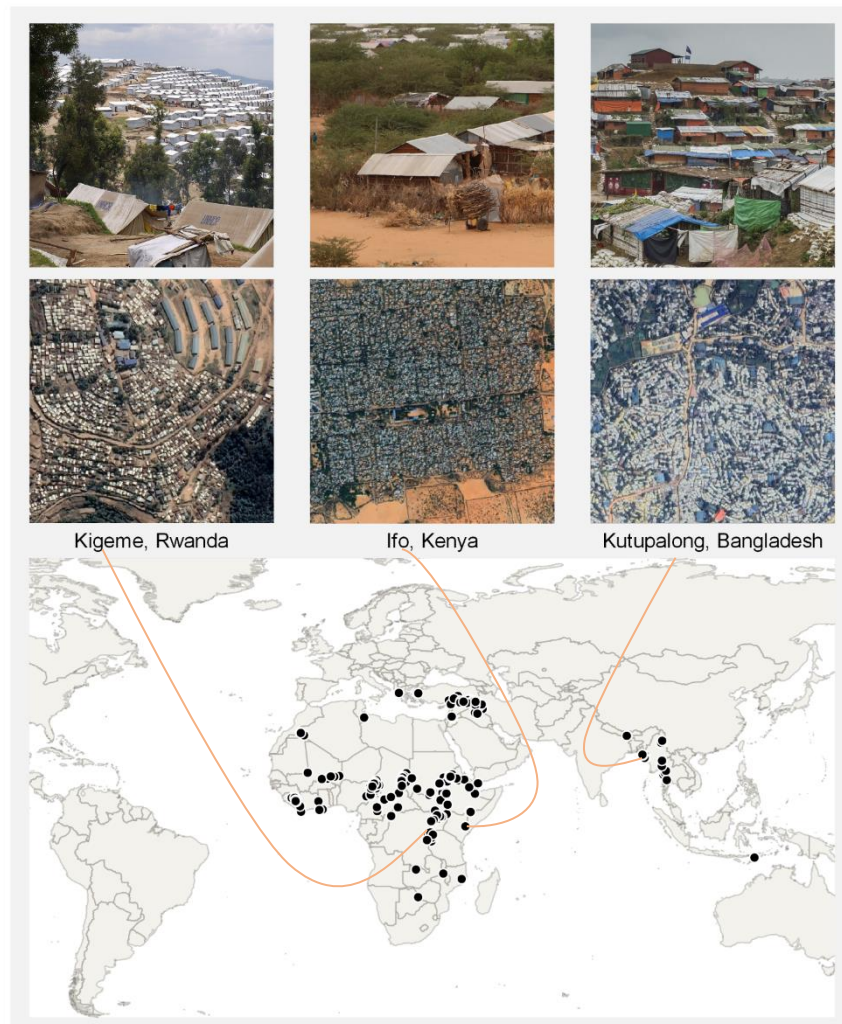
Figure 2: Locations and examples of refugee camps. The top shows examples of three refugee camps, and their appearance in VHR satellite imagery (middle row). The map shows all refugee camp locations included in this study. Satellite data: © Google Earth, 2023 CNES / Airbus. Images: (top-left) ©Oxfam East Africa, CC 2.0; (top-mid) ©UN-Habitat/Julius Mwelu; (top-right) ©Tanvir Murad Topu / World Bank, CC BY-NC-ND 2.0.

This study examined 275 refugee and IDP camps across 40 countries (see Figure *2*). The selection of camps was based on the Structural Catalogue dataset developed by Weigand et al. (2023), which analysed 285 camp locations worldwide. For a camp location to be included in my analysis, one of two criteria had to be met. Either, a meaningful OSM geometry had to exist, describing the extent of the camp. Alternatively, the boundary of the refugee camp had to be visibly identifiable from VHR satellite data, and camps should be clearly distinguishable from formal settlement.

The selected camps are distributed across three continents - Africa, Asia, and Europe. The majority of the camps (70 %) are located in Africa, 28 % in Asia, and only three camps are located in Europe, specifically in Greece. The largest number of camps analysed in this study are located in Nigeria, followed by Chad. For further details, refer to Appendix 1.

This study included several of the largest refugee camps worldwide. As of 2023, the biggest refugee camp in the world is Kutupalong, situated in the Cox's Bazar district of Southern Bangladesh. This camp complex houses over 800 thousand individuals, primarily Rohingya people (Benz et al. 2019). The Muslim minority was displaced from Myanmar, where they are facing widespread persecution. This camp has expanded significantly since the violence in Myanmar reached its peak in 2017, forcing 742,000 of people to leave Myanmar (USA for UNHCR 2023b). A list of the world's largest camps included in this study can be found in Appendix *2*.

The studied refugee camps are very diverse in their age, settlement structures, dwelling types, or settlement size. No overarching pattern was used for selecting the camps. The selection of camps is further explained chapter 3.1.1.

2.2 Neural Networks, Deep Learning and Semantic Segmentation

The following chapter aims to introduce the methodology applied in this study and therefore gives an overview on the basic concepts of DL, neural networks, and semantic segmentation.

DL is a method of Machine Learning (ML) and describes "a learning algorithm based on neural networks" (Ma et al. 2019, p. 167). The term 'deep' in Deep Learning refers to the stacking of multiple layers in a network architecture. The depth of the network increases with a growing number of layers. In contrast to common ML techniques, neural networks enable end-to-end learning, where they learn and extract rich and complex features from the training data, without the need of manual feature engineering (Zhu et al. 2017; Höser and Künzer 2020).

DL has been tested effective for Computer Vision (CV) tasks, which describe the field of study on image processing (Goodfellow et al. 2016, p. 447). One common task in CV is semantic segmentation, and it is often applied with DL. Therein, the goal is to categorize each pixel of an image, by learning contextual information from the input. The result is a segmentation map of the same size and resolution as the input image, where each pixel is assigned to a specific class. It differs from image classification, which describes the assignment of a label for a whole image. CNNs are most commonly used for semantic segmentation tasks (Chai et al. 2021) and were therefore applied in this study.

For a proper understanding of the methods, this section introduces the basic concept of NNs and CNNs, and finally the utilized model architectures U-Net and FPN with the encoders MobileNet-V2, ResNet-18 and EfficientNet-B0.

2.2.1 Introduction to Neural Networks

A neural network (NN) is a linear network consisting of connected nodes, called neurons (Skansi 2018, 79f.). Different kinds of neural networks exist, like the classical Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN) for processing sequential data, Generative Adversarial Networks (GAN) which are mainly used to generate data, and Convolutional Neural Networks (CNN) (Höser and Künzer 2020). The functionality of NNs is visualized in Figure 3 and described in the following.
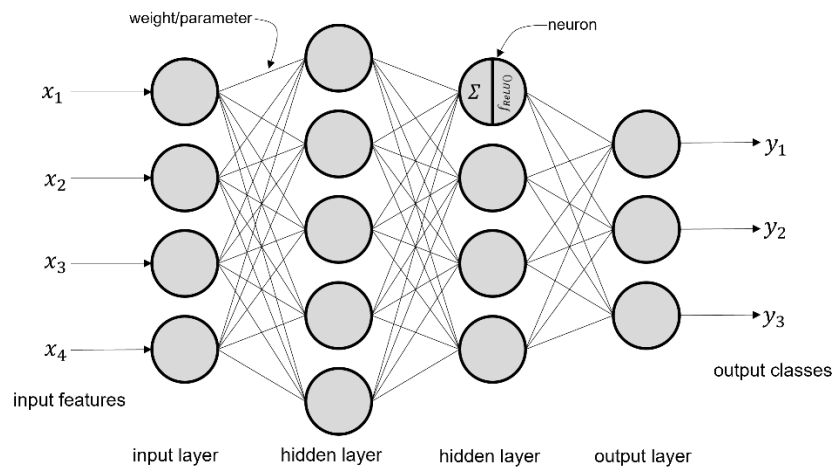
Figure 3: Schematic representation of a fully connected neural network. Adapted from Höser and Künzer (2020).

A NN consists of a sequence of fully connected layers. One layer holds neurons which are connected to each neuron of the next layer by weights or parameters. The values of the neurons get multiplied by the weights and the resulting values are passed onto the next layer via the connections. All incoming values are summed up by each neuron and a (non-) linear function, the activation, is executed (Skansi 2018, 80f.).

Neural networks are trained by using iterative, gradient-based optimization algorithms which aim to minimize the cost function, i.e., the loss. During forward-propagation, the input x provides initial information and propagates through the network and finally produces an output $\hat{y}$. A scalar cost, the loss, is produced. With backpropagation, the information of the loss flows backwards through the network to compute the gradient. On basis of the gradient, the network can be optimized with the stochastic gradient descent. Here, the learning rate

is a hyperparameter which controls the degree to which the weights get adjusted with respect to the loss gradient (Goodfellow et al. 2016, p. 424). The overarching goal in neural network training can be defined as the reduction of the loss function to increase the model performance.

2.2.2   Convolutional Neural Networks

CNNs are a specific kind of neural network, designed to process structured arrays of data such as images (LeCun et al. 2015). CNNs are widely used among different applications, ranging from natural language processing to image segmentation and object detection (Höser and Künzer 2020).

The core strength of a CNN lies in the convolutional layer. Here, a convolutional kernel slides over the 2D input array like a moving window and thus is able to detect patterns. The local sensitivity of the kernel function enables analysing images not only on pixel-level, but also taking local connectivity and neighbouring relations of the pixels into. By sequentially stacking convolutional layers, the CNN can learn hierarchical features, such as textures, in the same way as natural signals are composed. Hence, representational features of image data can be learned to an increasingly abstract degree, while also being aware of their values and local arrangement (Höser and Künzer 2020).
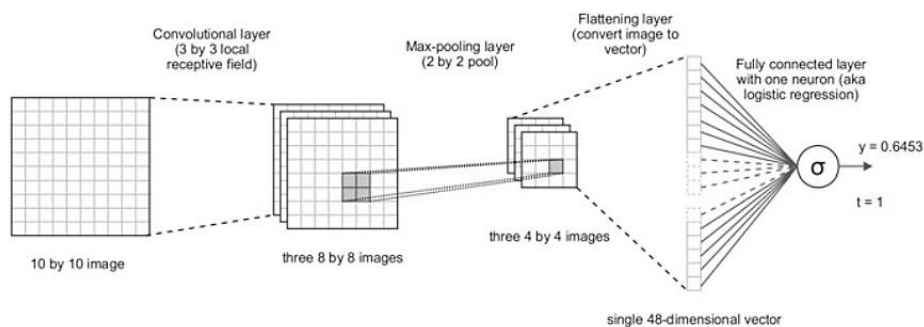
Figure 4: Convolutional Neural Network. Reprinted from Skansi (2018), p.126.

Convolutional Neural Networks consist of the three parts: Input, convolutional backbone and classifier head (see Figure 4). The input, a 2D array, is passed through a sequence of hidden layers. Those layers include convolutions, activations, and pooling operations, aiming to extract high level features. The pooling operation reduces the resolution of the input by factor 2 (if applied with a stride of 2) and introduces translation invariance. The subsequent classifier consists of fully connected layers and is located at the end of the

backbone. The features extracted in the convolutional backbone are then used to classify them into output classes (Ma et al. 2019; Höser and Künzer 2020).

There are multiple advantages of applying CNNs. CNNs have less parameters than fully connected networks, which were used for image analysis prior to CNNs, because network parameters and weights are shared during convolutions (Goodfellow et al. 2016, p. 251). Having less weights allows CNNs to converge faster to a minimum loss, while reducing the probability of overfitting, using less memory and taking less training time (Skansi 2018, p. 126). Another advantage is, that CNNs allow for effective transfer learning. With transfer learning, the weights learned in one network are utilized in another network to improve generalization there. This can reduce training time significantly. As CNNs are observed to learn simple and generic features in the early layers, they are widely applied for transfer learning (Goodfellow et al. 2016, 543f.).

In the classical CNN architecture, the feature maps are resized to a lower spatial resolution due to the pooling layers. At the end of the network stands a fully connected layer, which has one output neuron, resulting in 1D outputs (Skansi 2018, p. 124). For image analysis targeting precise localisation of features in an output map, several challenges arise.

Firstly, in semantic segmentation, where accurate predictions are made at pixel level, it is necessary to localise the learned features in a 2D output map, in contrast to the 1D outputs of the network. However, accurate information on the feature location in the image gets lost due to the resizing of feature maps in the convolutional and pooling layers. Additionally, for pixelwise classification, understanding the contextual relationships between pixels with their neighbouring pixels and segments is crucial (Höser and Künzer 2020). However, as the feature depth increases, the image resolution decreases, posing a trade-off for remote sensing research.

To overcome these challenges, encoder-decoder networks constitute a suitable architecture (Zhu et al. 2017). They combine the advantages of CNNs while still considering the spatial location and contextual relationships of the features. As the goal of this study is to map refugee camps with semantic segmentation from Sentinel-2 images, the encoder-decoder architecture was identified as ideal solution. In the following part, two encoder-decoder architectures are presented: The U-Net and the Feature Pyramid Network (FPN).

2.2.3   U-Net

The U-Net architecture was created by Ronneberger et al. (2015), originally implemented and applied for biomedical image segmentation. It is now widely used in all kinds of segmentation tasks (Höser and Künzer 2020). The name U-Net originates from its U-form, which is created by the encoder and decoder paths (see Figure 5).

Figure 5: U-Net architecture, reprinted from Ronneberger et al. (2015).

The encoder part serves as a convolutional backbone, responsible for extracting features from input data, by applying a sequence of convolutions, activations, and pooling layers. The Rectified Linear Unit (ReLU) is the here applied activation function and also in other applications widely used. With its' function $f(x) = \max(0, x)$, with x as input value, positive input values will be directly output, whereas negative values will be passed as zero. Models which used ReLU appear to be easier to train and often show better performance (Goodfellow et al. 2016, p. 189).

It is common practice to utilize existing CNN architectures as convolutional backbones in the U-Net architecture. The subsequent decoder path enables precise feature localisation. The semantically rich feature maps, resulting from the feature extraction, first get upsampled or deconvoluted to a larger spatial size. Additionally, features get localised by combining the semantically rich feature maps with spatially higher resolution maps from earlier stages in the encoder part via shortcuts or skip connections. Once the input resolution is restored, pixelwise predictions are made to create the final segmentation map (Ronneberger et al. 2015).

2.2.4 FPNs

The Feature Pyramid Network was developed by Facebook AI Research, Cornell University and Cornell Tech as a "pyramidal hierarchy of deep convolutional networks" for object detection (Lin et al. 2016). The FPN architecture consists of a feature extracting convolutional backbone and a top-down hierarchical path, creating feature maps at every level. The network possesses semantically rich features at all levels, by applying lateral connections (see Figure 6).
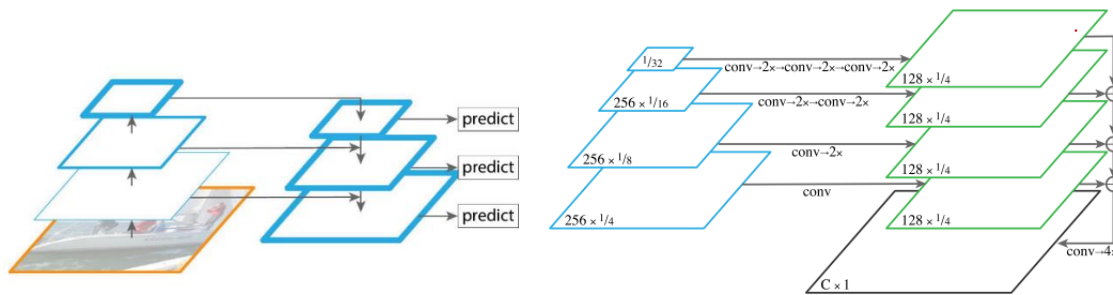
Figure 6: Classical FPN and the Panoptic FPN. FPN reprinted from Lin et al. (2016) (left); on the right the extension for the Panoptic FPN for semantic segmentation by (Kirillov et al. 2019). The left part of the Panoptic FPN corresponds to the feature maps in the classical FPN (right part).

The bottom-up path of the network computes a feature hierarchy with feature maps at different scales. Thus, layers producing output maps of the same size belong to the same network stage, one pyramid level exists for each stage. The output of the last layer is used as reference set for the feature maps within one stage, since the deepest layer of the stage possesses the richest and most complex features. As the feedforward computation of the backbone CNN is independent of the top-down process, common convolutional architectures can be chosen for the bottom-up path.

In the top-down path, higher pyramid levels get upsampled to spatially coarse but semantically rich feature maps. The upsampling method applied is nearest neighbour interpolation (Höser and Künzer 2020). By using lateral connections, feature maps of the same size are merged from the bottom-up and top-down path. The high resolution but lower-level semantic feature map of the bottom-up path helps localise the semantically strong features derived from the top-down feature maps.

For object detection applications, a classification map is created at every stage in the pyramid by a final 3x3 convolution, resulting in proportionally sized maps at multiple levels (see Figure 6, left). For semantic segmentation, the methodology was adapted by Kirillov et

al. (2019) with their Panoptic FPN (see Figure 6, right). Therein, each top-down feature map gets upsampled by convolutions and bilinear upsampling to ¼ of the original image size. The outputs are merged into one stack and transformed into a final output layer (Kirillov et al. 2019).

2.2.5 Encoder architectures

In this study three commonly used CNN architectures were compared. The architectures serve as convolutional backbones in the U-Net and FPN models. They differ in their model architectures and therefore are differently suited for the task of mapping refugee settlement extents.

ResNet

The Residual Network (ResNet) architecture and residual blocks were developed by He et al. (2015) to solve a common problem with CNNs. When increasing the depth of the model, the accuracy will start decreasing at a certain point. Therefore, a ResNet learns residual functions with reference to the layers input, instead of learning the underlying mapping. These residual blocks enable the ResNets to become deeper without leading to a degradation model (Höser and Künzer 2020).

Figure 7(a) shows the core of the ResNets, a residual learning block Figure 7. Two or more stacked convolutional layers are bypassed by the skip connection, which adds from the input layer to the output (Zhu et al. 2021). Thus, the function describing the convolutional operations of the main trunk $F(x, W_i) = H(x, W_i)$ is reformulated to a residual function $H(x, W_i) = F(x, W_i) + x$, where x is the input from the residual connection. Since x is already known due to the residual connection, approximating the weights for $F(x, W_i)$ is considered easier (Höser and Künzer 2020). Thus, there is empirical evidence that these models are easier to optimize and are gaining accuracy from increasing depth (He et al. 2015). ResNets are constructed by stacking the residual blocks sequentially. In this study, ResNet-18 was used, where the name already indicates that the network consists of 18 layers.
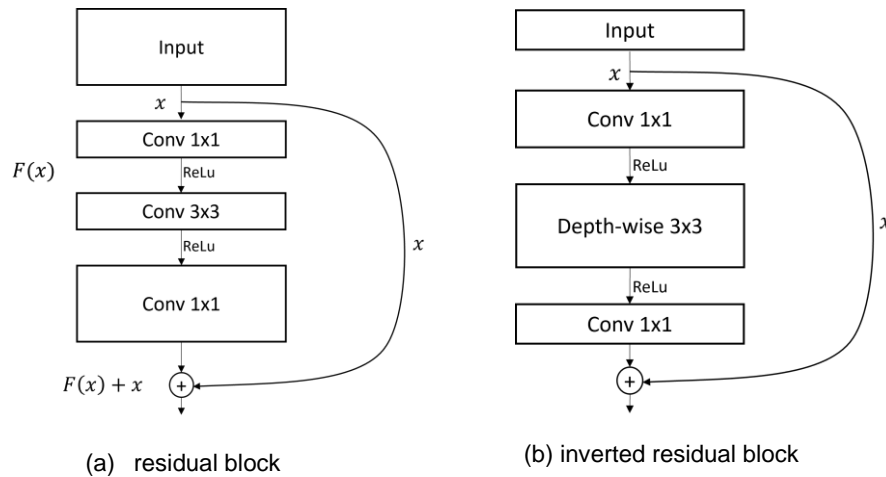
Figure 7: The basic building blocks of ResNet and MobileNet. (a) shows a residual block, (b) an inverted residual block, as used in the MobileNet -V2. Adapted from (a) He et al. (2015) and (b) Sandler et al. (2019).

MobileNet

The MobileNet networks were designed by Sandler et al. (2019) to work on small, portable devices like mobile phones or robots. Thus, they aimed for a small number of parameters and the first version MobileNet-224 had only 4.2 million parameters. The second version constitutes an improvement to the first MobileNet, by using inverted residual blocks (Figure 7b). This means, the first 1x1 convolution in the block is expanding the depth of the feature map for the adjacent 3x3 convolution, in contrast to the first convolution in a residual block (see Figure 7a). This results in thinner outputs of the blocks. By connecting the bottleneck layers, the surrounding residual connection adds input to the output (Höser and Künzer 2020).

EfficientNet

The EfficientNets are a network family introduced by Tan and Le (2019). Similar to He et al. (2015), they aimed to create a network architecture which can be upscaled efficiently. The goal was to upscale depth, width, and image resolution all at once. They found the compound scaling method, which uniformly upscales all three objectives with a set of fixed scaling coefficients. Already existing networks can be upscaled very efficiently with this method, but they also built their own architecture, the EfficientNet. The base model EfficientNet-B0 was constructed by multi-objective neural architecture search that optimizes both accuracy and computer performance. On basis of the B0, the versions B1-B7 could be upscaled very efficiently by applying the compound scaling method. The EfficientNets achieve much better accuracies and outperformed state of the art CNNs for image recognition tasks (Tan and Le 2019).

2.2.6 Accuracy Assessment for Imbalanced Datasets

Class imbalance is a widespread challenge that must be taken into account in DL applications (Johnson and Khoshgoftaar 2019). An imbalance in the dataset occurs, when the distribution of samples among the classes is not equal. When class imbalance arises in training data, the majority class is typically overclassified because of its major occurrence. Consequently, the minority class, often the class of interest, is misclassified and underrepresent (He and Garcia 2009). In DL, class imbalance can be attempted to be reduced through, for example, data-level methods, where data sampling techniques are applied. Balancing the datasets helps to increase the accuracies (Stark et al. 2019). Furthermore, when evaluating performance and classification accuracies of deep neural networks, this phenomenon needs to be taken into account (Johnson and Khoshgoftaar 2019).

A **confusion matrix** is a common and widely used error matrix presenting the number of correct and incorrect predictions made by a classifier. The matrix size complies to n*n, where n is the number of classes in the dataset. Several commonly used accuracy metrics can be derived from this matrix (He and Garcia 2009).

The **Overall Accuracy (OA)** (a) analyses the number of correctly predicted pixels (True Positives, TP) against all pixels included. Thus, this metric does not consider class imbalance as it is biased towards the major classes, which undermines the importance of the minor classes (He and Garcia 2009).

$$\text{(a) } OA = \frac{TP+TN}{TP+TN+FP+FN} = \frac{correct\ predictions}{all\ predictions}$$

with TP = true positives; TN = true negatives; FP = false positives and FN = false negatives

Different, class-based accuracy metrics assess the accuracy for each class involved in the analysis. By that, class imbalance induced biases in the accuracy values can be prevented (He and Garcia 2009). The following paragraphs will introduce some commonly used class-based accuracy metrics that allow for deeper understanding of how individual classes perform. This is especially relevant in imbalanced datasets, especially if it is to assume the class(es) of interest are in minority.

The **Intersection over Union (IoU)** (b) is calculating the area of intersection between a prediction and the label divided by the area of union between the prediction and label. The metric values range between 0 and 1, where 1 means perfect overlap (Taha and Hanbury 2015).

$$\text{(b) } IoU = \frac{Area\ of\ Intersection}{Area\ of\ Union}$$

Precision (c) is a metric for exactness, that indicates the proportion of the positively predicted samples that are true positives (He and Garcia 2009). The precision can be seen as an indicator for overclassification, as a low precision indicates a high number of false positive predictions. Hence, a high precision value describes a model with a lot of true positives and low overclassification.

$$(c) \; Precision = \frac{TP}{TP+FP}$$

Recall (d), also called True Positive Rate or Sensitivity, describes the number of predicted positives among all pixels that are true positives. It is a measure of completeness (He and Garcia 2009). Low values of recall points to underclassification, where many actual positives are not assigned as positive. A trade-off between precision and recall exists, which needs to be evaluated individually by application (Johnson and Khoshgoftaar 2019).

$$(d) \; Recall = \frac{TP}{TP+FN}$$

The **F1 Score** (e), also known as the Dice score, can be understood as the harmonic mean between precision and recall. It describes twice the area of intersection between prediction and label divided by the total number of pixels in both prediction and label (Johnson and Khoshgoftaar 2019). There is some ambiguity in case of a medium F1 score, as information gets lost if either precision or recall possess high values. Hence, recall and precision need to be considered additionally to correctly interpret the F1 score.

$$(e) \; F1 = 2 * \frac{precsion*recall}{precision+recall}$$

The False Positive Rate (FPR) (f) describes the amount of incorrectly labelled samples among the total number of negative samples. The metric shows the number of true negatives mislabelled by the classifier and can therefore suggest overclassification and highlight the model's confusion among classes. Higher values are indicating worse classification accuracy (Taha and Hanbury 2015).

$$(f) \; False \; Positive \; Rate = \frac{FP}{FP+TN}$$

3 Data and Methods

3.1 Data

For mapping the extents of refugee settlement with DL, it is essential to generate training data. It is used by the DL models to learn meaningful features and to validate the performance of the models. In semantic segmentation tasks, images with two-dimensional reference labels are required.

The camp locations addressed in this thesis were derived from the work by Weigand et al. (2023). In this study, reference data for refugee camps outlines were collected from various sources, including OSM[1], and visual image interpretation from Sentinel-2 and Google Earth[2] VHR imagery. Leveraging the full stack of historical data, the temporal dynamics of camp extents were captured, too. The goal was to provide an extensive reference dataset composed of all possible Sentinel-2 scenes for all camps as well as label images referring the land use classes. The following subchapters present the data products included in this work and describe the creation of the dataset holding the camp settlement boundaries.

3.1.1 Structure Catalogue and People of Concern database

The SC was developed by Weigand et al. (2023) and describes the morphological settlement structure of 285 refugee and IDP camps worldwide. The camp locations were derived from the UNHCR spatial database on locations for People of Concern (PoC), which held over 13,000 sites worldwide as of October 2020 (Weigand et al. 2023; UNHCR and UNHCR GIS DATA 2023). Weigand et al. (2023) subset this dataset by several criteria:

- All entries which were annotated with the category 'camp' were chosen.
- The coordinates of a camp location had to be located within or in proximity to a camp-like structure.
- The morphological structure exhibits a coherent, but temporary pattern of settlement.
- For locations close to a city, camp structures and formal urban settlement are clearly distinguishable.
- The camp location was referenced in any other data (UNHCR documents, OSM, Google Maps, …).

---

[1] openstreetmap.org
[2] earth.google.com/web/

The resulting 285 refugee and IDP camp locations were analysed by visual image interpretation of historical and recent VHR satellite imagery. The authors assessed several morphological features on building, block and camp level and validated them against reference data. These features were aggregated and simplified into the two major categories compactness and geometric arrangement. Both variables can take on the values of low, medium or high (Figure 17).

The measurement of compactness (g) takes into account the parameters shelter area ($AREA$), building distance ($DIST$), shelter density ($DENS$) and homogeneity of densities in the blocks (HOM). It describes the compactness and density within a camp.

$$(g)\ Compactness = \frac{\frac{1}{3}(AREA+(|DIST-4|)+DENS+HOM)}{4}$$

$$(h)\ Arrangement = \frac{\frac{1}{3}(ORI+STR+PATH)}{3}$$

The geometric arrangement (h) gets derived by combining the morphological features of building orientation (ORI), camp structure (STR) and path structure ($PATH$). The variable describes the geometric orientation along those features.

In this work, the SC served as basis for identifying potential camp locations suitable for CV analysis. Furthermore, the different structure categories were used to quantify the extent to which structural variations across camps worldwide would influence the applicability of CV methods for mapping camps. The expectation was that certain types of camps were better detectable in Sentinel-2 imagery than others.

Weigand et al. (2023) describe the structural types as the combination of the metrics compactness and geometric arrangement. For simplicity, the resulting 9 categories were numbered, to be able to refer to them (see Figure 17).

3.1.2 OpenStreetMap and Refugee Settlement Boundaries

Refugee settlement extents for 275 refugee and IDP camps were collected and created in this study. Potential camp locations were derived from the SC, and OpenStreetMap (OSM) data as well as manual visual image interpretation (MVII) were used to find the extents.

OSM is a database for free and open geographic data, with global coverage, thematic richness and temporal depth (van den Hoek et al. 2021). The data is collected via open collaboration by a community of voluntary contributors. The database is broadly used for humanitarian applications (see Herfort et al. 2021). Each feature provided by OSM possesses a tag, describing a geographic attribute, next to its basic data structure. For this

study, polygon geometries representing refugee camp settlement extents were searched. Thus, the tag 'amenity = refugee_site' was identified as the suitable tag.

The first step was to query the OSM database for polygon geometries with the tag 'amenity = refugee_site', via the Overpass API[3]. The resulting polygon collection was filtered to only include those geometries within a 1 km proximity to the camp location. Additionally, a visual assessment was performed by comparing the geometries to VHR Google Earth imagery. Criteria for inclusion here were (a) the geometry spatially covers the camp and (b) the size of the geometry matches the visually apparent camp. While also digitizing camp extents from the OSM basemap, 180 useful geometries could be found.
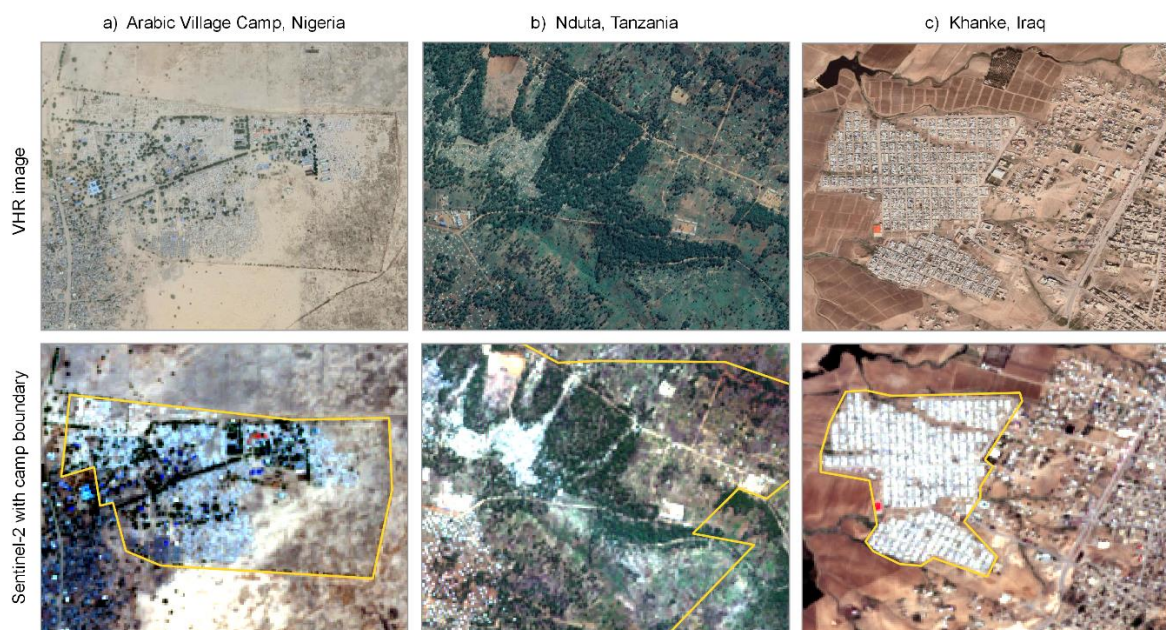
Figure 8: Examples showing MVII for camp settlement extents. (a) shows that the camp is clearly demarcated by a wall or something similar, but the area of built-up is visibly smaller than the extent; (b) shows different dwelling densities and the challenge of delimiting the camp; (c) shows the visible difference between camp and formal settlement. VHR imagery from Google Earth; (a) ©2023 Airbus, (b) ©2023 CNES/Airbus, (c) ©2023 Maxar Technologies.

In the next step, missing extent geometries were digitized manually. For reproducibility, it is important to define a guide on how to identify or rather define camp settlements and their extents in satellite imagery. As mentioned before, refugee camps around the world do not follow any coherent pattern or structure. Therefore, they do not exhibit the same dwelling types or any consistent type of boundary. Thus, in this work, a refugee camp is defined (in satellite imagery) as spatially coherent settlement, which can differ spectrally and

---

[3]overpass-turbo.eu

geometrically from formal settlement, as shown in Figure 8c. Camps are spatially not only defined by the dwellings, large fields of bare ground or infrastructural houses like schools can be a part of the camp, as shown in Figure 8a. Extent boundaries can be found through apparent demarcations like fences and walls, streets, and rivers, but also apparent land cover changes. For camp locations with unclear boundaries, the extent of built-up area can be helpful (Figure 8b).

With this guide, new geometries were created and existing geometries were adapted according to their change over time. Due to camp growth, deconstruction or change of dwelling type e.g., from tents to containers, camp settlement extents can change over time. In these cases, multiple boundary polygons were collected to account for this multitemporal dynamic. By comparing the extents with VHR Google Earth and HR Sentinel-2 imagery, validity periods could be identified for each boundary. The temporal validity period was noted per month in the attribute table of the geometries.

In total, the resulting dataset holds 369 polygon geometries describing camp settlement extents for 275 camps around the world. For 10 camps included by Weigand et al. (2023), accurate boundaries could not be created. The observation period is from January 2016 until September 2022.

3.1.3 Satellite imagery

Optical, multispectral satellite imagery was used extensively for the training data generation in this study. Google Earth offers high to very-high resolution satellite imagery and combines VHR commercial satellite images such as Worldview, with freely available data from the Landsat and Copernicus missions. A function of Google Earth is the possibility to display historical data, however, the data availability over time strongly differs among regions. Google Earth is freely accessible for visual image interpretation or via the software program Google Earth Pro[4].

Besides Google Earth imagery, Sentinel-2 data was used in this work. The satellite imagery collected by the European Sentinel-2 satellites of the Copernicus mission offers freely available data on the Earth's surface. With up to 10 m spatial resolution, 13 bands and a swath width of 290 km, the high resolution imagery is desirable for large scale applications (Weigand et al. 2020). The constellation of two satellites allows a high revisit time of 5 days at the equator. Optical data is available since mid to end of 2015, depending on the region. This study included Sentinel-2 imagery acquired from both satellites between January 2016

---

[4] google.com/earth/versions/#earth-pro

until September 2022. To reduce the influence of atmospheric interference, Level-2A (L2A) data was employed (ESA 2015).

In this work, satellite data was downloaded in an automatized way, by browsing the Sentinel-2 catalogue in the SpatioTemporal Assets Catalogue (STAC)[5]. There, the users can access data items or only subsets, as well as their metadata in form of Cloud Optimized GeoTiffs (COGS), without the need of downloading the whole scene. This benefits this study as the camps were located all around the globe, which required to access many unique Sentinel-2 scenes. Not having to download each image in its entirety reduced the necessary data volume immensely.

For each camp location, STAC items were scanned for imagery with less than 10 % cloud cover. These images were subset by a bounding box of spatial size of 10x10km around the camp location. As cloud effects could still decrease the quality and usability of the subsets, the Scene Classification Layer (SCL) of the Sentinel-2 L2A product was used to assess the impact. High quality subsets were then downloaded, with a maximum restriction of 100 images per location due to memory limitations. This resulted in a final number of 15.112 RGB Sentinel-2 scenes of 10x10 km^2 size for 275 camp locations, for the period of January 2016 to September 2022. For the remainder of this thesis, the 10x10 km^2 satellite imagery subsets will be referred to as scenes.

3.1.4 World Settlement Footprint 2015

The World Settlement Footprint (WSF) is a global binary mask depicting the extent of human settlement in 10 m spatial resolution, developed by Marconcini et al. (2020). The extents were derived from approximately 300,000 scenes of multitemporal Landsat-8 optical and Sentinel-1 radar imagery, referring to the years 2014 and 2015. The authors argue that the WSF2015 outperforms similar datasets like the Global Urban Footprint (Esch et al. 2017) and is very robust and accurate. In 2021, DLR published the WSF 2019, together with ESA and Google Earth Engine. The updated settlement mask was derived globally from temporal means of Sentinel-1 and Sentinel-2 data (Marconcini et al. 2021).

The WSF product was chosen for this analysis, as both versions (WSF2015 and WSF2019) are the newest and most accurate state-of-the-art settlement products (Marconcini et al. 2021). As the training data was created with a focus on temporal validity, the first version was chosen over the WSF2019 for this analysis, to limit bias introduced by the WSF. Settlement which existed in 2019 but not, for example in 2016, would bring bias to the

---

[5] stacspec.org/en/

training data. By supporting the thesis of majorly growing human settlements in the years 2016 to 2022 (Taubenböck et al. 2019), the WSF2015 would underestimate settlement extents. However, this is more acceptable than an overestimation of urban settlement and thus possible misclassification.

The WSF depicts settlement on a building to block level, describing settlement as extents of built-up. However, the camp geometries, in this study, describe the settlement extents as contextual information derived from not only built-up areas but also infrastructure and other land cover types. Under the assumption of simplifying the feature learning process of the DL models, the WSF was processed to generate an extensive layer of settlement, by applying morphological closing operations (Figure 9). This transformation aimed to create a more contextual representation of the formal settlement.
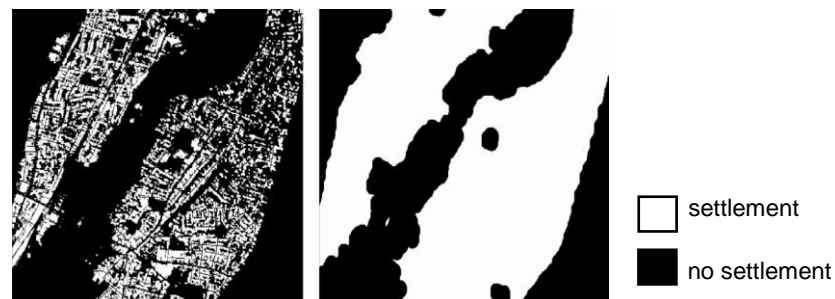
Figure 9: WSF2015 (left) and morphologically closed WSF2015 (right). WSF2015 by Marconcini et al. (2020).

3.2 Methods

The objective of this work is to apply DL models for mapping the extents of refugee settlement using Sentinel-2 data. Figure 10 gives an overview on the methodological application. The approach follows a standard process for training DL models. First, it is crucial to obtain accurate training data. In this study, a combination of open source and manually collected data was employed. The collected data then needs to be formatted in a way that allows the DL model to effectively learn from it. Finally, six different models are trained, validated, and analysed. The subsequent chapter will provide a detailed explanation of the underlying steps involved in this process.
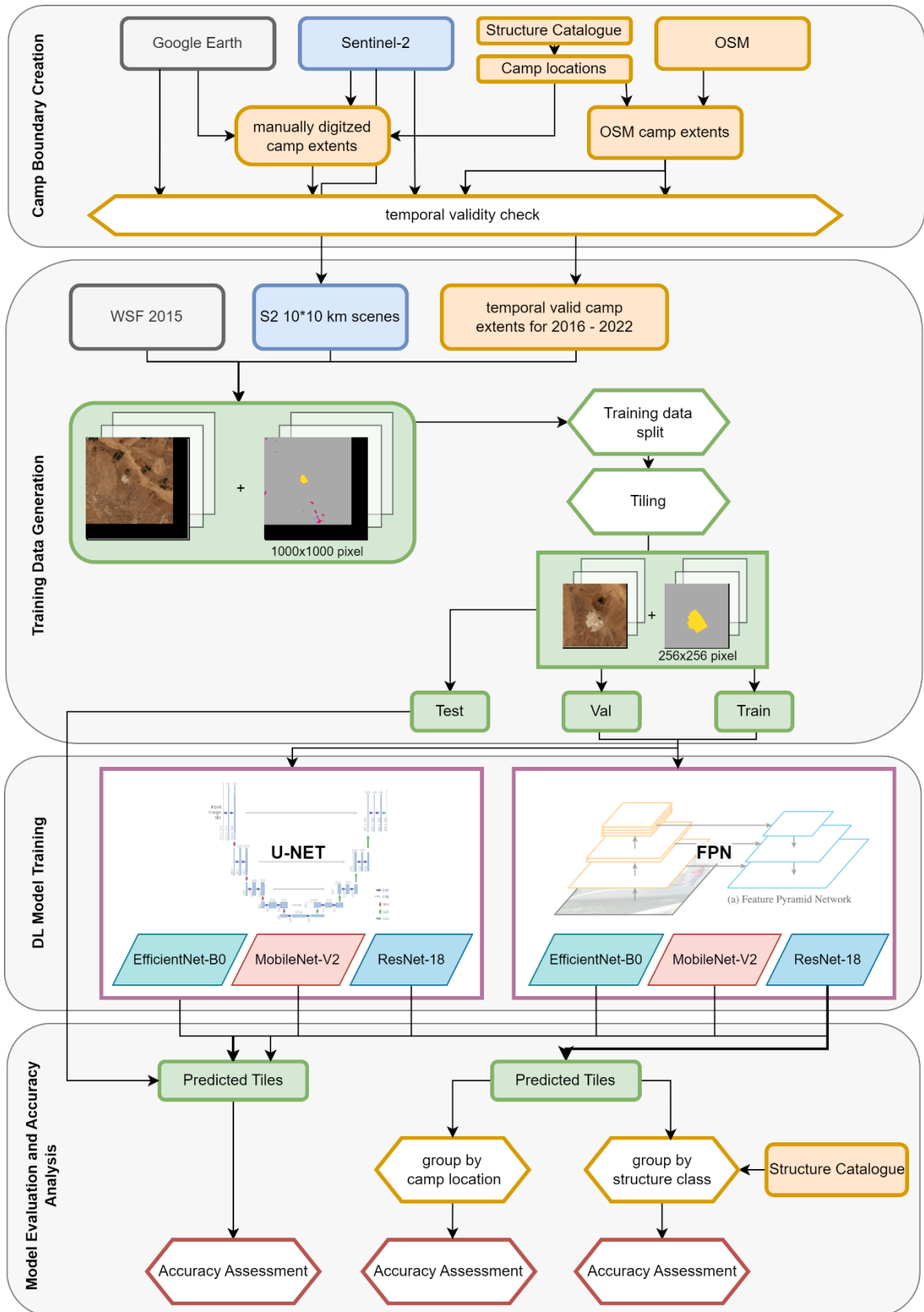
Figure 10: Workflow.

3.2.1 Training Data Preparation

Deriving the refugee camp settlement boundaries has already been explained in chapter 3.1.2, and the automated download for hundreds of Sentinel-2 scenes was described in section 3.1.3. The following section will explain the creation of a reference label corresponding to each Sentinel-2 scene by leveraging the camp boundaries and the modified WSF2015.

Image-Label Pairs

For creating the corresponding labels to the Sentinel-2 scenes, an automatized approach was used. First, four relevant semantic classes were determined: 'camp' (3), spatially covering the extents of refugee and IDP camp settlements; 'formal settlement' (2) describing all residual settlement structures; 'other' (1) land cover classes (e.g., water, vegetation) combined to one, and 'noData' (0) depicting all pixels with no value. See Table 1 for further information on the classes.

Table 1: Description of the classes.

| Class Nr.	Class Name	Description
0	NoData	No information included in Sentinel-2 scene; edge pixel, padded image edges
1	Other	All land cover types that are not associated with settlement in any kind. Vegetation or bare ground within settlements not included.
2	Formal settlement	Formal settlements which are not camp-like structures.
3	Refugee camp	Refugee and IDP camps, all housing types, delimited by natural borders (e.g., rivers) and anthropogenic borders (e.g., streets, fences) or visible delimitation in satellite imagery.

During label creation, one 2D label image was built for each downloaded Sentinel-2 scene. For this process, empty rasters were created, which were filled with pixel values step by step. First, the camp boundary geometries were utilized to label the refugee settlements, described by the class 'camp'. Corresponding to the time stamp of the Sentinel-2 scene, the temporal valid camp extent was rasterized. All other settlement structures occurring in the Sentinel-2 scene, which were not camp-like, were labelled by the modified WSF2015. The remaining pixels were labelled as class 'other' landcover classes, and the edge pixels to class 'noData'.

Some camps lie in spatial proximity to each other, resulting in more than one camp per image-label pair. In these cases, the temporal validity of all included camp geometries needed to be considered during the label creation.

Training data split

For model learning and the subsequent validation and testing, the created training data was distributed among three datasets. In this study, a 70/15/15 split was approached, where the training dataset holds 70 %, and validation and test datasets each 15 % of the data. The image-label pairs were assigned to the dataset according to their geographic location. Image-label pairs belonging to the same camp were all assigned to one dataset. Because of the spatial proximity of some camps, a careful split with spatially exclusive datasets needed to be established. This is to ensure the training, validation and test samples were completely independent. By ensuring the model's evaluation was based on unseen camps rather than reiterating those already encountered, but in different scenes, the evaluation process was able to quantify the DL model's ability to generalize and abstract information. The therefore required approach grouped camp locations by their proximity by surrounding them with a 5 km buffer and tested for intersection between the buffers. For those intersecting, the whole group was assigned to the same dataset. The non-intersecting camps were distributed among the three datasets aiming to achieve the 70/15/15 split. Finally, the training dataset held 9,015 scenes, the one for validation consisted of 1,689 and the testing dataset of 2,019 image-label pairs.

Tiling

At last, the image-label pairs were brought into a size format which can be handled by the network. It is common practice to train DL models on images with sizes ranging from 64x64 up to 512x512 pixels (Wurm et al. 2019; Wickert et al. 2021). Often, a simple resizing is applied, with a resulting loss in spatial resolution. In approaches of DL with remote sensing it is important to keep the spatial resolution to not lose information, as the pixels refer to an actual spatial extent on the ground. Thus, the data was split into smaller tiles, in order to maintain the high spatial resolution. As the Sentinel-2 scenes came with slight differences in their sizes, the image-label pairs first were padded to a size of 1024x1024 pixels. The padded pixels were assigned to class 'noData'. The padded scenes where then cropped into 256x256 pixels sized tiles, by applying a moving window approach with an overlap of 128 pixels. This resulted in 49 tiles per original scene, as no tiles overshooting the original boundaries of the scenes were included.

3.2.2 Deep Learning Implementation

The objective of this work is to assess the applicability of DL methods for mapping refugee settlement extents with Sentinel-2 data. Therefore, the two model architectures U-Net and FPN are trained with three different encoders and tested for their performance.

The DL framework made use of the popular ML libraries Pytorch (Paszke et al. 2019) and Pytorch Lightning[6]. Pytorch and Tensorflow are, "today's most popular and widely used [DL libraries] in research and practice" (Höser and Künzer 2020). The model and encoder architectures were accessed and used via the library Semantic Segmentation Models Pytroch (SMP)[7], which implements several state-of-the-art model and encoder architectures.

Six models were trained, three U-Nets and three FPNs, each with the encoders MobileNet Version 2, ResNet-18 and EfficientNet-B0. The performances of different model-encoder architectures were explored, assuming that the ability of mapping refugee settlement extents would differ across the models. Due to different architectures, the models and the encoders possessed different amounts of parameters for the feature learning (see Table 2), which potentially affected their respective abilities in mapping refugee settlement extents.

Table 2: Number of parameters by network configuration.

Encoder	Model architecture	Parameters
U-Net	EfficientNet-B0	6.3 Mio
	MobileNet-V2	6.6 Mio
	ResNet-18	14.3 Mio
FPN	EfficientNet-B0	5.8 Mio
	MobileNet-V2	4.2 Mio
	ResNet-18	13 Mio

Each encoder was initialized with ImageNet weights, since the model performance can benefit from the transfer learning (Gella et al. 2022). The models were trained with an NVIDIA RTX A5000 GPU with 24 GB VRAM and a CPU with 32 cores and 200 GB memory.

For model training, several pivotal hyperparameters on the model's performance were identified. It was outside of the scope of this study to find perfect values for all hyperparameters. Rather, the most impactful parameters were manually tested, in particular the batch size, the learning rate and the loss function. In addition, pretrained encoder

---

[6] pytorchlightning.ai
[7] github.com/qubvel/segmentation_models.pytorch ©2019, Pavel Iakubovskii

weights and regularizing parameters like dropout, weight decay and early stopping were implemented, next to regularizing learning rate schedulers such as a 'learning rate warm-up' and 'reduce learning rate on plateau'. The warm-up begins the training process with a smaller LR, which then increases after a set waiting time, here after 5 epochs. This can result in a plateau of the loss curves at the beginning. The 'reducer on plateau' decreases the learning rate when the validation loss starts to stagnate, to give the model the chance to learn even more. Both regularizing parameters help to automatically tune the learning rate. The models were trained with the Adam optimizer (Kingma and Ba 2014). The applied hyperparameters are listed in Table 3. Dropout could only be applied for the FPNs, as it was not reasonably implemented in the SMP library for the U-Net model.

Table 3: Hyperparameters used for model trainings.

Hyperparameter	U-Net	FPN
Batch_size	98	128
Dropout	not implemented	0.3
Early Stopping Min-Delta	0.01	0.01
Early Stopping Patience	25	10
Learning Rate	1e-6	1e-6
Loss function 1	CE_W	CE_W
Loss function 2	CE	CE
Pretrained weights	ImageNet	ImageNet
Weight Decay	0.1	0.5

The six models were trained with the large datasets consisting of 441,735 tiles for training and 82,761 tiles for validation. The prepared data, with a tile size of 256x256 pixels, were normalized by the Z-score normalization before feeding them to the models for training. Due to the already large dataset, which created an overhead and several bottlenecks, data augmentation was not applied. A strong class imbalance could be found in the datasets (see Figure 11), where class 'noData' and 'other' are overrepresented, whereas urban and refugee camp classes are of major interest for this study. Therefore, a data sampler was implemented while feeding the data to the model. It sampled tiles with the appearance of the classes camp and/or urban and dropped tiles which only held class 0 and 1. Additionally, the weighted cross entropy CE_W was chosen as loss function, which considers the class imbalance. Furthermore, the unweighted cross entropy loss CE was logged as a comparison to the very harsh metric of CE_W.
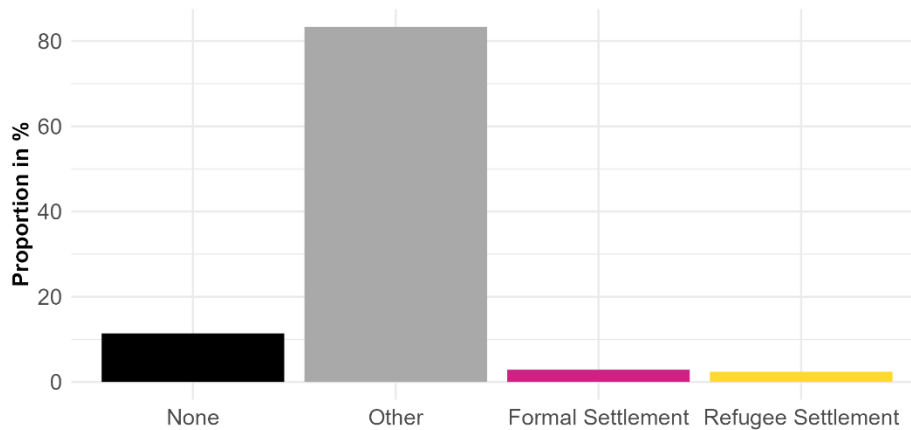
Figure 11: Distribution of pixels per class among whole training dataset. A strong class imbalance towards the class 'other' is apparent.

3.2.3 Model Evaluation and Accuracy Analysis

After all training processes finished, the model performance of the six models were assessed in a comparative analysis. First, the learning performance of all six models was analysed by examining the learning curves of the models. The course of the training curve shows how well a model learned the provided training data. Furthermore, the validation curve shows the models ability to generalize the learned features. The resulting graphs therefore showed training and validation loss, as well as the average IoU accuracy derived from predicting the validation dataset at the end of each epoch.

In the second step, for each trained model, the state of the model with the lowest loss got identified and used for the predictions. Then, the testing dataset, containing 98,931 image-label pairs, was predicted with the trained models. The predicted images were compared to the existing reference labels and served as basis for the accuracy assessment. The test dataset is independent from the training and validation datasets, therefore meaningful evaluations on the model's performance were made. From this, the confusion matrix as well as the accuracy metrics were calculated. The OA was derived as average for all classes, to give an overview on the model performances. The class-based metrics IoU, recall, precision and F1-sorce were calculated to provide an insight into the model performances across the classes. This is especially important in imbalanced datasets, as the majority classes can cover up the accuracies of smaller classes. The FPR was only calculated among the classes 'formal' and 'refugee settlement', to observe the models' ability of distinguishing between those two classes. Moreover, the resulting prediction maps were visually assessed by comparing the maps of the six models with each other and with the label. Based on the

visual and numeric assessment, the best performing model among the different architectures and encoders was identified and utilized for further analysis.

In the analysis part, the best model's ability to predict refugee settlement extents was evaluated for each camp location and by morphological structure (derived from Weigand et al. 2023). The predicted output maps were grouped by camp location and by structure type, for which all accuracy metrics were calculated separately. This allowed for detailed analysis of the segmentation performance in relation to the underlying structure types. The analysis by morphological structure required some preprocessing. The SC holds information at an observed point in time, therefore the temporal validity of the structure type had to be matched with the timestamp of the predicted Sentinel-2 scene. By assuming that the observed structural category was valid until a change was noticed, start and end timestamps for the structural types were created. The start time was the observed timestamp, and for the end date one day before the next observation was chosen. When the successive observation was missing, the 30th of September 2022 was picked since it is the last day of this work's study period. By comparing the prediction's timestamp with the structure type time periods, each prediction tile was assigned with their valid structure type.

4 Results

This chapter presents the results of the six DL models, which were trained in the scope of this work. The two model architectures FPN and U-Net were trained each with the three CNN encoders EfficientNet-B0, MobileNet-V2 and ResNet-18. Their performances during model training, as well as their accuracy results for predicting on the test dataset are displayed in the following chapter. Lastly, the results of the analysis by camp location and by morphological structure are presented.

4.1 Model Training

As explained in chapter 2.2.1, the primary goal of DL models is to minimize their loss function during training. The training and validation loss are monitored while training over time. When plotted as learning curves, the model performance can be evaluated and problems with learning can be diagnosed. In the following paragraphs, the results of the model trainings are presented.

4.1.1 U-Nets

The plots in Figure 12 show the learning of the models over the training epochs. Because of applied parameters like early stopping and learning rate regulators, the models stopped learning after different numbers of epochs. The early stopping set in when no more improvement above the set threshold of 0.01 was detected.

The MobileNet stopped earliest after 32 epochs, and the EfficientNet learned the longest for 79 epochs. Still, the ResNet model, which learned for over 47 epochs, achieved the lowest CE_W with the validation dataset and the highest IoU accuracy of 0.43 (f).
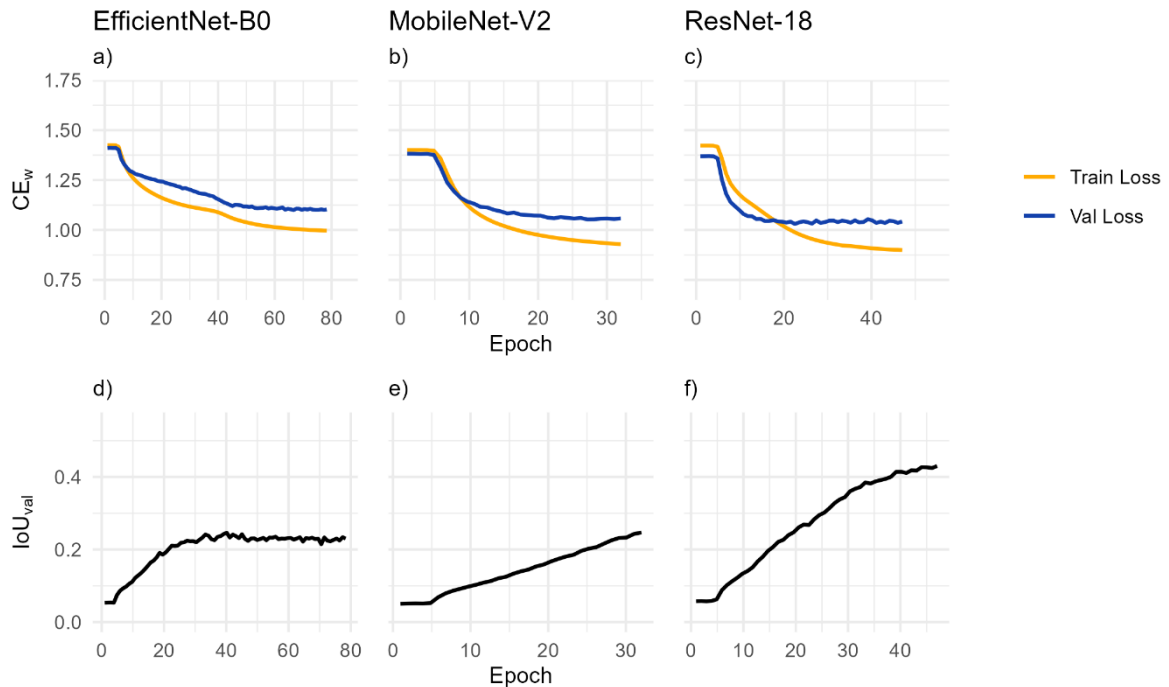
Figure 12: Learning Curves of U-Nets with CE_W and validation IoU accuracy.

It is striking that the validation loss curves of all three models reached a point of stagnation, while the training loss continued to decrease. With this behaviour, the graphs do not conform with a good fitting curve (Goodfellow et al. 2016, 423f.). The initial plateaus can be associated with the regularizing parameter of learning rate warm-up.

The improvement of the EfficientNet training curve (a) gradually slowed down until, after 41 epochs, it reached a point where the slope of the curve increased once again. The validation loss meanwhile continued to sink until the gradient suddenly decreased significantly. It is interesting to note, that the IoU (d) stabilized after approximately 40 epochs. During the stabilization of the ResNet validation loss (c), a slight fluctuation can be observed, albeit within a small range. Like the other two models, the loss curves of MobileNet (b) decreased until the validation loss started to stabilize. The gradient of the training loss also decreased while the model gained experience, although the curve continued to sink. In contrast, the IoU graph of the MobileNet model (e) consistently rose with no sign of stabilization. This indicates that there still was potential in model learning, but the training was aborted too early.

4.1.2 FPNs

Figure 13 shows the learning progress of the FPN models over time. The ResNet model trained the longest for 62 epochs, while the EfficientNet and MobileNet were trained for approximately the same duration, ending at 47 and 46 epochs, respectively. The ResNet model achieved the lowest validation loss (c) and the highest IoU accuracy of 0.51 (f), making it the most successful among the three models.
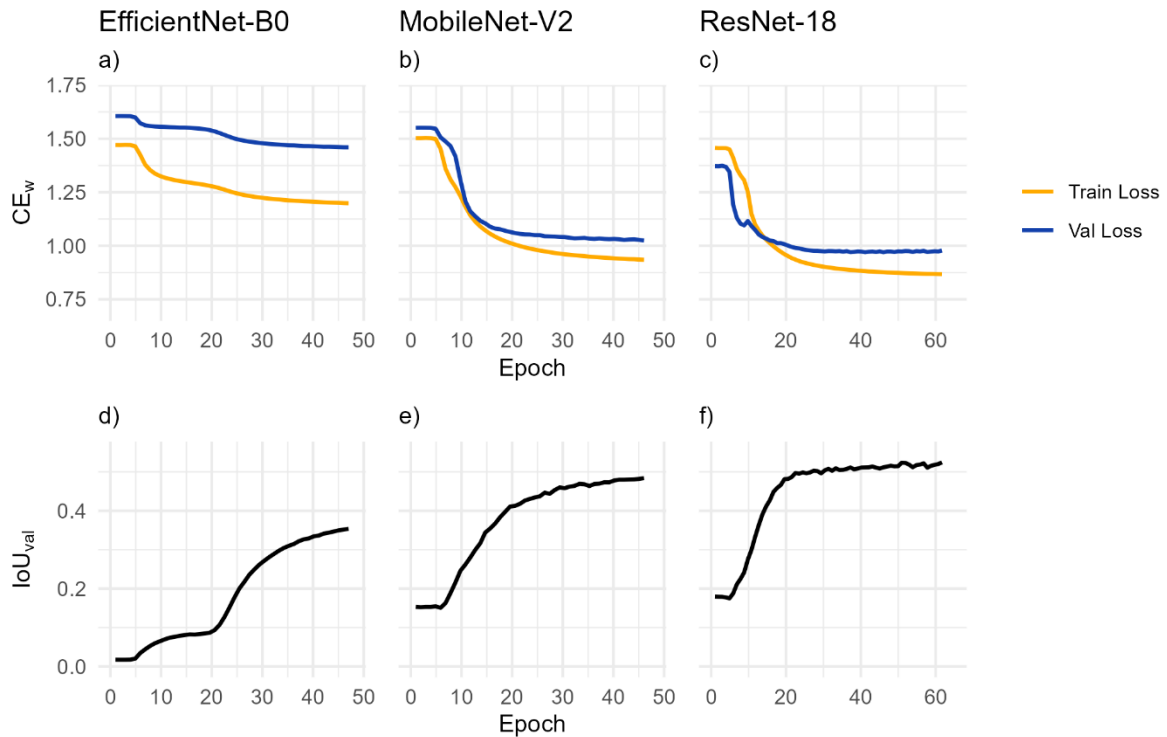
Figure 13: Learning Curves of FPNs with CE_W and validation IoU accuracy.

The behaviour of the EfficientNet graphs stands out. While the training and validation curves decreased (a), their decrease was not as significant as compared to the other models. The training and the validation curve reached plateaus twice, indicating loss stabilization after 14 and 26 epochs. After the 14th epoch, the loss gradient started to increase again, and after the 26th epochs the graphs began to stabilize. Interestingly, the IoU curve exhibits moderate growth until the 20th epoch, where it experienced a substantial increase followed by an incipient stagnation (d).

In contrast, the MobileNet and ResNet graphs show decreasing training and validation curves, with notable stabilization. The training loss of MobileNet (b) stabilized after 15 epochs, while the training loss improvements of ResNet (c) gradually slowed down after more than 20 epochs. The behaviour of the ResNet validation loss curve is particularly

striking. After only 8 epochs, the validation loss peaked for one epoch before resuming to decline, and it stabilized after the 25th epoch.

4.2 Validation/Testing

Once the six models completed their learning, the test dataset was predicted to validate the models' performances. To assess the models' abilities to apply the learned features on previously unseen data, a confusion matrix and the in 2.2.6 introduced state-of-the-art accuracy metrics were computed. The accuracy metrics for all six models are listed in Table 4.

4.2.1 U-Nets

The overall accuracies scored by the U-Net models differed significantly. The highest accuracy of 61 % was reached by the model with the ResNet-18 encoder, while the MobileNet encoder scored 44 %. The lowest overall accuracy was achieved by the EfficientNet encoder with 32 % of correctly classified pixels.

It is striking, that the accuracy values for the 'refugee camp' class were low for all three models, with the highest IoU value of just 0.12 and the highest F1-score of 0.221, both were achieved by the ResNet model. Comparing the confusion matrices of all three models (see Appendix 3, Appendix 4, Appendix 5), an overclassification of the 'refugee settlement' class among all models can be observed. The low precision values among all models indicate that more pixels were labelled as 'camp' in the prediction than are true positive. All three models misclassified many pixels labelled 'other' as 'camp'. Nonetheless, the models succeeded to predict a majority of actual 'camp' pixels correctly, as evidenced by high recall values ranging between 0.69 for MobileNet, 0.72 for ResNet and 0.75 for EfficientNet.

Similar patterns can be observed for the 'formal settlement' class. The predictions tended to overclassify the occurrence of formal settlements, although the degree of overclassification was less pronounced compared to the 'camp' class. Similar to the 'camp' class, there was confusion between the 'other' class and 'formal settlement'. Among the three models, the MobileNet-model performed best for 'formal settlement', with an IoU of 0.12 and a F1-score of 0.207.

The higher FPR values for the 'refugee camp' class show that among the two settlement classes, the models performed better for formal settlements than for refugee camps. The ResNet performed best for 'formal settlement', while at the same time reached a very high

FPR of 0.9 for 'refugee settlement'. For this class, the MobileNet encoder was performing best with the lowest value of 0.4.

The class 'other' achieved remarkably high precision values of over 0.99 in all models. As previously mentioned, the true labels were highly misclassified, resulting in lower to moderately high recall values. The highest recall of 0.56 for an U-Net model was achieved by the ResNet model. Notably, the ResNet model misclassified the fewest 'other' pixels as refugee settlement with only 13 %. The model also attained the highest IoU and F1 values for this class, which were notably higher compared to the previously mentioned classes.

The class 'noData' was predicted best by the EfficientNet model, achieving an IoU of 0.65 and a F1 score of 0.75. It also experienced overclassification, mostly at the cost of class 'other'.

Table 4: Accuracy assessment results for U-Nets and FPNs. *FPR only among classes 2 and 3.

	U-Net			FPN		
	EfficientNet	**MobileNet**	**ResNet**	**EfficientNet**	**MobileNet**	**ResNet**
OA	0.32	0.44	0.61	0.71	0.78	0.83
IoU						
Class 0	0.649	0.364	0.393	0.743	0.867	0.895
Class 1	0.212	0.354	0.560	0.693	0.755	0.823
Class 2	0.061	0.115	0.070	0.055	0.124	0.158
Class 3	0.040	0.057	0.124	7e-7	0.110	0.137
Precision						
Class 0	0.656	0.365	0.394	0.744	0.871	0.900
Class 1	0.997	0.995	0.995	0.993	0.995	0.994
Class 2	0.064	0.126	0.082	0.056	0.149	0.184
Class 3	0.040	0.058	0.131	0.014	0.114	0.146
Recall						
Class 0	0.984	0.996	0.996	0.998	0.995	0.994
Class 1	0.212	0.355	0.561	0.696	0.757	0.828
Class 2	0.586	0.573	0.501	0.942	0.431	0.528
Class 3	0.746	0.695	0.724	7e-7	0.760	0.681
FPR*						
Class 2	0.247	0.271	0.188	0.999	0.191	0.256
Class 3	0.405	0.398	0.904	(7e-7)	0.543	0.413
F1						
Class 0	0.787	0.534	0.564	0.852	0.929	0.995
Class 1	0.350	0.523	0.718	0.819	0.860	0.903
Class 2	0.115	0.207	0.141	0.105	0.221	0.273
Class 3	0.076	0.107	0.221	1e-7	0.198	0.241
Color Coding:	*worst value*	*best value in architecture*		*overall best value*		
Class 0: NoData	*Class1: Other*	*Class 2: Formal settlement*		*Class 3: refugee settlement*		

4.2.2 FPNs

The FPN models reached high accuracy values, with the ResNet model achieving again the highest value of 83 %, MobileNet with 77.8 % and EfficientNet with 71.3 %. However, the accuracy values for the 'camp' class were relatively low, with maximum IoU of 0.14 and F1-score of 0.24 achieved by the ResNet model. Overclassification can be observed among the 'refugee camp' class, primarily due to the confusion of 'other' pixels being classified as refugee settlements. The low precision and high recall values highlight the overclassification, while also indicating that a significant proportion of pixels labelled as 'refugee camp' was classified correctly. A striking observation from the error matrices of the three FPN models is the apparent inability of the EfficientNet model to predict refugee settlements (see Appendix 6, Appendix 7 and Appendix 8). The IoU, recall, precision and F1-score values for the 'camp' class were notably low, as they were approaching zero. Furthermore, the proportion of pixels classified as 'camp' was almost zero.

The ResNet model achieved the highest F1 and IoU accuracies for the 'formal settlement' class, indicating a better performance compared to the other models. However, it is noteworthy, that the values, similar to the 'camp' class, were still relatively low. These low values already indicate the presence of overclassification in all three models. The EfficientNet here notably misclassified over 24 % of the 'formal settlement' pixels. Low precision values among all models further highlight the overclassification, while the high recall values indicate a high amount of correctly classified pixels. The FPR among the two settlement classes showed lower values for 'formal settlement' than for 'refugee camp'. A special case is the EfficientNet encoder, where the 'formal' class was strongly overclassifying 'refugee camps', which showed in an FPR of 0.99 for 'formal settlement' and almost 0 for 'camp'. Besides that, the best values for 'formal settlement' were achieved by MobileNet and for 'refugee settlement' by ResNet.

The class 'other' experienced underclassification, as it was prone to confusion with the other classes, particularly with the 'refugee camp' class. Despite this, all models achieved very high precision values of over 0.99, pointing out that almost no non-'other' pixels were misclassified as class 'other'. The recall values, although slightly lower, are ranging between 0.7 for EfficientNet and 0.83 for ResNet. As a result, the 'other' class reached high accuracy values, with the highest F1-score of 0.9 and the highest IoU of 0.82 achieved by the ResNet model.

The highest accuracy values for class 'noData' were reached by the ResNet model, with an IoU of 0.9. High precision and high recall values resulted in high F1-scores, the highest value of 0.95 scored by the ResNet model.

4.3 Comparison of all model performances

The visual results of all six models for ten sites of different structural types are presented in Figure 14 and Figure 15. For each predicted model they show predicted tiles of the test dataset. Single looks of the approximate size 2.5x2.5 km^2 were chosen. This chapter aims to combine the presented numerical results with their visual representation. To present them for different kinds of refugee camps, examples for each structural type were chosen.

Comparing the performances of the trained U-Nets and FPNs, the FPN models achieved significantly higher overall accuracies: The highest OA of 83 % was reached by FPN ResNet compared to 61 % for the U-Net ResNet. The FPNs achieved higher numerical accuracy metrics, and the visual results showed more accurate predictions of refugee settlement extents.

The U-Net models didn't show expected learning behaviour: The visual predictions contained much noise, because of relatively small segments and high misclassification of the classes 'noData', 'formal settlement', and 'refugee camp'. All three U-Net learning curves showed underfitting behaviour, which indicates that the model has not been able to learn the training data sufficiently.

The FPN models, on the other hand, showed good fitting learning behaviours in their loss curves. Visually, they showed the best performance. In most cases, the presence of a camp got detected, as for lines 4, 5 and 8 (in Figure 14 and Figure 15). Small camps, which are small enough to be completely shown in one tile and are detached from any urban area, were detected by the MobileNets and ResNets in both architectures. Still, the area git overestimated for every camp (see lines 4, 5, 8). Line 7b shows a refugee camp enclosed in urban structure. All models failed to detect this, and rather mapped camps in other locations.

The FPN with EfficientNet encoder did not perform as expected, in contrast, it failed to map refugee settlement. The model loss improvements for both training and validation were small, suggesting the models did not learn sufficiently. The accuracy metrics and images show that the model was not able to learn features related to refugee camps and confused them with 'formal settlement'. Nevertheless, the visual assessment showed that built-up structures were detected well by the model. As in lines 0, 6 and 7b, the model captured the extents of built-up well, but failed to distinguish between formal and informal built-up.

The visual assessment showed high overclassification of the 'camp' class of all models, except FPN EfficientNet, resulting in low accuracy and precision values. Thereby, the confusion between 'camp' and 'formal settlement' was maximum high for the U-Net

MobileNet model, with an FPR of 0.904 for the 'camp' class. Among the FPNs, the highest FPR was 0.543, indicating some incorrect classification but mostly successful distinction between 'camp' and 'formal settlements'. Nonetheless, the correct localisation of refugee settlement can be observed, which is evidenced by high recall values over 0.68.

Classification errors between the 'other' and the 'noData' class were observed frequently in the visual results. For example, water in lines 1 and 4 was misclassified as 'noData', resulting in accuracy errors.

Figure 14: Visual results of the six trained DL models, part1. Showing comparative result examples for refugee camps with different structural morphologies.
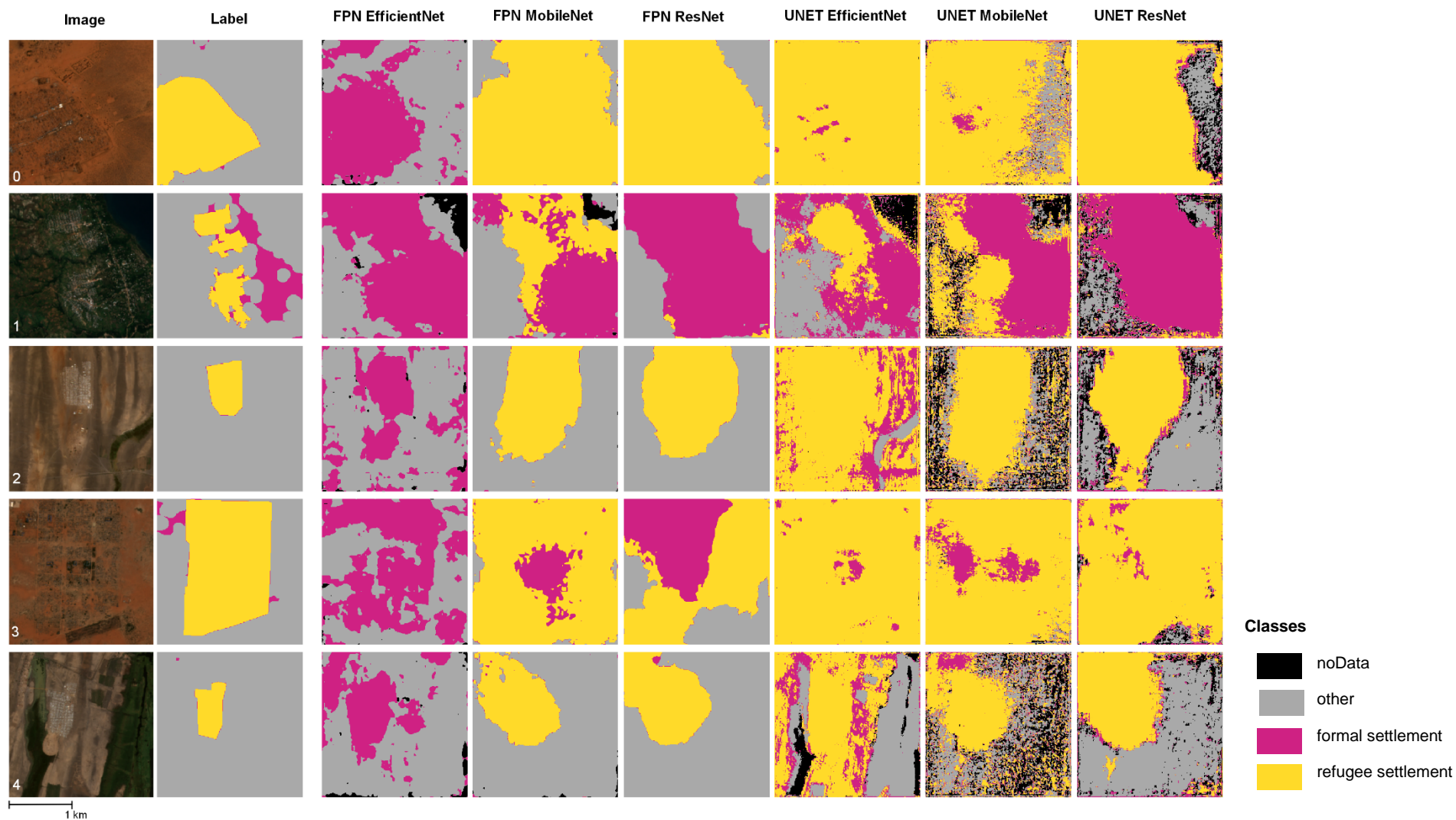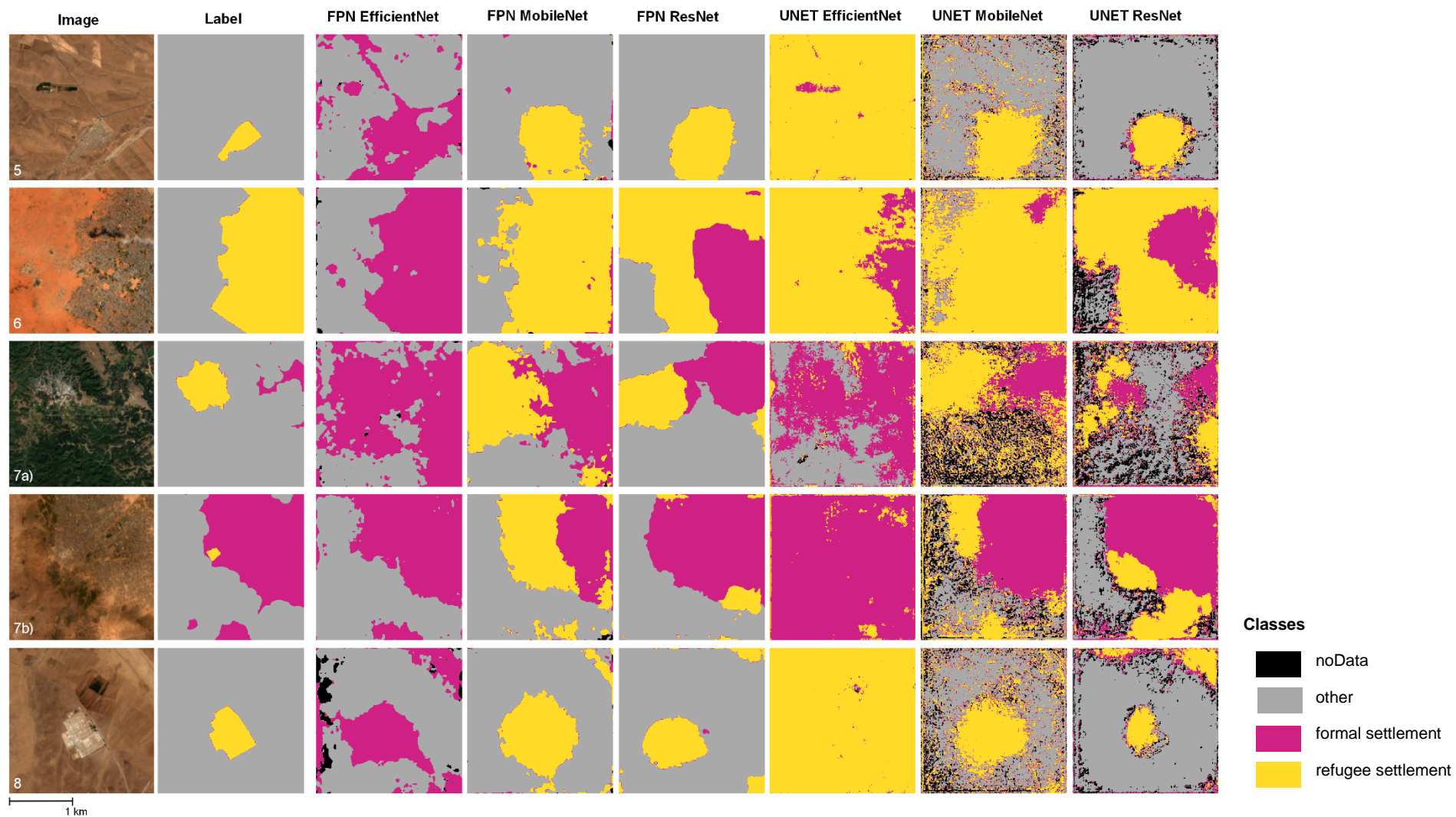
Figure 15: Visual results of the six trained DL models, part2. Showing comparative result examples for refugee camps with different structural morphologies.

4.4 Analysis by camp location and structure type

The FPN model with ResNet-18 encoder showed the highest accuracies and in total the best performance in comparison to the other models. Therefore, the model was chosen for the following analysis. Firstly, the models' ability to map refugee settlement extents was assessed per refugee camp location existing in the test dataset. This is to unveil potential differences in detection capabilities across different geographical regions or depending on the physical-geographic setting. Secondly, the model performance was analysed among the structure types identified by Weigand et al. (2023).

4.4.1 Model performance per refugee camp location

Figure 16 shows the spatial distribution of the camps contained in the test dataset. The colors present the accuracy metric values achieved by assessing the performance for all tiles associated with the camp.

It is striking that only two camps hold the highest values in most categories: Mamrashan in Iraq and Smara in Algeria. Two predicted examples of these camps are included in Figure 16. Mamrashan achieved the highest accuracies in the categories mean IoU of 0.56, mean precision of 0.59 and mean F1-score of 0.64. Smara scored best for the 'refugee camp' class' IoU of 0.27, precision of 0.27 and F1-score of 0.43.

When examining the global maps, no striking large-scale regional pattern became apparent. Upon examination of areas with clusters of refugee camps, certain patterns emerged. Case area 1 shows camps located in Sudan, in proximity to the South Sudanese border. The camps in this region exhibited relatively low IoU and precision values, but high recall accuracies for the 'refugee settlement' class. The camp with the highest global recall value for 'refugee settlement' was Jourie camp and was located in this case region (see Figure 14, 1).

The second case area enclosed three camps in Turkey, close to the Syrian border. The accuracy results for 'refugee camps' IoU and precison were notably heterogeneous. For the camp Saricam reatively high IoU of 0.18, precision 0.21 and recall 0.56 values were achieved. In contrast to that, the camp Islahiye was predicted with relatively low accuracies with IoU and precision of 0.04, but a recall of 0.87.

The third case area shows four camps located in Bangladesh, close to the border to Myanmar. The accuracies per camp were relatively homogeneous, all achieving low values. However, the accuracies of the most southern camp, Camp 22, fall out of the pattern as higher IoU and recall values were reached (compare Figure 15, 7a).
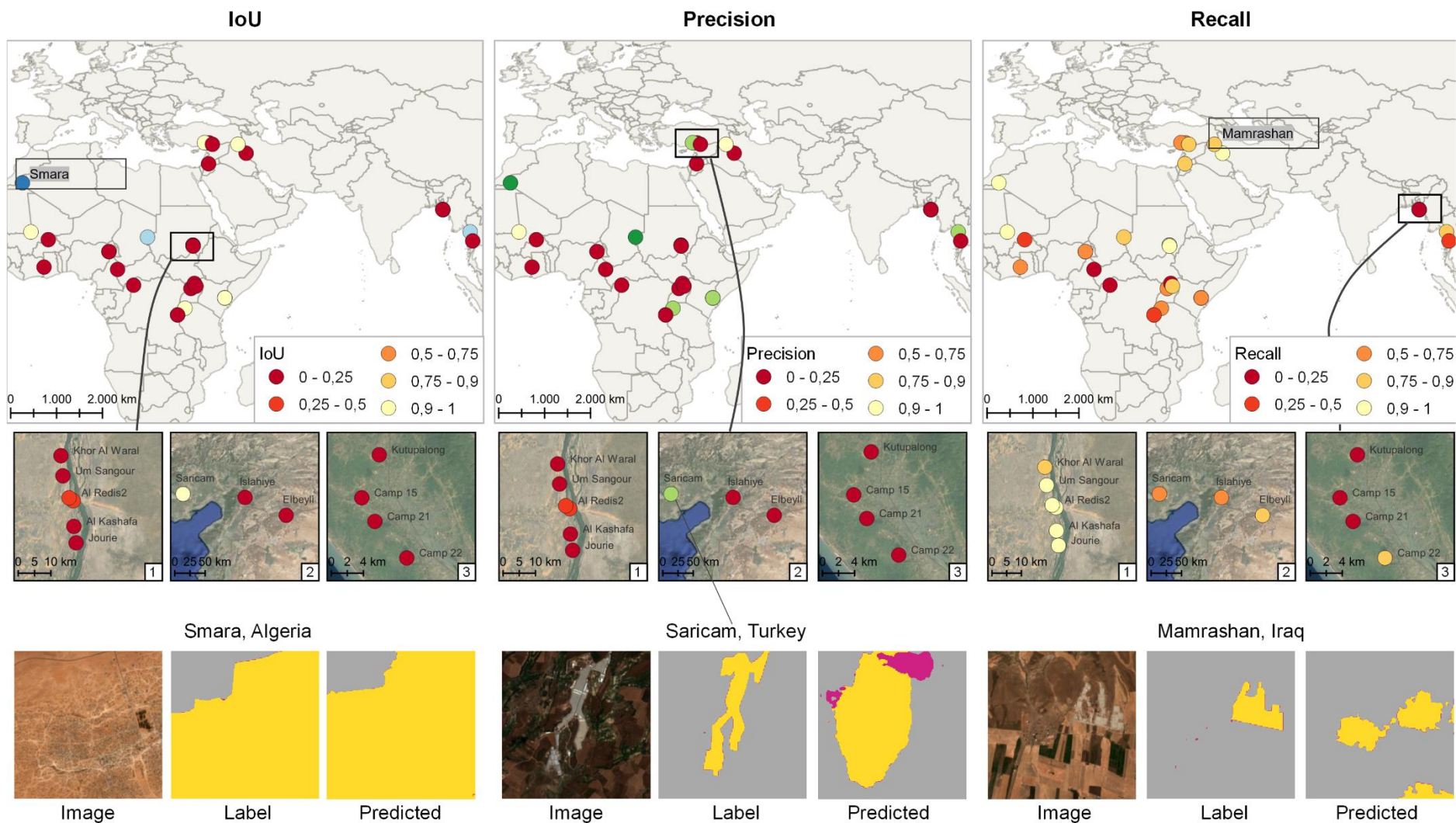
Figure 16: Accuracy metrics by camp location, with examples. Showing the accuracy metrics IoU, precision and recall for the 'refugee camp class'.

Basemap: Google Earth

4.4.2 Model performance for structure types

In the next analysis step, the predicted tiles were grouped by their morphological structure types, while considering their structural changes over time. For the analysis, the outputs of the prediction with FPN ResNet-18 were chosen. It exhibited the highest accuracies, as discussed in the previous chapter.
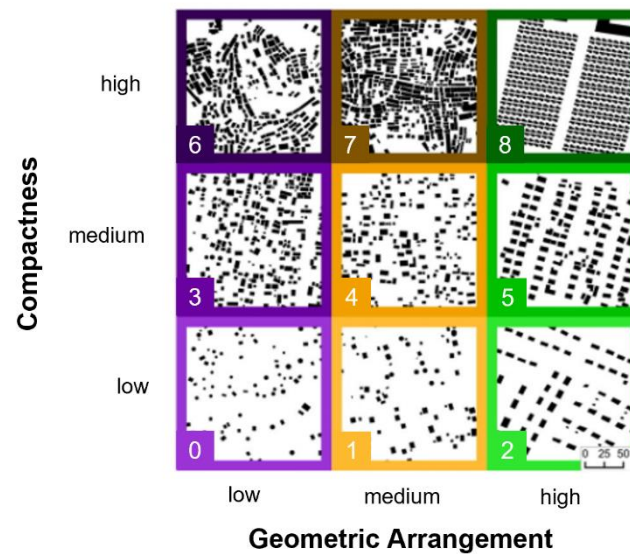
Figure 17: Structure types by Weigand et al. (2023), with my numbering of classes. Reprinted and adapted from Weigand et al. (2023).

Figure 18 shows the distribution of OA values achieved for each structure type. They varied from minimum 71 % for type 6 (high compactness, low arrangement, see Figure 17) to the highest of 91 % in type 5 (medium compactness, high arrangement). Even more variation between the type accuracies can be seen in the class-based accuracy metrics per structure types, which are illustrated for the 'refugee settlement' class (Figure 18). It is noteworthy, that the accuracy metric values were rather low among all types. Only the recall metric achieved high values, with the highest of 0.97 in type 5. Among the structure types, the highest values for IoU of 0.2 and F1-score of 0.33 were achieved for type 4. The category represents refugee camps characterized by medium compactness and medium geometric arrangement. Type 3, with medium compactness and low geometric arrangement, achieved the highest precision value of 0.2 and performed well, although it had a low recall value.

The lowest accuracy values were attained by type 1, which represents camps with low compactness and medium arrangement. Similarly, type 8, describing highly compact and geometrically arranged camps, did not achieve high accuracy values either.

Figure 18: Accuracy metrics per structure type. IoU, precision, recall and F1-score show the accuracy for the 'refugee camp' class.

5 Discussion

This work introduced an approach for mapping refugee settlement extents from HR satellite data. The study employed six different DL models to analyse 275 camps worldwide. The findings revealed an overall success of localising refugee camps, but with a tendency to overestimate their extents. The six DL models performed 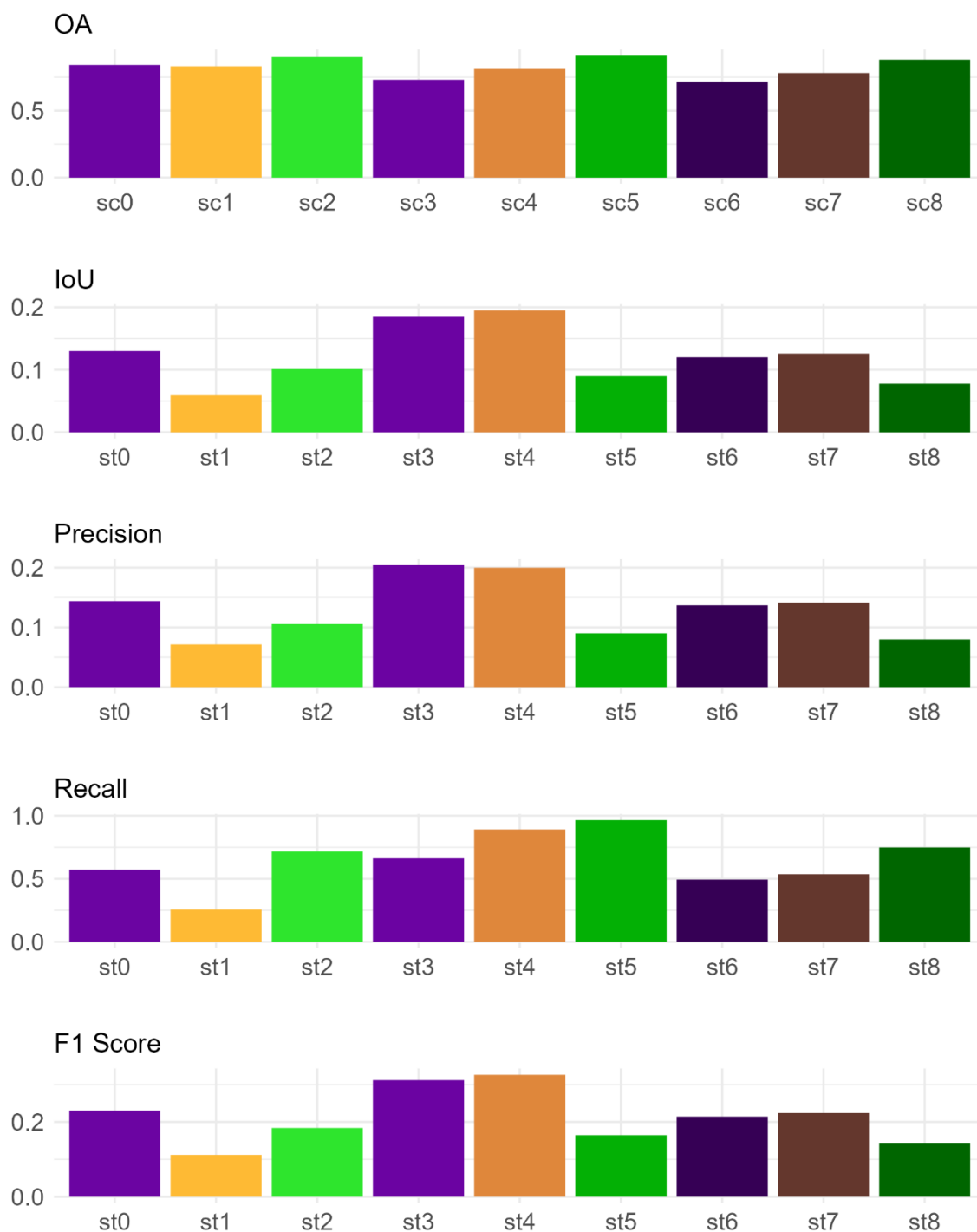very differently on the task, with the FPN models achieving higher accuracies in total. Among all models, the FPN with ResNet-18 encoder was identified as the most suitable for mapping refugee settlement extents. Regional patterns in the model's ability to map refugee settlement extents were found, however no large-scale pattern became apparent. In the following chapter, the results of applying the six DL models are discussed in detail, while addressing the presented research questions.

1) Is it possible to map refugee camps from satellite imagery across the globe?

Despite medium to high OA results for all six models, the class-based accuracies for the 'refugee' and 'formal settlement' classes were comparably low. This mismatch between OA and class-based accuracies for camps can be explained by the stark class imbalance of the task. The maximum IoU achieved for the 'refugee camp' class is 0.14, and for 'formal settlement' a similar value of 0.16. Striking is the strong overclassification of the two settlement classes evidenced by low precision values and overestimated visual results. There are several parameters which pose a challenge to the mapping of refugee camps from satellite imagery. On the one hand, refugee camps across the globe have very different and complex morphological appearances, which makes it difficult for the models to find a generalized way of mapping them. As Weigand et al. (2023) showed with their Structure Catalogue, the morphological structures of refugee camps vary in their density, building homogeneity, shelter areas and other variables. On the other hand, refugee camps can spectrally be very similar to their surroundings. The models trained in this work strongly confused 'other' land cover types with the settlement classes. While building materials in refugee camps like tarp, wood and thatch spectrally differ from formal building materials, they can be difficult to distinguish from the immediate surroundings (UNHCR 2016; Quinn et al. 2018; van den Hoek and Friedrich 2021). Distinction becomes especially challenging in 10 m resolution satellite imagery, as small structures and dwellings are not detectable (Wendt et al. 2017). Since the FPRs among the settlement classes were medium to low, a promising distinction between formal and refugee settlement can be evidenced. Lastly, the existence of an official or visible demarcation of the camp does not necessarily describe the location of dwellings within the camp (van den Hoek and Friedrich 2021). Rather, in addition to shelters and buildings, the camp extents do enclose other land cover types like bare ground, vegetation, and streets, which also exist outside the camps. Hence, the models needed to learn contextual information by extracting "spatial

features capable of capturing long-range pixel dependency in the image" (Persello and Stein 2017, p. 2325) in order to accurately map settlement extents. Usually, this makes CNNs especially suited for the task over pixel-based classification methods (e.g. Kuffer et al. 2016; Wurm et al. 2017), as CNNs are able to abstract contextual information without the need of explicit feature engineering (Höser and Künzer 2020). However, the models seemed to struggle localising the extents, suggesting that the models failed to learn contextual information.

Lastly, it is crucial to discuss the importance of some numerical metrics in regard to the obtained results. The study revealed relatively low precision values for the settlement classes, indicating strong overclassification. However, the recall reached high accuracies for all models expect U-Net EfficientNet, showing that a substantial number of pixels labelled as settlement were correctly predicted. Quinn et al.'s (2018) emphasized the importance of high accuracies and precision, as false detection of dwellings can result in more work than missing detection. However, it is essential to recognize the informative value of high recall values since the evaluation and prioritisation of the recall and precision metric depends on the desired end product. In this work, the recall values, in conjunction with visual assessment, demonstrate that the models did not randomly classify arbitrary areas as refugee camps but accurately detected their location and correctly identified camp structures. Although the precision may not be high, the high recall indicates a success of detecting refugee camps to some degree.

2) Which model performed best?

The DL models trained in this study performed very differently in their training performance and in their accuracies. The FPN models achieved the highest values in 16 out of 19 accuracy categories. They also performed better in the visual results, stating that the FPN models perform superior for mapping refugee settlement extents than the U-Net models (see Table 4). The findings of Gao et al. (2022c) stand in contrast to that, as their study for dwelling detection with different model and encoder architectures did not find any significant differences in performance, albeit a different task. Furthermore, as all three U-Net models trained in this thesis show underfitting, the superior accuracies for FPNs can be explained with the model training performances. The learning curves of the U-Nets show that the models had difficulties learning the training data sufficiently (see Goodfellow et al. 2016, p. 423). The analysis was conducted under the assumption that employing more regularization would improve the learning process. Due to the time constrains of the study, only the U-Nets were trained with improved regularization parameters compared to the FPNs, like reducing the LR when hitting a loss plateau or increasing the patience of the early stopping function. In contrast, regularization can prevent the model from learning, as the regularizing parameters are mostly applied for preventing overfitting. For example, a weight decay value set too high can trap the

network in a bad local minimum (Goodfellow et al. 2016, p. 424). This suggests that the U-Nets could possibly have achieved higher accuracies when trained with less regularizing parameters, but further testing would be necessary to verify this.

Another approach to address underfitting is by increasing the complexity of the model, as deeper models have the capacity to learn more complex features (Goodfellow et al. 2016, p. 110). In the context of the study, the ResNet encoder, which features the highest number of parameters, achieved the best accuracy values among all six evaluated models (Table 2). This suggests that the increased complexity of the encoder could have contributed to the model's abilities to abstract and generalize the learning task (Goodfellow et al. 2016, p. 21; Mboga et al. 2017). In support of this idea, Ghorbanzadeh et al. (2022) achieved significantly better accuracies for dwelling detection with a U-Net with residual connections over a standard U-Net. The residual connections allow the model to increase in complexity while preventing scaling problems such as a vanishing gradient (Höser and Künzer 2020). Since the U-Net models evaluated in this work possessed even more parameters than the equivalent FPN architecture with the same encoder, three findings can be identified. First, one must see the possibility that the U-Net architecture is just not suitable for mapping refugee camp settlement extents from Sentinel-2 data with the applied set of hyperparameters. The visual results argue against this, as some signs of predicted extents can be seen in the visual data (e.g., Figure 15, example 8). Second, the FPN architecture required less parameters to achieve even better accuracy results, suggesting the architecture can solve the task in a computationally more efficient way. And finally, the results show that the choice of encoder and architecture has a great influence on the segmentation accuracy and requires a special selection of a suitable architecture.

Regarding the FPN models, the difference in performance between the models with ResNet and MobileNet encoder and the FPN EfficientNet is significant. The latter shows an even more extreme underfitting behaviour than the U-Nets, as it almost completely failed to learn any features of refugee camps. The explanations for the low U-Net performances listed above cannot be applied on this model, as the model was trained with less regularization and possesses medium numbers of parameters in terms of complexity. Besides architectural differences between the encoders, a less extensive manual hyperparameter search can be assumed as reason for the bad performance. Due to the time constrains of this work, the hyperparameters for the EfficientNet trainings could not be tested as extensively as for the other two encoders, possibly leading to the observed underfitting behaviour.

3) Are there any patterns among regions or structure types?

In the analysis, the accuracy metrics distributed among camp locations and structure types were compared. Among the accuracy metrics by camp location, no global, large-scale pattern became apparent. However, some regional patterns could be observed, for example along the Bangladesh-Myanmar border, where the accuracy metrics performed relatively similar for all camps (see Figure 16, case area 3). As the SC shows, these camps exhibit the same trends in their structural morphologies, as they are highly compact settlements with low to medium arrangement. In comparison to VHR-imagery, it becomes apparent that similar building materials were used (see Figure 19). Roofing material in this region is often blue or red metal sheets, that provides a good contrast from the vegetation or ground, and compact building arrangement led to high homogeneity of the spectral signature. The same trend can be observed in Sudan, where six camps are aligned in proximity along the White Nile. The accuracy metrics for these camps are relatively similar, while their appearance in space also exhibits similarities in building structure and dwelling materials. However, the information in the SC is not supporting the findings in Sudan, as the camps were categorized into different classes, with medium to high compactness and all ranges of geometric arrangement. But as Weigand et al. (2023) stated, there is "no one-size-fits-all description of the complex and diverse morphological parameters", so even if the camps are categorized into different classes, they can still share similar morphologies.
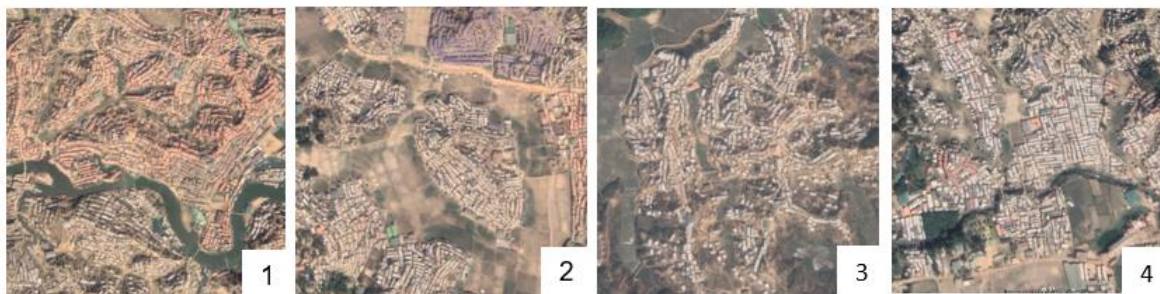
Figure 19: VHR imagery of four camps in Bangladesh. 1: Kutupalong Balukhali Camp; 2: Camp 15; 3: Camp 21; 4: Camp 22. Images: Google Earth ©2023 Airbus.

Weigand et al. (2023) found a cluster of camps with similar morphologies close to the Turkish-Syrian border. The camps show mostly highly compact and highly arranged structures. Despite their structural similarities, the model's ability to map the extents differed. The Saricam camp with a planned, strictly aligned layout reached relatively high accuracies (see Figure 16), whereas the Islahiye camp with similar structure did not perform as well. The SC shows that the Islahiye camp lost compactness over time, which is evidenced by gradual deconstruction and removal of some dwellings, as observable in historic VHR-imagery from Google Earth.

This finding suggests that the DL model had difficulties mapping the camp extents during times of deconstruction.

To conclude, the model's ability to map various camp locations differed, and some relationships to their structural morphologies could be found. As similar camp structures can occur in some regions, regional patterns in model performance were found. Still, different types of structures can be found in all regions. Therefore, no overarching statements describing supra-regional trends can be deduced from the findings of this study.
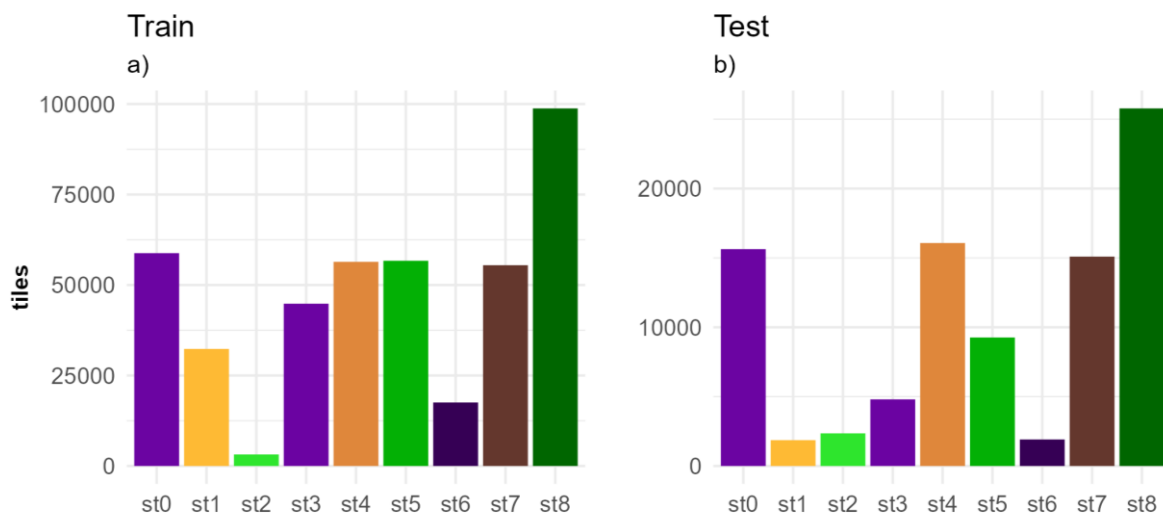
Figure 20: Histograms of tiles per structure types, shown in the training and testing dataset.

When comparing the model's performance among the different morphological structure types, the results were rather surprising. The categories with medium compactness and low to medium arrangement achieved the highest accuracy values, while the type with high compactness and high geometric arrangement performed rather bad. Furthermore, the type 0 with low compactness and low arrangement performed, in average, better than the type 8 with high compactness and high arrangement. To analyse these results, the distribution of data among the types needs to be considered. Figure 20 shows a strong imbalance in appearance of the structure types among the training (a) and test dataset (b), since spatial exclusivity was considered for the dataset split instead. Categories 0, 4, 5 and 7 are highly represented in the datasets, and type 8 is the most represented, among both datasets. When comparing the accuracy results for these five types, st4 still performed best and st0 second best. No clear patterns of improved performance of more compactness or more geometric arrangement can be found. Additionally, it is rather surprising, that even though type 8 is the major category, it performed under average in most categories. In contrast to this, common behaviour of model learning with class imbalanced datasets show bias towards the majority class as it was learned

the best (Johnson and Khoshgoftaar 2019). An attempt at explaining the low performance by their similar appearance to formal settlements becomes dismissed by the low misclassifications among the two settlement classes.

In summary, the results revealed that some morphological structure types of camps could be learned and predicted better than others. More detailed statements on the influence of structure on the results do not seem appropriate (at this point), however, as the training and test data showed there is a strong class imbalance.

4) Challenges & Future Research

It is common knowledge that the model performance depends on the quality of the training data, besides the model's ability to generalize the training data (Goodfellow et al. 2016, p. 414). As this works training data relied on, for the most part, manually digitized camp extents of 275 refugee camps worldwide, the manual visual image interpretation (MVII) did most certainly introduce bias. Like Kraff et al. (2020) showed in their study, MVII executed by only one interpreter can introduce uncertainties into the data. Furthermore, refugee camps are complex settlements and are difficult to map due to unclear boundaries or loose structure. Also, their appearance can be very similar to formal settlements nearby. Future research could analyse the effects of consulting various people to digitize borders to reduce the subjectivity and resulting uncertainties from the training data. Another approach could include the automatized creation of labels. Gao et al. (2022b) showed the potential of OBIA generated labels for refugee dwelling detection, as the models pretrained on the OBIA labels and finetuned with manual labels outperformed all other models.

Another factor challenging the model performances and limiting the quality of this work's training data is the selection of the WSF2015 as formal settlement layer. By preferring its sole usage above incorporating the version of 2019, it was accepted that the formal settlement was presented as static while the refugee settlement was dynamic. Furthermore, by creating a morphologically closed settlement layer, the formal settlement structure could have been overestimated for some regions. This could have affected the model's ability to learn the class formal settlement. The exact influence of the morphologically closed WSF could not be analysed in the scope of this work. In future research, model performances with the original WSF could be compared with the modified settlement layer. Additionally, an analysis including the dynamics of formal settlement could be realised by incorporating the WSF Evolution, which presents settlement expansion from 1985 until 2015 (Marconcini et al. 2021).

Another limiting factor is the class imbalance, introduced by the large satellite image scene. The size of 10x10 km^2 exceeded the refugee camps area sizes in most cases, resulting in an underrepresentation of the camp class in the scenes. This introduced a strong class imbalance

towards the class 'other', as many camps are surrounded by other land cover types which are not settlement (see Figure 11, Johnson and Khoshgoftaar 2019,). Further research could examine the effects of smaller Sentinel-2 scenes on the model's performance. For this work, a data sampling method as well as a weighted loss function was applied to soften the class imbalance effects. While testing, no data sampling was applied on the testing dataset. Consequently, the class imbalance was reintroduced here. It can be assumed that the models strongly overclassified the expansion of refugee settlements as the models did not learn the low abundance of settlement pixels in 10x10 km^2 scenes, due to the balancing methods. Another challenge introduced by the usage of Sentinel-2 scenes is their spatial resolution of 10 m. With a spatial resolution of 10 m, mixed pixels and the mapping small entities are a common problem (e.g., Wurm et al. 2017). This becomes apparent in the model's confusion of other land cover types with refugee camps, as the camps also exhibit similar land cover between buildings. Thus, the network needed to generalize on texture and contextual features, not only on spectral information. Higher resolution satellite imagery could help improving the segmentation results. Moreover, including near-infrared information into the training data could possibly improve performance, such as Gao et al. (2022a) showed.

A challenge often encountered during this work was the occurrence of memory bottlenecks due to the large amount of training data. Even though I had access to computationally strong hardware, the process of tuning the data loading and model training was very time-intensive and accompanied by many memory errors and bottlenecks. However, the high amounts of data did possibly increase the model's accuracies (Stark et al. 2020).

Due to the scope of this thesis and the available computational power, it was challenging to find the optimal hyperparameters for each one of the six DL models. Therefore, the U-Nets as well as the FPN-EfficientNet showed underfitting behaviour. With further optimization of hyperparameters, e.g., by structural hyperparameter testing using grid search, these model trainings could be improved. Especially the FPN-EfficientNet shows promising behaviour. Even though it almost completely failed to learn features for distinguishing camps and formal settlement, the visual results showed promising capabilities of delineating the boundary and the shape of settlements (see Figure 14 and Figure 15, examples 0, 6, 7b). This shows that the FPN Efficient-Net has potential for further analysis.

The findings examined in this thesis can act as a steppingstone for refugee camp research with remote sensing. They can function as a starting point for tackling the presented knowledge gaps in research, like the strong underrepresentation of refugee camps in satellite-based settlement products. Furthermore, to my knowledge, no study assessed the ability of mapping refugee camps in a large-scale approach yet. Therefore, the finding of this work can strongly contribute to the current research. The findings show that there does not exist a one-fits-all

model architecture or encoder tackling the complex problem of mapping refugee settlement extents. Future research can focus on finding fitting approaches for different structure types, in order to receive the best possible mapping results by morphological structure type. As this work identified the FPN architecture to be better suited for mapping refugee camps, this architecture could be examined further. Since the ResNet-18 performed best as encoder with the most parameters, future analysis should focus on applying deeper models to further investigate the influence here. Additionally, a combined approach of architectures can be proposed. While the FPN ResNet successfully localised refugee camps and distinguished them from formal settlement, the FPN EfficientNet was able to map settlement extents accurately. By combining those architectures or deploying them successively, it could lead to accurately mapped refugee settlement extents. Despite the constrains of Sentinel-2 data, we were able to map refugee settlement extents with only freely available data. This could be extended by additionally incorporating Volunteered Geographic Information (VGI), for example OSM building footprints.

6 Conclusion and Outlook

Refugee camp settlements are highly underrepresented in current satellite-based settlement products (van den Hoek and Friedrich 2021). Moreover, the research is lacking approaches for large-scale mapping of refugee settlement extents. Therefore, this work tested the applicability of six DL models for mapping refugee settlement extents for 275 camps worldwide with Sentinel-2 satellite imagery. The two DL encoder-decoder architectures U-Net and FPN were trained with three different encoders EfficientNet-B0, MobileNet-V2 and ResNet-18. The training data consisted of over 620,000 Sentinel-2 satellite images and their corresponding labels, with camp extents obtained from OSM and manual digitization, and the WSF2015 as the representative layer for formal settlement.

The results showed that all six models had difficulties mapping the settlement extents, as they all output low IoU, Precision and F1 accuracies for refugee camps. However, the results still have distinguishable levels of usability: While the models vastly overestimated the camp extents, some were able to localise the refugee camps. Furthermore, they successfully distinguished between formal and refugee settlement. The comparative analysis identified the FPN architecture and the ResNet-18 encoder as the most suitable in comparison for mapping refugee settlement extents. This model was used for analysing the accuracies across space and among different structure types of refugee camps, which were defined by Weigand et al. (2023). Spatial patterns in model accuracy were found in some clusters of camps that exhibit similar accuracy values and also show similarities in their morphological structure.

Nevertheless, the accuracies among regions were still heterogeneous. Furthermore, no large-scale pattern in the model's performance across space could be observed. The analysis of accuracies among different structure types identified variations among the model's performance to map refugee settlements.

The results highlight the complexity of mapping refugee settlement extents worldwide from HR multispectral imagery alone, as there is no perfect approach that fits all structures. Refugee camps are complex settlements, with building materials spectrally similar to their surroundings, unclear boundaries and structures varying across the globe. This work can serve as first advancements for further research on mapping the extents of refugee camps in an automatized way. Future analysis can optimize the approach presented here by e.g., further tuning hyperparameters and applying deeper and more complex models. In this regard, it is of importance to consider the structural differences between the refugee camps, as there is no one-fits-all solution for mapping refugee camp extents. The knowledge on refugee settlement extents can be valuable to NGOs, humanitarian aid organizations or governmental institutions for localising camps, estimating needs of inhabitants and infrastructural requirements, and thereby helping the most vulnerable population groups.

Publication bibliography

Aravena Pelizari, P.; Spröhnle, K.; Geiß, C.; Schöpfer, E.; Plank, S.; Taubenböck, H. (2018): Multi-sensor feature fusion for very high spatial resolution built-up area extraction in temporary settlements. In *Remote Sensing of Environment* 209, pp. 793–807. DOI: 10.1016/j.rse.2018.02.025.

Avtar, R.; Kouser, A.; Kumar, A.; Singh, D.; Misra, P.; Gupta, A. et al. (2021): Remote Sensing for International Peace and Security: Its Role and Implications. In *Remote Sensing* 13 (3), p. 439. DOI: 10.3390/rs13030439.

Benz, S.; Park, H.; Li, J.; Crawl, D.; Block, J.; Nguyen, M.; Altintas, I. (2019): Understanding a Rapidly Expanding Refugee Camp Using Convolutional Neural Networks and Satellite Imagery. In : 2019 15th International Conference on eScience (eScience), pp. 243–251.

Braun, A.; Fakhri, F.; Hochschild, V. (2019): Refugee Camp Monitoring and Environmental Change Assessment of Kutupalong, Bangladesh, Based on Radar Imagery of Sentinel-1 and ALOS-2. In *Remote Sensing* 11 (17), p. 2047. DOI: 10.3390/rs11172047.

Chai, J.; Zeng, H.; Li, A.; Ngai, E. W.T. (2021): Deep learning in computer vision: A critical review of emerging techniques and application scenarios. In *Machine Learning with Applications* 6, p. 100134. DOI: 10.1016/j.mlwa.2021.100134.

Chan, C. Y.-C.; Weigand, M.; Alnajar, E. A.; Taubenböck, H. (2022): Investigating the capability of UAV imagery for AI-assisted mapping of Refugee Camps in East Africa. Zenodo. In Minghini. M., P. Liu, H. Li, A. Y. Grindberger, L. Juhasz (Eds.): Proceedings of the Academic Track at State of the Map. Zenodo. Academic Track at State of the Map. Florence, Italy.

Chi, M.; Sun, Z.; Qin, Y.; Shen, J.; Benediktsson, J. A. (2017): A Novel Methodology to Label Urban Remote Sensing Images Based on Location-Based Social Media Photos. In *Proc. IEEE* 105 (10), pp. 1926–1936. DOI: 10.1109/JPROC.2017.2730585.

Cornebise, J.; Oršolić, I.; Kalaitzis, F. (2022): The WorldStrat Dataset: Open High-Resolution Satellite Imagery With Paired Multi-Temporal Low-Resolution.

ESA, European Commission (2015): Sentinel-2. Data Access and Products. Available online at https://sentinel.esa.int/documents/247904/1848117/Sentinel-2_Data_Products_and_Access, checked on 5/10/2023.

Esch, T.; Heldens, W.; Hirner, A.; Keil, M.; Marconcini, M.; Roth, A. et al. (2017): Breaking new ground in mapping human settlements from space – The Global Urban Footprint. In *ISPRS journal of photogrammetry and remote sensing : official publication of the International Society for Photogrammetry and Remote Sensing (ISPRS)* 134, pp. 30–42. DOI: 10.1016/j.isprsjprs.2017.10.012.

Fransen, S.; Haas, H. (2022): Trends and Patterns of Global Refugee Migration. In *Population and development review* 48 (1), pp. 97–128. DOI: 10.1111/padr.12456.

Friedrich, H. K.; van den Hoek, J. (2020): Breaking ground: Automated disturbance detection with Landsat time series captures rapid refugee settlement establishment and growth in North Uganda. In *Computers, Environment and Urban Systems* 82, p. 101499. DOI: 10.1016/j.compenvurbsys.2020.101499.

Gao, Y.; Gella, G. W.; Liu, N. (2022a): Assessing the Influences of Band Selection and Pretrained Weights on Semantic-Segmentation-Based Refugee Dwelling Extraction from Satellite Imagery. In *AGILE GIScience Ser.* 3, pp. 1–6. DOI: 10.5194/agile-giss-3-36-2022.

Gao, Y.; Lang, S.; Tiede, D.; Gella, G. W.; Wendt, L. (2022b): Comparing OBIA-Generated Labels and Manually Annotated Labels for Semantic Segmentation in Extracting Refugee-Dwelling Footprints. In *Applied Sciences* 12 (21), p. 11226. DOI: 10.3390/app122111226.

Gao, Y.; Lang, S.; Tiede, D.; Gella, G. W.; Wendt, L. (Eds.) (2022c): Comparing the robustness of U-Net, LinkNet, and FPN towards label noise for refugee dwelling extraction from satellite imagery. 2022 IEEE Global Humanitarian Technology Conference (GHTC). Santa Clara, CA, USA, 08-11 September 2022.

Gella, G. W.; Tiede, D.; Lang, S.; Wendt, L.; Gao, Y. (2023): Spatially transferable dwelling extraction from Multi-Sensor imagery in IDP/Refugee Settlements: A meta-Learning approach. In *International Journal of Applied Earth Observation and Geoinformation* 117, p. 103210. DOI: 10.1016/j.jag.2023.103210.

Gella, G. W.; Wendt, L.; Lang, S.; Braun, A.; Tiede, D.; Hofer, B. et al. (2021): Testing Transferability of Deep-Learning-Based Dwelling Extraction in Refugee Camps. In *giforum* 1, pp. 220–227. DOI: 10.1553/giscience2021_01_s220.

Gella, G. W.; Wendt, L.; Lang, S.; Tiede, D.; Hofer, B.; Gao, Y.; Braun, A. (2022): Mapping of Dwellings in IDP/Refugee Settlements from Very High-Resolution Satellite Imagery Using a Mask Region-Based Convolutional Neural Network. In *Remote Sensing* 14 (3), p. 689. DOI: 10.3390/rs14030689.

Ghorbanzadeh, O.; Crivellari, A.; Tiede, D.; Ghamisi, P.; Lang, S. (2022): Mapping Dwellings in IDP/Refugee Settlements Using Deep Learning. In *Remote Sensing* 14 (24), p. 6382. DOI: 10.3390/rs14246382.

Goodfellow, I.; Bengio, Y.; Courville, A. (2016): Deep Learning. Cambridge, Mass.: MIT Press Ltd (Adaptive Computation and Machine Learning Series). Available online at https://www.deeplearningbook.org/.

Hassan, M.; Smith, A.; Walker, K.; Rahman, M.; Southworth, J. (2018): Rohingya Refugee Crisis and Forest Cover Change in Teknaf, Bangladesh. In *Remote Sensing* 10 (5), p. 689. DOI: 10.3390/rs10050689.

He, H.; Garcia, E. A. (2009): Learning from Imbalanced Data. In *IEEE Trans. Knowl. Data Eng.* 21 (9), pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

He, K.; Zhang, X.; Ren, S.; Sun, J. (2015): Deep Residual Learning for Image Recognition. Available online at http://arxiv.org/pdf/1512.03385v1.

Herfort, B.; Lautenbach, S.; Porto de Albuquerque, J.; Anderson, J.; Zipf, A. (2021): The evolution of humanitarian mapping within the OpenStreetMap community. In *Scientific reports* 11 (1), p. 3037. DOI: 10.1038/s41598-021-82404-z.

Hornby, A. S. (2023): Oxford advanced learner's dictionary of Current English. B2-C2. 10th Edition. Oxford: Oxford University Press.

Höser, T.; Künzer, C. (2020): Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends. In *Remote Sensing* 12 (10), p. 1667. DOI: 10.3390/rs12101667.

Hossain, F.; Moniruzzaman, Dr.Md. (2021): Environmental change detection through remote sensing technique: A study of Rohingya refugee camp area (Ukhia and Teknaf sub-district), Cox's Bazar, Bangladesh. In *Environmental Challenges* 2, p. 100024. DOI: 10.1016/j.envc.2021.100024.

iDMC (2022): GRID 2022 - Global Report on Internal Displacement. iDMC. Available online at https://www.internal-displacement.org/sites/default/files/publications/documents/IDMC_GRID_2022_LR.pdf, checked on 4/25/2023.

iDMC (2023): GRID 2023 - Global Report on Internal Displacement and Food Security. iDMC, checked on 6/9/2023.

Johnson, J. M.; Khoshgoftaar, T. M. (2019): Survey on deep learning with class imbalance. In *J Big Data* 6 (1). DOI: 10.1186/s40537-019-0192-5.

Kingma, D. P.; Ba, J. (2014): Adam: A Method for Stochastic Optimization. Available online at https://arxiv.org/pdf/1412.6980.

Kirillov, A.; Girshick, R.; He, K.; Dollár, P. (Eds.) (2019): Panoptic Feature Pyramid Networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Kraff, N. J.; Wurm, M.; Taubenböck, H. (2020): Uncertainties of Human Perception in Visual Image Interpretation in Complex Urban Environments. In *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 13, pp. 4229–4241. DOI: 10.1109/JSTARS.2020.3011543.

Kraff, N. J.; Wurm, M.; Taubenböck, H. (2022): Housing forms of poverty in Europe - A categorization based on literature research and satellite imagery. In *Applied Geography* 149, p. 102820. DOI: 10.1016/j.apgeog.2022.102820.

Kuffer, M.; Pfeffer, K.; Sliuzas, R.; Baud, I. (2016): Extraction of Slum Areas From VHR Imagery Using GLCM Variance. In *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 9 (5), pp. 1830–1840. DOI: 10.1109/JSTARS.2016.2538563.

Lang, S.; Füreder, P.; Riedler, B.; Wendt, L.; Braun, A.; Tiede, D. et al. (2020): Earth observation tools and services to increase the effectiveness of humanitarian assistance. In *European Journal of Remote Sensing* 53 (sup2), pp. 67–85. DOI: 10.1080/22797254.2019.1684208.

LeCun, Y.; Bengio, Y.; Hinton, G. (2015): Deep learning. In *Nature* 521 (7553), pp. 436–444. DOI: 10.1038/nature14539.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. (2016): Feature Pyramid Networks for Object Detection. Available online at http://arxiv.org/pdf/1612.03144v2.

Liu, R.; Kuffer, M.; Persello, C. (2019): The Temporal Dynamics of Slums Employing a CNN-Based Change Detection Approach. In *Remote Sensing* 11 (23), p. 2844. DOI: 10.3390/rs11232844.

Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B. A. (2019): Deep learning in remote sensing applications: A meta-analysis and review. In *ISPRS journal of photogrammetry and remote sensing : official publication of the International Society for Photogrammetry and Remote Sensing (ISPRS)* 152, pp. 166–177. DOI: 10.1016/j.isprsjprs.2019.04.015.

Marconcini, M.; Metz-Marconcini, A.; Esch, T.; Gorelick, N. (2021): Understanding Current Trends in Global Urbanisation - The World Settlement Footprint Suite. In *giforum* 1, pp. 33–38. DOI: 10.1553/giscience2021_01_s33.

Marconcini, M.; Metz-Marconcini, A.; Üreyen, S.; Palacios-Lopez, D.; Hanke, W.; Bachofer, F. et al. (2020): Outlining where humans live, the World Settlement Footprint 2015. In *Sci Data* 7 (1), p. 242. DOI: 10.1038/s41597-020-00580-5.

Mboga, N.; Persello, C.; Bergado, J.; Stein, A. (2017): Detection of Informal Settlements from VHR Images Using Convolutional Neural Networks. In *Remote Sensing* 9 (11), p. 1106. DOI: 10.3390/rs9111106.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G. et al. (2019): PyTorch: An Imperative Style, High-Performance Deep Learning Library. Available online at http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Persello, C.; Stein, A. (2017): Deep Fully Convolutional Networks for the Detection of Informal Settlements in VHR Images. In *IEEE Geosci. Remote Sensing Lett.* 14 (12), pp. 2325–2329. DOI: 10.1109/LGRS.2017.2763738.

Pesaresi, M.; Huadong, G.; Blaes, X.; Ehrlich, D.; Ferri, S.; Gueguen, L. et al. (2013): A Global Human Settlement Layer From Optical HR/VHR RS Data: Concept and First Results. In *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 6 (5), pp. 2102–2131. DOI: 10.1109/JSTARS.2013.2271445.

Quinn, J. A.; Nyhan, M. M.; Navarro, C.; Coluccia, D.; Bromley, L.; Luengo-Oroz, M. (2018): Humanitarian applications of machine learning with remote-sensing data: review and case study in refugee settlement mapping. In *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 376 (2128). DOI: 10.1098/rsta.2017.0363.

Ramadan, A. (2013): Spatialising the refugee camp. In *Transactions of the Institute of British Geographers* 38 (1), pp. 65–77. DOI: 10.1111/j.1475-5661.2012.00509.x.

Ronneberger, O.; Fischer, P.; Brox, T. (2015): U-Net: Convolutional Networks for Biomedical Image Segmentation. Available online at https://arxiv.org/pdf/1505.04597.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. (2019): MobileNetV2: Inverted Residuals and Linear Bottlenecks. Available online at http://arxiv.org/pdf/1801.04381v4.

Skansi, S. (2018): Introduction to Deep Learning. From Logical Calculus to Artificial Intelligence: Springer.

Stark, T.; Wurm, M.; Taubenböck, H.; Zhu, X. X. (2019): Slum Mapping in Imbalanced Remote Sensing Datasets Using Transfer Learned Deep Features. In : 2019 Joint Urban Remote Sensing Event (JURSE). 2019 Joint Urban Remote Sensing Event (JURSE). Vannes, France, 22.05.2019 - 24.05.2019: IEEE, pp. 1–4.

Stark, T.; Wurm, M.; Zhu, X. X.; Taubenböck, H. (2020): Satellite-Based Mapping of Urban Poverty With Transfer-Learned Slum Morphologies. In *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 13, pp. 5251–5263. DOI: 10.1109/JSTARS.2020.3018862.

Taha, A. A.; Hanbury, A. (2015): Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. In *BMC medical imaging* 15, p. 29. DOI: 10.1186/s12880-015-0068-x.

Tan, M.; Le, Q. V. (2019): EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning*. Available online at http://arxiv.org/pdf/1905.11946v5.

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? In *Remote Sensing of Environment* 232, p. 111353. DOI: 10.1016/j.rse.2019.111353.

Turner, S. (2016): What Is a Refugee Camp? Explorations of the Limits and Effects of the Camp. In *Journal of Refugee Studies* 29 (2), pp. 139–148. DOI: 10.1093/jrs/fev024.

UNHCR (2010): Convention and Protocol relating to the Status of Refugees. Available online at https://www.unhcr.org/media/28185, checked on 5/13/2023.

UNHCR (2016): Shelter Design Catalogue. United Nations Division of Programme Support and Management. Geneva.

UNHCR (2022a): Global Trends Report 2021. UNHCR. Available online at https://www.unhcr.org/62a9d1494/global-trends-report-2021, checked on 2/8/2023.

UNHCR (2022b): Refugee Population Statistics Database. UNHCR. Available online at https://www.unhcr.org/refugee-statistics/, updated on 10/27/2022, checked on 2/8/2023.

UNHCR; UNHCR GIS DATA (2023): Refugee camps and other people of concern's locations. Available online at https://www.arcgis.com/home/webmap/viewer.html?webmap=24cad2271eaf4219832bf82da5803193, checked on 5/10/2023.

USA for UNHCR (2021): Refugee Camps Explained. USA for UNHCR. Available online at https://www.unrefugees.org/news/refugee-camps-explained/, updated on 4/6/2021, checked on 5/13/2023.

USA for UNHCR (2023a): Refugee Camps. UNHCR. Available online at https://www.unrefugees.org/refugee-facts/camps, checked on 5/12/2023.

USA for UNHCR (2023b): Rohingya Refugee Crisis. UNHCR. Available online at https://www.unrefugees.org/emergencies/rohingya/, checked on 5/12/2023.

van den Hoek, J.; Friedrich, H. K. (2021): Satellite-Based Human Settlement Datasets Inadequately Detect Refugee Settlements: A Critical Assessment at Thirty Refugee Settlements in Uganda. In *Remote Sensing* 13 (18), p. 3574. DOI: 10.3390/rs13183574.

van den Hoek, J.; Friedrich, H. K.; Ballasiotes, A.; Peters, L. E. R.; Wrathall, D. (2021): Development after Displacement: Evaluating the Utility of OpenStreetMap Data for Monitoring Sustainable Development Goal Progress in Refugee Settlements. In *IJGI* 10 (3), p. 153. DOI: 10.3390/ijgi10030153.

Weigand, M.; Staab, J.; Wurm, M.; Taubenböck, H. (2020): Spatial and semantic effects of LUCAS samples on fully automated land use/land cover classification in high-resolution Sentinel-2 data. In *International Journal of Applied Earth Observation and Geoinformation* 88, p. 102065. DOI: 10.1016/j.jag.2020.102065.

Weigand, M.; Worbis, S.; Sapena, M.; Taubenböck, H. (2023): A structural catalogue of the settlement morphology in refugee and IDP camps. In *International Journal of Geographical Information Science*, pp. 1–27. DOI: 10.1080/13658816.2023.2189724.

Wendt, L.; Lang, S.; Rogenhofer, E. (2017): Monitoring of Refugee and Camps for Internally Displaced Persons Using Sentinel-2 Imagery – A Feasibility Study. In *giforum* 1, pp. 172–182. DOI: 10.1553/giscience2017_01_s172.

Wickert, L.; Bogen, M.; Richter, M. (2021): Lessons Learned on Conducting Dwelling Detection on VHR Satellite Imagery for the Management of Humanitarian Operations. Fraunhofer-Gesellschaft. Available online at https://publica.fraunhofer.de/handle/publica/266645.

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. In *ISPRS journal of photogrammetry and remote sensing : official publication of the International Society for Photogrammetry and Remote Sensing (ISPRS)* 150, pp. 59–69. DOI: 10.1016/j.isprsjprs.2019.02.006.

Wurm, M.; Weigand, M.; Schmitt, A.; Geiß, C.; Taubenböck, H. (2017): Exploitation of textural and morphological image features in Sentinel-2A data for slum mapping. In : 2017 Joint Urban Remote Sensing Event (JURSE). 2017 Joint Urban Remote Sensing Event (JURSE). Dubai, United Arab Emirates, 06.03.2017 - 08.03.2017: IEEE, pp. 1–4.

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. In *IEEE Geosci. Remote Sens. Mag.* 5 (4), pp. 8–36. DOI: 10.1109/MGRS.2017.2762307.

Zhu, Y.; Geiß, C.; So, E.; Jin, Y. (2021): Multitemporal Relearning With Convolutional LSTM Models for Land Use Classification. In *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing* 14, pp. 3251–3265. DOI: 10.1109/JSTARS.2021.3055784.

Figure Sources

Figure 2, p.7:

Kigeme, Rwanda: ©Oxfam East Africa, CC 2.0. Available at:
https://upload.wikimedia.org/wikipedia/commons/5/50/Kigeme_refugee_camp_%288073663190%29.jpg. Last checked: 12.06.2022

Kutupalong, Bangladesh: ©Tanvir Murad Topu / World Bank. Available at:
https://www.flickr.com/photos/worldbank/41405642400/in/photolist-265SJWd-BMgcb3-Z1fuVm-GfSFwu-24qUqiJ-9QR8cx-DKRF7o-2hvJ6Um-ejznua-Zm8QAj-ejSd3F-Xsc7Vt-M4uaVW-Zm7F5h-24jKcQy-SnpkFj-Yu94MN-21Y7R6D-9Hp8mj-XZjfUK-2fZDFL2-ZoRh54-JoAvkX-egknK6-2fV43CN-6i5TjP-XSV3R9-265SJML-JHQmrT-a15inK-f7nAso-GRU1sd-265SK7U-KbS1a2-ZoRgJK-ZoRhrr-ZjmzRJ-2erb9Wp-ZwLFW4-2aGipbK-265SK4s-23ErzNe-29oYmVQ-29oYmKu-29oYmx5-dufBqq-29cc1nZ-dufpjU-fxjWHH-DFuyyt   . Last checked: 12.06.2023.

Ifo Dadaab, Kenya: ©UN-Habitat/Julius Mwelu. Available at:
https://www.flickr.com/photos/unhabitat/51240091563/ .Last checked 12.06.2023.

Appendix

Appendix 1: Distribution of refugee camps across the training data.

Index	Country	Number of camps in this study
1	Nigeria	49
2	Chad	23
3	Sudan	21
4	Ethiopia	20
5	Turkey	19
6	Iraq	17
7	Thailand	11
8	Guinea	9
9	Myanmar	8
10	Dem. Rep. Congo	8
11	South Sudan	8
12	Rwanda	6
13	Kenya	6
14	Bangladesh	6
15	Algeria	5
16	Burundi	4
17	Ghana	4
18	China	3
19	Greece	3
20	Syria	3
21	Djibouti	3
22	Jordan	3
23	Cameroon	3
24	Niger	3
25	Tanzania	2
26	Burkina Faso	2
27	Congo	2
28	Liberia	2
29	Timor-Leste	2
30	Nepal	2
31	Botswana	1
32	Indonesia	1
33	Central African Republic	1
34	Zambia	1
35	Yemen	1
36	Mauritania	1
37	Malawi	1
38	Mozambique	1
39	Tunisia	1
40	Somalia	1

Appendix 2: Largest refugee camps included in this study. USA for UNHCR (2023a).

Name	Country	Inhabitants (date)	Majority origin country
Kutupalong	Bangladesh	877,710 (Feb 2021)	Rohingya, Myanmar
Dadaab	Kenya	224,462 (Feb 2021)	Primarily Somali refugees
Kakuma	Kenya	163,299 (Feb 2021)	Primarily South Sudan, Somalia, Sudan and the Democratic Republic of the Congo
Azraq	Jordan	37,775 (Feb 2021)	Syria
Zaatari	Jordan	80,000 (July 2022)	Syria

Appendix 3: Confusion Matrix for U-Net with MobileNet-V2 encoder.

	pred noData	pred other	pred formal	pred camp	actual%
actual noData	10.69	0.014	0.02	0.02	10.74
actual other	18.53	30.14	5.36	30.92	84.94
actual formal	0.03	0.04	0.88	0.58	1.54
actual camp	0.04	0.09	0.72	1.94	2.79
pred%	29.29	30.29	6.98	33.46	100

Appendix 4: Confusion Matrix for U-Net with ResNet-18 encoder.

	pred noData	pred other	pred formal	pred camp	actual%
actual noData	10.69	0.009	0.01	0.02	10.73
actual other	16.38	47.66	8.06	12.85	84.94
actual formal	0.04	0.08	0.77	0.64	1.54
actual camp	0.05	0.13	0.58	2.02	2.79
pred%	27.17	47.88	9.42	15.53	100

Appendix 5: Confusion Matrix for U-Net with EfficientNet-B0 encoder.

	pred noData	pred other	pred formal	pred camp	actual%
actual noData	10.56	0.003	0.05	0.12	10.73
actual other	5.52	18.05	12.54	48.84	84.94
actual formal	0.002	0.02	0.90	0.61	1.54
actual camp	0.005	0.02	0.68	2.08	2.79
pred%	16.08	18.09	14.17	51.65	100

Appendix 6: Confusion Matrix for FPN with EfficientNet-B0 encoder.

	pred noData	pred other	pred formal	pred camp	actual%
actual noData	10.71	0.01	0.01	0.0	10.73
actual other	3.67	59.13	22.14	0.00014	84.94
actual formal	0.002	0.09	1.45	0.000001	1.54
actual camp	0.01	0.30	2.47	0.000002	2.79
pred%	14.40	59.53	26.07	0.00014	100

Appendix 7: Confusion Matrix for FPN with MobileNet-V2 encoder.

	pred noData	pred other	pred formal	pred camp	actual%
actual noData	10.68	0.04	0.01	0.002	10.73
actual other	1.58	64.33	3.29	15.75	84.94
actual formal	0.001	0.08	0.66	0.79	1.54
actual camp	0.002	0.17	0.50	2.12	2.79
pred%	12.26	64.62	4.46	18.66	100

Appendix 8: Confusion Matrix for FPN with ResNet-18 encoder.

	pred noData	pred other	pred formal	pred camp	actual%
actual noData	10,67	0,04	0,01	0,006	10.73
actual other	1,19	70,33	2,93	10,50	84.94
actual formal	0,005	0,15	0,81	0,57	1.54
actual camp	0,002	0,23	0,66	1,90	2.79
pred%	11,86	70,75	4,41	12,97	100

Statement of Originality

Thesis Title: Deep Learning for Refugee Camps - Mapping Settlement Extents with Sentinel-2 Imagery and Semantic Segmentation.

Topic provided by: Prof. Dr. Hannes Taubenböck (DLR, JMU), Jakob Schwalb-Willmann (JMU), Matthias Weigand (DLR)

Submitted by: Katrin Wernicke

Ich versichere, dass ich die vorstehende Arbeit selbstständig und ohne fremde Hilfe angefertigt und mich keiner anderer als der in den beigefügten Verzeichnissen angegebenen Hilfsmittel bedient habe. Alle Textstellen, die wörtlich oder sinngemäß aus Veröffentlichungen Dritter entnommen wurden, sind als solche kenntlich gemacht.

Weitere Personen waren an der geistigen Leistung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe eines Ghostwriters oder einer Ghostwriting-Agentur in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar Geld oder geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Arbeit stehen.

Der Durchführung einer elektronischen Plagiatsprüfung stimme ich hiermit zu. Die eingereichte elektronische Fassung der Arbeit ist vollständig. Mir ist bewusst, dass nachträgliche Ergänzungen ausgeschlossen sind. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht. Ich bin mir bewusst, dass eine unwahre Erklärung zur Versicherung der selbstständigen Leistungserbringung rechtliche Folgen haben kann.

I hereby confirm that this thesis is my own work and that I have used no sources or aids other than the ones stated. All passages in my thesis for which other sources, including electronic media, have been used, be it direct quotes or content references, have been acknowledged as such and the sources cited.

No other persons were involved in the intellectual work on this thesis. In particular, I did not use the help of a ghostwriter or ghostwriting agency. Third parties have not received money or monetary benefits from me, either directly or indirectly, for work related to the content of the submitted work.

I hereby agree to an electronic plagiarism check being carried out. The submitted electronic version of the thesis is complete. I am aware that subsequent changes are not permitted. This thesis has not been submitted to any other examination authority and has not been published. I am aware that an untrue statement regarding this declaration may have legal consequences.

Place, Date, Signature