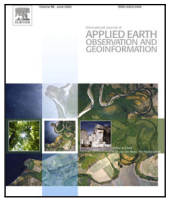
International Journal of Applied Earth Observation and Geoinformation 128 (2024) 103731

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



Interpretable deep learning for consistent large-scale urban population estimation using Earth observation data

Sugandha Doda [a], Matthias Kahl [a], Kim Ouan [a], Ivica Obadic [a], Yuanyuan Wang [a], Hannes Taubenböck [b,c], Xiao Xiang Zhu [a,*]

[a] Chair of Data Science in Earth Observation, Technical University of Munich, Arcisstraße 21, Munich, 80333, Germany
[b] German Remote Sensing Data Center, German Aerospace Center, Münchener Straße 20, Weßling, 82234, Germany
[c] Institute for Geography and Geology, Julius-Maximilians-Universität, Würzburg, 97074, Germany

ARTICLE INFO

Keywords:
Population estimation
Urbanization
Remote sensing
Deep learning
Interpretability
Explainable AI

ABSTRACT

Accurate and up-to-date mapping of the human population is fundamental for a wide range of disciplines, from effective governance and establishing policies to disaster management and crisis dilution. The traditional method of gathering population data through census is costly and time-consuming. Recently, with the availability of large amounts of Earth observation data sets, deep learning methods have been explored for population estimation; however, they are either limited by census data availability, inter-regional evaluations, or transparency. In this paper, we present an end-to-end interpretable deep learning framework for large-scale population estimation at a resolution of 1 km that uses only the publicly available data sets and does not rely on census data for inference. The architecture is based on a modification of the common ResNet-50 architecture tailored to analyze both image-like and vector-like data. Our best model outperforms the baseline random forest model by improving the RMSE by around 9% and also surpasses the community standard product, GHS-POP, thus yielding promising results. Furthermore, we improve the transparency of the proposed model by employing an explainable AI technique that identified land use information to be the most relevant feature for population estimation. We expect the improved interpretation of the model outcome will inspire both academic and non-academic end users, particularly those investigating urbanization or sub-urbanization trends, to have confidence in the deep learning methods for population estimation.

1. Introduction

1.1. Motivation

In 2015, the United Nations embraced the 17 Sustainable Development Goals (SDGs), which aim to support saving humankind and to provide a better future for all by 2030 (UN, 2015). Population distribution data have been identified as a crucial source of data to keep the SDGs on track (UN, 2022). The estimation and mapping of population distribution are also essential to much-informed decision-making issues such as hunger, poverty, education, climate, disaster control, civil protection, etc., and influence the policy-making, planning, and fund allocation of a government (Hay et al., 2005; Hu et al., 2019). Traditionally, national censuses are conducted to gather information about the population count, distribution, and demographics. This data has been extensively used by the government to plan for the future and the required services in a region. Censuses are time-consuming and expensive, thus, they typically take place once a decade, while in some nations, they are held every few decades due to political strife and financial challenges (UN, 2022; Wardrop et al., 2018). They play important roles in many strategic economic developments, business decisions, and planning. However, the rapid change in our society due to climate change and urban migration poses many new challenges such as informal settlement, regional flooding, and infectious diseases. Their requirements on population distribution are beyond the scope of existing census data collection procedures (Tatem et al., 2011). An alternative, less expensive data set with resolution beyond the typical administrative units is needed. There have been efforts to develop large-scale gridded population products that according to the potential location of settlements, redistribute population estimates from census units to grid cells. These gridded population products include the Global Human Settlement Population Grid (GHS-POP) (Freire et al., 2016), Oak Ridge National Laboratory's LandScan (Bhaduri et al., 2002), (WorldPop, 2018), and High-Resolution Settlement Layer (HRSL) (Layer, 2016), among others. Most of these products rely on an

* Corresponding author.
E-mail address: xiaoxiang.zhu@tum.de (X.X. Zhu).

https://doi.org/10.1016/j.jag.2024.103731
Received 11 August 2023; Received in revised form 31 January 2024; Accepted 18 February 2024

external settlement layer to redistribute the known census population counts to grid cells. Also, the difference in their methodology and supplementary data utilized leads to different results, implying that their applicability depends on the context and geographic extent of the application (TReNDS, 2020; Hierink et al., 2022).

1.2. Related work

In recent times, the availability of high-resolution satellite imagery and a surge in deep learning (DL) methods enable more accurate and rapid population estimation (Doupe et al., 2016; Hu et al., 2019; Klemmer et al., 2020; Robinson et al., 2017; Sapena et al., 2022; Tian et al., 2005). Doupe et al. (2016) employ a Convolutional Neural Network (CNN) based method to estimate population densities directly from satellite data. They trained their algorithm on Tanzanian data and predicted Kenya's population at 8 km spatial resolution. Robinson et al. (2017) used Landsat images to estimate the population in United States (US) counties at 1 km resolution. Hu et al. (2019) used multi-source satellite images and a DL technique to estimate India's population density. Huang et al. (2021) used existing population grids from LandScan (Dobson et al., 2000) and Sentinel-2 MultiSpectral Instrument (MSI) to train a DL model to map population trends in two cities in the US using a number of alternative state-of-the-art architectures. Gervasoni et al. (2018) employed a CNN-based method to disaggregate weights into 200 × 200 m grid cells based on urban variables retrieved from the OSM, such as building area and POI count and uses the census data to estimate the population count of a few French cities. The majority of the studies described above are evaluated in only a few cities and rely on census data. Some of the most recent research has been able to function without census data. For example, Metzger et al. (2022) used a DL model to perform population distribution and prediction without always relying on census data. Similar to this, Georganos et al. (2022) suggested a DL-based methodology to estimate the population in three sub-Saharan nations. However, these methodologies have only been evaluated in a comparable geographical environment and morphologies. Some other studies utilized microcensus or census data to automate the population estimation using a very high resolution satellite imagery. Jacobs et al. (2018) trained a CNN to predict fine-grained population maps using very high resolution satellite images and census data. Weber et al. (2018) estimate the population of two northern Nigerian states by identifying human settlements with very high resolution satellite imagery and mapping the population using a micro census. Neal et al. (2022) used very high resolution satellite data to extract building footprints using a representation learning approach combined with a regional micro census to estimate the population of two districts in Mozambique. However, scaling these methods is challenging because of the restricted availability of extremely high resolution satellite data and thus limits their scope of application. Additionally, all these methods lack transparency because their black box models are not unwired to explain the results of their methods.

1.3. Contribution of this article

In this paper, we present an interpretable DL framework to predict the population at a large scale without using census data. The deep learning architecture is based on a modified ResNet (He et al., 2016), which has been widely used in remote sensing and achieving promising results in population mapping (Georganos et al., 2022; Huang et al., 2021; Klemmer et al., 2020). Experiments are done using our recently published openly available So2Sat-POP data set (Doda et al., 2022). This data set consists of 98 cities across Europe and comprises multiple data sources that serve as population indicators. Therefore, we investigate the relevance of different data sources and evaluate our trained model on 18 unseen test cities from this data set. We also analyze our approach in a different continent to include the geographical heterogeneity and compare the results with one of the popular referenced gridded population products in the literature, GHS-POP (Stathama et al., 2021), and the baseline Random Forest (RF) model proposed in the data set paper (Doda et al., 2022).

Although deep learning models have shown tremendous success in various applications, their black-box nature hinders intuitive understanding of the important factors for their predictions (Adadi and Berrada, 2018; Tuia et al., 2021). This raises questions about the transparency and trustworthiness of these models and can limit their usage in sensitive applications like population estimation. Therefore, to shed light on the workings of our ResNet model for population estimation, we apply the Integrated Gradients explainable AI approach (Sundararajan et al., 2017) that highlights the salient features for the model predictions. Uncovering the relevant features for population estimation not only improves the transparency of the proposed deep learning model but also points to some specific difficulties when relying only on remote sensing data for population estimation, such as the inability to distinguish between the different types of built-up areas.

The contributions of this article are listed below:

- We developed an interpretable DL approach that exclusively employs the publicly available Earth observation data at a large scale, with some modalities such as local climate zones and relational statistics among OpenStreetMap (OSM) nodes that have not been used in prior DL-based population estimation studies.
- The model is customized to handle the raster and vector data simultaneously, commonly required for population estimation, and can estimate the population even when no census counts are available by generalizing across countries.
- The relevant features in a multi-source data set and further insights into the model were analyzed using an explainable AI method that is being employed for the first time in population estimation studies. This unboxing of the proposed model improves its transparency and, at the same time, reveals its limitations.
- The accuracy of our new population maps outperforms GHS-POP, a popular state-of-the-art large-scale gridded population product at a spatial resolution of 1 km.
- The proposed method is evaluated using a carefully curated large-scale data set that covers 98 European cities plus three add-on non-European cities. This geographically heterogeneous evaluation illustrates the method's transferability, which has been lacking in the majority of previous population estimation studies.

2. Data

2.1. So2Sat-POP data set

We used the publicly available So2Sat-POP data set (Doda et al., 2022). It is spread across 98 cities in Europe, with 80 cities constituting the training set and the remaining 18 constituting the test set. It is a multi-source data set comprising the digital elevation model, classifications of local climate zones and land use, data on nighttime light emissions, four seasons of Sentinel-2 imagery, and data from the OpenStreetMap initiative. The above-mentioned input data has been prepared for each city and is then processed using the city's population grid. A grid cell in the population grid is 1 × 1 km in size and each cell represents the population count per square kilometer of the cell. All other input data is cropped for each grid cell. As a result, a total of 9 patches, one from each data source, have been created and assigned a population count as of the corresponding population grid cell and a population class based on which bin the grid cell's population count falls into. If a grid cell has a population count of zero, it has been assigned a Class 0, otherwise, it has been assigned a population Class k + 1, if the population count of the cell lies in the range $[2^k, 2^{k+1})$ where $k \in \mathbb{N}$. Therefore, this data set might be used to create both regression and classification models. Fig. 1 depicts a patch-set from Class 10 and an actual population count of 755 as the ground-truth

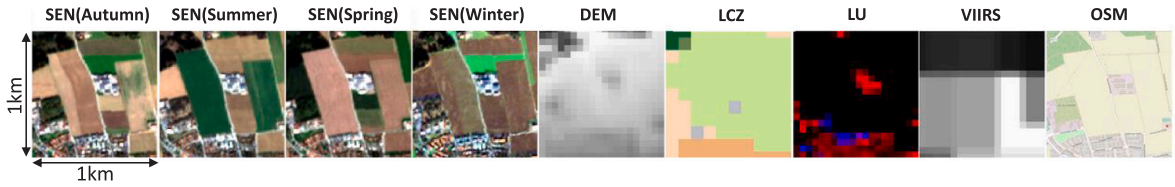


Fig. 1. A patch-set from Class 10 and an actual population count of 755 as the ground-truth labels. Each such patch set consists of 9 patches, one from each input data source.

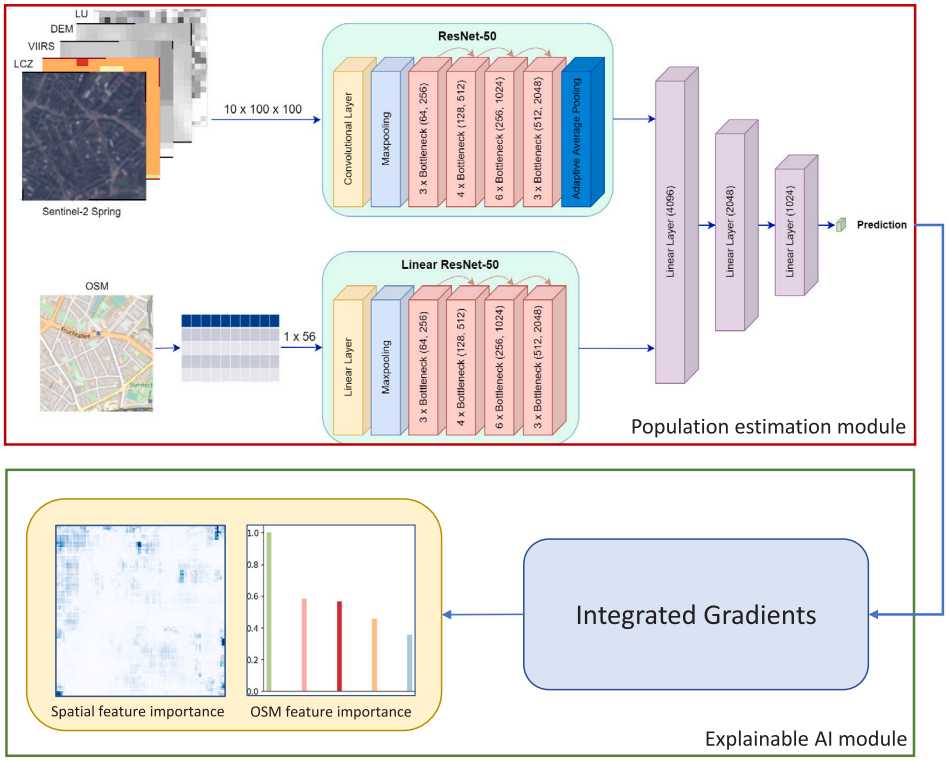


Fig. 2. Our proposed interpretable DL framework for population estimation.

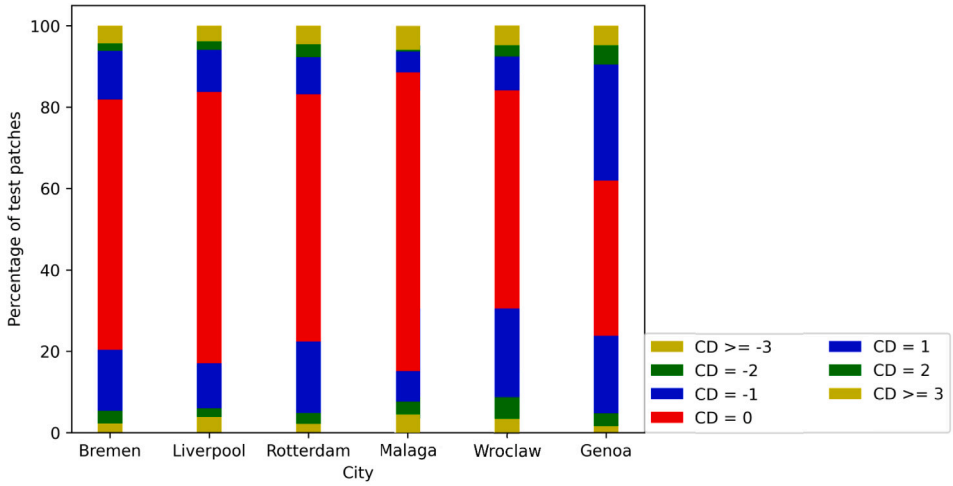


Fig. 3. Class distance (CD) plots of two top-performing cities (Bremen, Liverpool), two average (Rotterdam, Malaga), and two bottom-performing cities (Wroclaw, Genoa) test cities.

labels. The data set consists of 276,172 patches in total. The test-set cities have an area of ~18 292 km^2, whereas the train-set cities have an extent of ~119 794 km^2, with London being the largest and Bilbao being the smallest, with areas of 11 306 km^2 and 54 km^2, respectively.

2.2. Supplementary data set

For the comparison analysis, we collected and processed the data from a popular community standard product for population estimation,

Table 1
A summary of all the data sets used in our work for training and comparison analysis.

Data set	Year	Resolution[a]	Purpose
So2Sat-POP			
Sentinel-2	2017	10 m	
Digital Elevation Model (TanDEM-X)	2016	90 m	So2Sat-POP is a collection of multi-data sources. It is used as the input data for our training pipeline.
Local climate zones (So2Sat LCZ v1.0)	2017	100 m	
Nighttime lights (NPP-VIIRS)	2016	500 m	
OpenStreetMap (OSM)	2017	–	
Population Grids (GEOSTAT- EU)	2011	1 km	
SEDAC Census Grids (US)	2010	1 km	Reference population grid used for the comparison study in the US.
GHS-POP	2015	1 km	Gridded population product compared with our estimates in the EU and US.

[a] To homogenize the input data, all the data sources in the So2Sat-POP data set have been resampled to 10 m (Doda et al., 2022).

Table 2
Comparison of our best models with the random forest model. Across all criteria, our model outperforms the random forest model.

Model	Regression			Classification			
	RMSE	MAE	R^2	Accuracy (%)	Bal. Acc. (%)	F1 score	MACD
Random forest	1276.26	463.35	0.827	59.13	37.95	0.383	0.896
Ours	1164.39	394.38	0.863	61.40	45.25	0.449	0.781

Table 3
Evaluation of different Sentinel-2 seasons on the test set.

Sentinel-2 Season	Regression			Classification			
	RMSE	MAE	R^2	Accuracy (%)	Bal. Acc. (%)	F1 score	MACD
Autumn	1579.59	548.93	0.747	56.52	35.43	0.355	1.01
Spring	1501.62	**545.12**	**0.785**	**57.63**	**37.34**	**0.378**	**0.977**
Summer	1776.90	562.74	0.680	57.58	36.32	0.369	0.981
Winter	**1453.54**	613.46	0.781	55.27	36.85	0.377	1.25

the GHS-POP (Freire et al., 2016). It is a global gridded population data set based on remotely sensed data that has been developed by the EU Joint Research Center. In this data set, each grid cell depicts the number of people at 250 m and 1 km resolution, which has been estimated on the basis of the Global Human Settlement Layer (GHSL) (Calka and Bielecka, 2020). And the reference population data for the comparison is collected from the European Statistical System (ESSnet) project in collaboration with the European Forum for Geography and Statistics (Eurostat, 2011), the same source as that in the So2Sat-POP data set, but unseen in our training, validation, and test data. In the ESSnet project, the grid statistics of the majority of countries were produced through aggregation or a hybrid method (EFGS, 2011). Aggregation (bottom-up approach) is assumed to be the best method for producing population grids (Gallego, 2010). Therefore, it best fits reference population data. For all the grid cells in the test cities where the reference population data was available, we extracted the GHS-POP population counts at a resolution of 1 km and to obtain the GHS-POP population classes, we binned the population counts into population classes as defined in the data set paper (Doda et al., 2022).

Additionally, we created a small data set on three randomly selected cities in the United States (US): New York City, San Jose, and Denver, to evaluate the transferability of our model. The extent of the cities has been determined by the algorithm used in our data set paper (Doda et al., 2022). This algorithm expands the city's extent in order to accommodate the city's fast urbanization. For each city, we collected the data from all the six input modalities utilized in the So2Sat-POP data set, processed and cropped to create 1×1 km patches. We used the Socioeconomic Data and Application Center's (SEDAC) 2010 US Census grids at the resolution of 1 km (http://sedac.ciesin.columbia.edu/) as the reference population data (CIESIN, 2010) for these cities. Table 1 summarizes the various data sets employed in our work, from training purposes to comparison studies in a few EU and US cities.

2.3. Data preprocessing

The So2Sat-POP data set combines data from multiple sources. The data from every input modality has a varying scale. So, we employ different preprocessing techniques and standardize the training data set. For the Sentinel-2 data (only RGB channels) and the VIIRS nighttime light emission data, channels are clipped to the 99.9th percentile, which is calculated channel-wise and is based on all training data samples. The Sentinel-2 images are subsequently normalized in accordance with the suggested preprocessing methods for images when training a ResNet or ResNet-like architecture by Li et al. (2021) to have a channel-wise mean of zero and a channel-wise standard deviation of one. The clipping of the VIIRS data is followed by a min–max normalization to a [0, 1] range to reduce bias caused by the variations in surface materials and seasonal effects.

Land use data values represent areas covered by the respective land use classes, i.e. commercial, industrial, residential, and other, within a raster pixel. To determine the proportion of a pixel that is covered by each land use class, the area is divided by the area of a pixel, resulting in a four-band raster with corresponding land-use proportions percentages for each land-use class. Thus, for each pixel, theoretically, the area proportions (%) across all the bands sum up to 1. We observed, however, that sometimes these values exceed this sum because of the buildings being on top of each other; for example, an underground station and a building on top of it. We normalized such invalid values so that their sum yields 1. The handcrafted features extracted from the OSM dump, such as street density, presence of public transport, number of highways, railways, etc., are in different ranges. We employ min–max normalization to bring all values on the same scale (Pedregosa et al., 2011). The local climate zone classes range from 1 to 17, where classes from 1 to 10 represent the built classes and classes ranging between 11 and 17 represent the natural classes (Stewart and Oke, 2012). To separate the built classes from the natural classes, we mapped all the natural classes (11–17) to 0, and the rest of the classes are mapped between the range of 0.1 and 1, starting from sparsely to densely built-up areas.

Table 4
Evaluation of the significance of various data modalities on the test set by omitting each modality once, except Sentinel-2 (spring). When the specific data modality is removed, blue represents the most affected, while red represents the least affected metrics.

Excluded modality	Regression			Classification			
	RMSE	MAE	R^2	Accuracy (%)	Bal. Acc. (%)	F1 score	MACD
None	**1164.39**	394.38	**0.863**	61.40	**45.25**	**0.449**	0.781
OSM	1216.65	422.18	0.849	61.68	44.25	0.442	0.791
DEM	1181.46	404.87	0.858	**61.71**	43.68	0.444	**0.778**
LU	1270.54	437.71	0.836	58.44	37.68	0.374	0.887
LCZ	1224.74	428.46	0.847	60.63	40.55	0.413	0.833
VIIRS	1168.89	**386.64**	0.861	61.69	42.87	0.436	0.779

Table 5
Quantitative comparison of our best Regression model with GHS-POP on two top-performing (Bremen, Liverpool), two average (Rotterdam, Wroclaw), and two bottom-performing (Malaga, Genoa) test cities.

Cities	Regression					
	Ours			GHS-POP		
	RMSE	MAE	R^2	RMSE	MAE	R^2
Bremen	537.67	272.89	0.933	1530.27	892.60	0.461
Liverpool	616.86	308.88	0.887	1039.60	607.16	0.679
Rotterdam	884.05	427.06	0.890	1871.76	1097.83	0.509
Wroclaw	1184.88	518.58	0.856	2199.25	1096.29	0.515
Malaga	3758.04	1710.72	0.777	6492.98	3901.72	0.334
Genoa	3724.55	2652.69	0.686	2732.62	1906.80	0.831

3. Method

3.1. Interpretable deep learning framework for population estimation

Our proposed framework for population estimation consists of a deep learning model and an explainable AI module based on the Integrated Gradients (IG) method that reveals the relevant features for the predictions of the trained model. Fig. 2 depicts our proposed framework.

The population estimation module is based on the ResNet-50 (He et al., 2016). This architecture excels not only at population estimation tasks (Doupe et al., 2016; Hu et al., 2019; Metzger et al., 2022; Robinson et al., 2017), but also at other multi-modal remote sensing tasks (Ebel et al., 2021; Qiu et al., 2019). Also, this architecture has proven itself to be a good trade-off between the model capacity and performance (Tan and Le, 2019). For the task at hand, a custom architecture with two branches is created to handle both image-like and vector-like features and concatenated before the first fully connected layer. The upper branch is modified to handle inputs of size $10 \times 100 \times 100$ (channels $\times$ width $\times$ height), whereas the lower branch is modified to handle the tabular data. Numerous recent studies have presented deep models for tabular data (Arik and Pfister, 2021; Gorishniy et al., 2021; Huang et al., 2021; Somepalli et al., 2021), and considering the success of ResNet in computer vision and NLP tasks (Sun and Iyyer, 2021), we adapted a linear ResNet-50-like architecture in the lower branch (Gorishniy et al., 2021), where the convolutional layers of the ResNet-50 architecture are replaced with fully connected layers. The two branches are merged following the decision-level fusion protocol of Hoffmann et al. (2019a). For the upper branch, land use and local climate zone classifications, the digital elevation model, nighttime light emissions, and Sentinel-2 imagery are the inputs, and for the lower branch, the OSM feature vector is fed as an input.

The explainable AI module relies on the Integrated Gradients (IG) saliency method (Sundararajan et al., 2017) to examine the outcomes of our black-box model for population estimation. While there are numerous explainability methods that can highlight the relevant input features for a DL model (Ras et al., 2022), in our framework we use the IG method due to the following reasons: Unlike other popular methods such as Grad-CAM (Selvaraju et al., 2017) that are specific for image inputs and convolutional layers, the IG method can attribute feature importance for multi-modal inputs, which also occur in our dataset that consists of both image and tabular data inputs. Moreover, in contrast to other approaches like saliency maps based on the vanilla gradient approach (Simonyan et al., 2013) which can also attribute features of multi-modal inputs, IG satisfies two fundamental axioms, namely sensitivity and implementation invariance. The sensitivity axiom ensures that in case the model outputs different scores for an input example and a baseline input that differs along one feature, then the IG method will assign a non-zero relevance score to this feature. On the other hand, the implementation invariance guarantees that identical feature attributions are assigned for two functionally equivalent models. To satisfy the sensitivity axiom, the IG method relies on a baseline input x' that signifies the absence of features. Next, it defines the feature importance for an input example x as the integral of the gradients for the model predictions on examples that lie along the path from x' to x. In practice, the integral is approximated with a summation, and the importance $I_d(x)$ for a feature d of the input example x is computed with the following equation (Sundararajan et al., 2017):

$$I_d(x) = \frac{(x_d - x'_d)}{m} \sum_{k=1}^{m} \frac{\partial f\left(x' + \frac{k}{m}(x - x')\right)}{\partial x_d} \quad (1)$$

In our framework for population estimation, x is the multi-modal input example consisting of satellite imagery and OSM tabular features, x' is the baseline input consisting of a black image and zero OSM vector, and f is the prediction of our ResNet-50 model for population estimation. Further, m is the number of steps in the path from x' to x.

3.2. Experimental setup

We have two architectural setups; in the first setup, the vector branch is omitted and the final layers are scaled to match the output shape of the image branch. This setup is used when the vector data is not utilized in the experiments and we called it a reduced setup. In the other setup, we utilized both the image and vector branch for training on all the data and called it a complete setup. For both setups, the training set is split into training (80%) and a validation set (20%), and the So2Sat-POP test set is used only to evaluate the trained model at the end. The normal Xavier initialization (Glorot and Bengio, 2010) is used to initialize weights and biases, and training is performed with a batch size of 32 and an ADAM optimizer (Kingma and Ba, 2014) with an initial learning rate of 1×10^{-4}. All the experiments ran for a maximum of 50 epochs, and whenever the training loss did not improve for five subsequent epochs, the learning rate decayed by a factor of 0.1. Batch normalization (Glorot and Bengio, 2010), dropout (Srivastava et al., 2014), and weight decay in combination, following Loshchilov and Hutter (2017) are used as the regularization techniques. We employed various data augmentation techniques. Random Flipping and Random Rotations have commonly used data augmentation techniques in deep learning for population estimation literature (Doupe et al., 2016; Hu et al., 2019). Additionally, we used random brightness and gamma adjustment, which considerably improved the robustness and the performance of the models (Sirazitdinov et al., 2019; Sun et al., 2021). All data augmentation techniques are applied with a probability of 50%.

Table 6
Quantitative comparison of our best Classification model with GHS-POP on two top-performing (Bremen, Liverpool), two average (Rotterdam, Malaga), and two bottom-performing (Wroclaw, Genoa) test cities.

Cities	Classification					
	Ours			GHS-POP		
	Accuracy (%)	Bal. Acc. (%)	MACD	Accuracy (%)	Bal. Acc. (%)	MACD
Bremen	54.15	46.67	0.096	23.42	17.28	1.89
Liverpool	53.03	41.78	0.179	19.13	13.56	2.38
Rotterdam	51.19	48.88	0.071	15.52	11.74	2.39
Malaga	42.26	41.19	0.989	25.77	20.66	1.63
Wroclaw	36.06	42.03	−0.08	19.91	15.46	2.23
Genoa	22.00	15.05	0.340	56.00	31.90	0.92

For the classification, the model output size is set to be 17 as defined by the data set, while for regression, the model output is a single value, a population count. Also, the loss function is set to Mean Squared Error (MSE) for regression and Focal Loss (Lin et al., 2017) for classification. The entire procedure has been implemented with Python 3.8 using the PyTorch 1.10 framework (Paszke et al., 2019). All models are trained on a single NVIDIA RTX 3090 GPU with 24 GB RAM.

3.3. Evaluation metrics

For regression, we employed the commonly used performance metrics, the root-mean-square error (RMSE) and the mean absolute error (MAE), as defined by Eqs. (2) and (3), respectively. To measure the proportion of variance of the population counts that is captured by the model, R^2 is calculated using Eq. (4), where y_i donates the ground truth, $\hat{y}_i$ denote the prediction and $\bar{y}_i$ is the mean of ground truth . For classification, we used balanced accuracy to evaluate the classification performance. We also calculated the accuracy, macro-averaged F1-score, and class distance (CD) as other intuitive classification evaluation metrics. The CD measures the class distance between the predicted and the reference class label. The metric considers the fact that a misclassification to a "nearby" class has a lower error than to a "far away" class due to the underlying regression task. The CD is calculated using Eq. (5).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \, y_i - \hat{y}_i| \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4}$$

$$CD = reference\_class_i - predicted\_class_i \tag{5}$$

4. Results and discussion

4.1. General model performance

Using the complete setup, we trained our models on both image and vector data. We compare our best regression and classification model with the RF model, proposed as a baseline on our data set (Doda et al., 2022). Table 2 shows that we improved the results across all metrics. The balanced accuracy has been improved by approximately 7.5%, and MAE and RMSE by 15% and 8%, respectively. To visually compare the classification model's performance, we plotted a normalized confusion matrix for the two top-performing, two average, and two bottom-performing test cities, shown in Fig. 4. The confusion matrices show that while our model performs poorly on the classes of lower population densities, it is confident in predicting the high-density classes (urban regions). The first four classes reflect areas with extremely low population counts, and it is difficult to tell these classes apart by their attributes. On the other hand, for most cities, the RF model overestimates in low-population classes and underestimates in high-population classes. Fig. 3 shows the CD plots for our model to help us understand the misclassification distance in our predictions. The CD plots indicate the percentage of patches that has a particular CD value in a given city. For instance, in Malaga, approximately ∼75% of the patches have a class distance of zero, and only for ∼8% of the patches the predicted class is three classes distant from its actual class. For regression, we plotted the scatter plots of the predicted population count versus the reference population count for each of these cities. Fig. 5 shows that in our DL model predictions, the scattering is closer to the perfect fitting line for lower population values, while it is more dispersed from the ground truth values for higher population counts. On the other hand, the RF model again tends to over-predict across low population values and under-predict in higher population ranges.

4.2. Importance of input data modalities

Since the So2Sat-POP data set consists of four seasons of Sentinel-2; autumn, spring, summer, and winter, we first analyzed which season is most important. Our reduced setup, without the vector branch, is utilized for this experiment. Individual Sentinel-2 seasonal images were fed into the model without any additional data. Table 3 shows that compared to using any other season, Sentinel-2's spring season was found to yield better performance on the MAE than autumn (by 0.7%), summer (by 3%), and winter (by 11%) and ∼6% on an average on R^2. Also, the classification achieved better-balanced accuracy and F1 scores in the spring season. The mean absolute class distance (MACD), which is basically the class distances averaged across all samples is also lowest in the spring.

Since the So2Sat-POP data set consists of data from six different sources, we also examine the importance of each modality. For both regression and classification, models are trained using different combinations of the data modalities following the leave-one-out principle in cross-validation except for the Sentinel-2 spring season. Using this principle, all data modalities are included in the initial experiment and then for the subsequent trials, each modality is removed once to test if its removal affects the outcomes. Each trained model is evaluated on the 18 unseen test cities. The predictions are on 1 × 1 km patch-sets and compared with the reference population count and population class for each patch-set. Results for this set of experiments are shown in Table 4, with blue-colored metrics representing the most affected when the specific data modality is removed and red-colored metrics representing the least impacted scores. We found that land use is the most crucial input among the other data sets, followed by LCZ, since they have the most influence on the outcomes of both classification and regression. For classification, the exclusion of land use decreased the balanced accuracy by 7.5%, and for regression, increased the mean

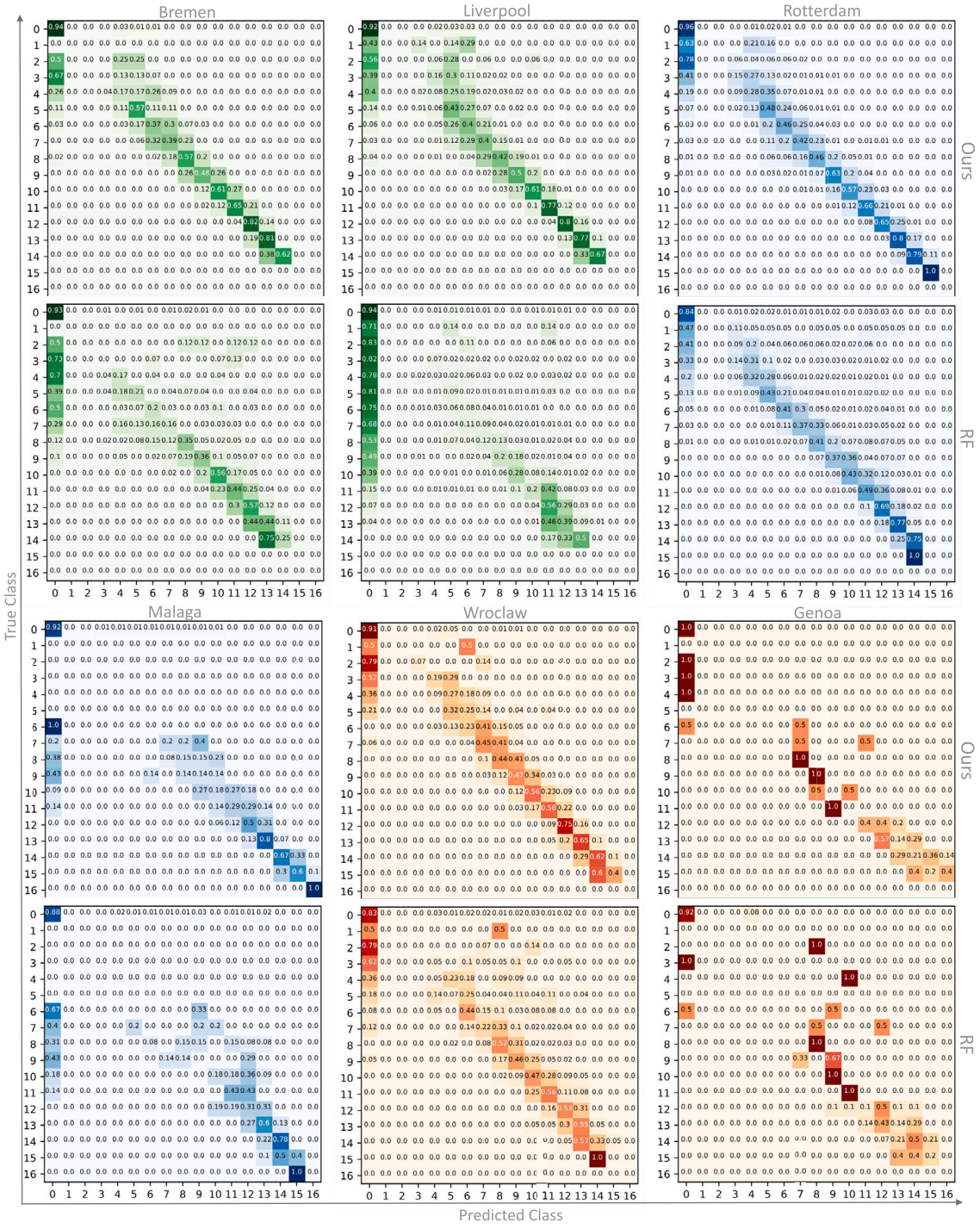


Fig. 4. Normalized confusion matrix of two top-performing cities (Bremen, Liverpool in green), two average (Rotterdam, Malaga in blue), and two bottom-performing cities (Wroclaw, Genoa in red) test cities for our DL and RF model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

absolute error by 11%. However, DEM in classification and VIIRS in regression are found to be the least important because their exclusion has little or no effect on the results. The exclusion of VIIRS increased the RMSE only by 0.3% and even marginally improved the MAE. Similarly, in the classification, excluding DEM slightly improved the accuracy and MACD while decreasing the F1 score by 1% and balanced accuracy by approximately 3%. We furthermore observed that when none of the data modalities were excluded, we achieved the best results on most metrics, so we concluded that all data modalities are important for both regression and classification models.

4.3. Comparison with GHS-POP

To further assess the accuracy of our method, we compare our results with GHS-POP on each of the two top-performing, average-performing, and bottom-performing test cities that have been extracted

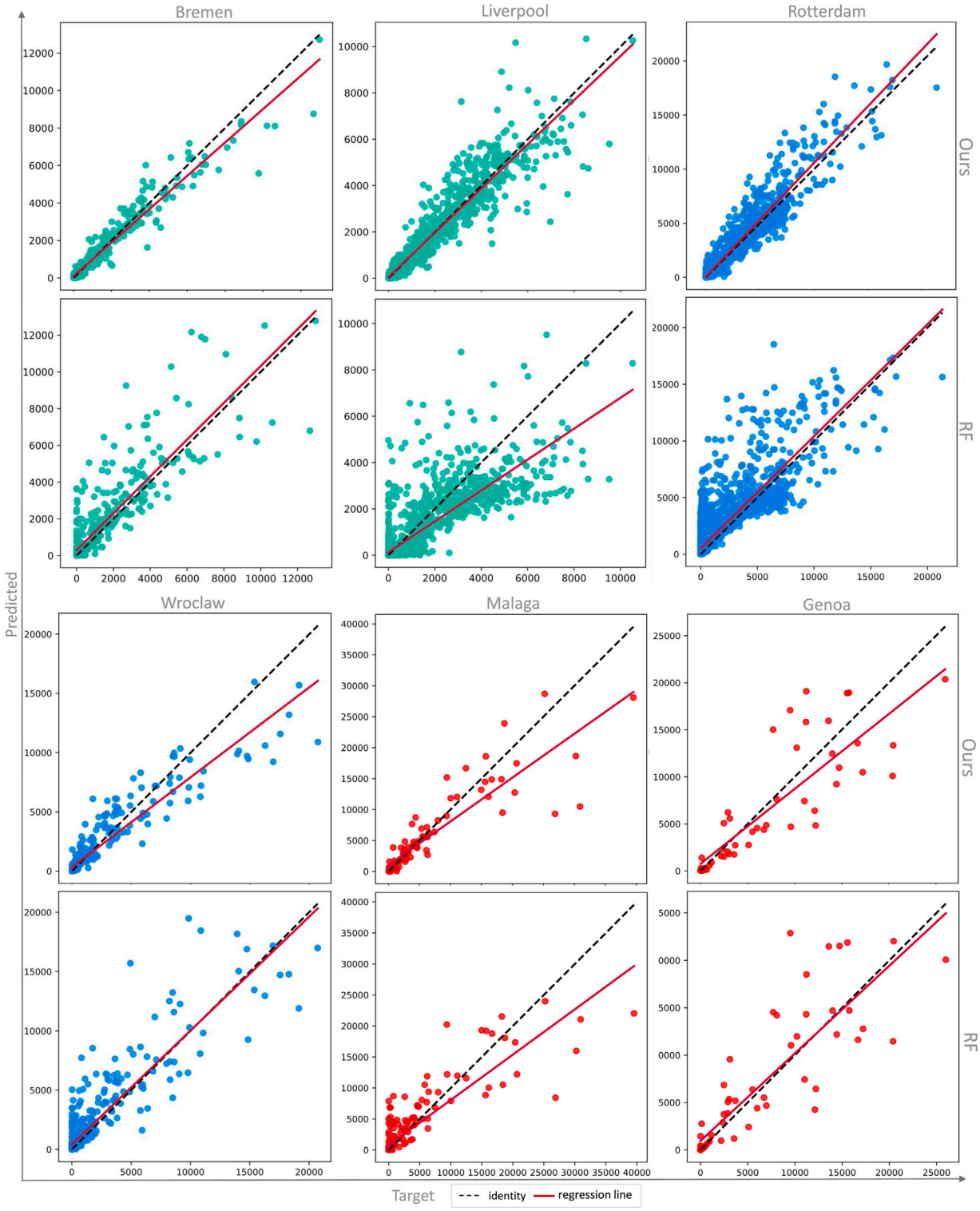


Fig. 5. Scatter plots of our DL model predictions and RF model at the grid level for two top-performing cities (Bremen, Liverpool in green), two average (Rotterdam, Wroclaw in blue), and two bottom-performing cities (Malaga, Genoa in red) test cities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and prepared as the supplementary data set. Tables 5 and 6 represent the quantitative comparison between our predictions and GHS-POP estimates and we have observed that on almost all the evaluation metrics, our approach surpasses the performance of GHS-POP. For example, among the two top-performing and two average-performing cities, the improvement in RMSE ranges from 50% to 70%, and among our two worst-performance cities, Malaga still outperforms the GHS-POP with 42% improvement in RMSE while in Genoa we underperformed by 36%. For classification, our model improved the balanced accuracy in these cities up to 35% except for Genoa. A visual comparison is shown in Figs. 6 and 7. Both regression and classification visualizations show that GHS-POP is not particularly good at capturing densely populated

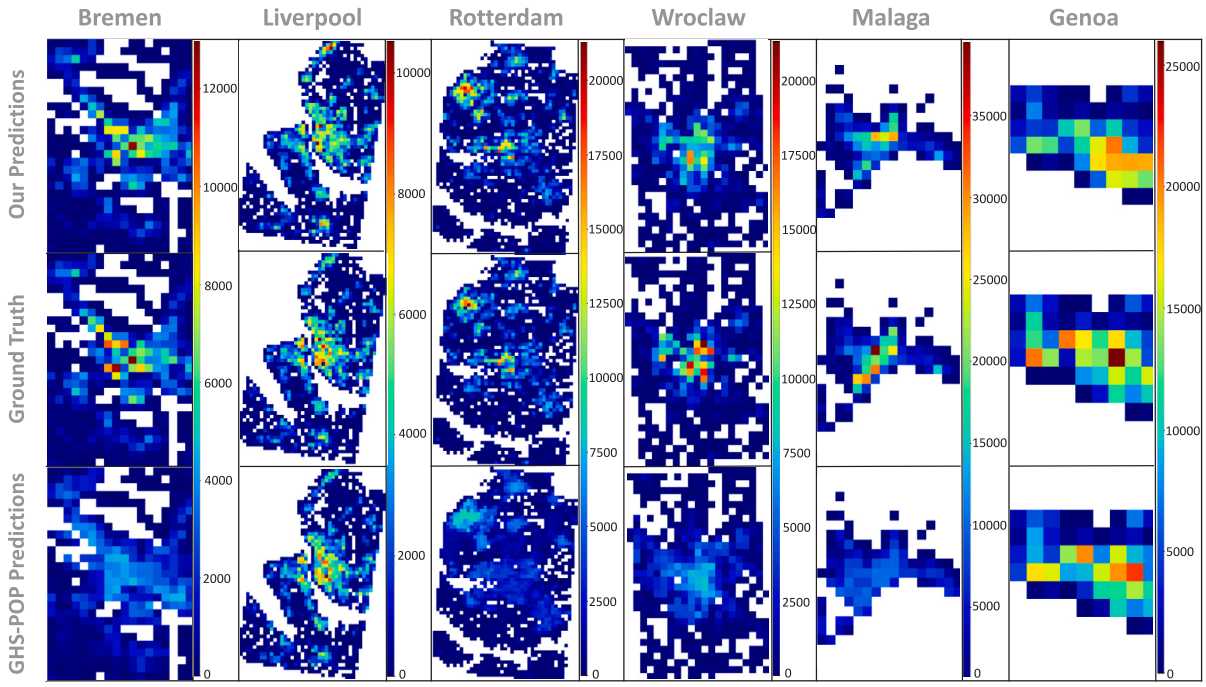


Fig. 6. Comparison of two top-performing cities (Bremen, Liverpool), two average (Rotterdam, Wroclaw), and two bottom-performing cities (Malaga, Genoa) with GHS-POP for regression.

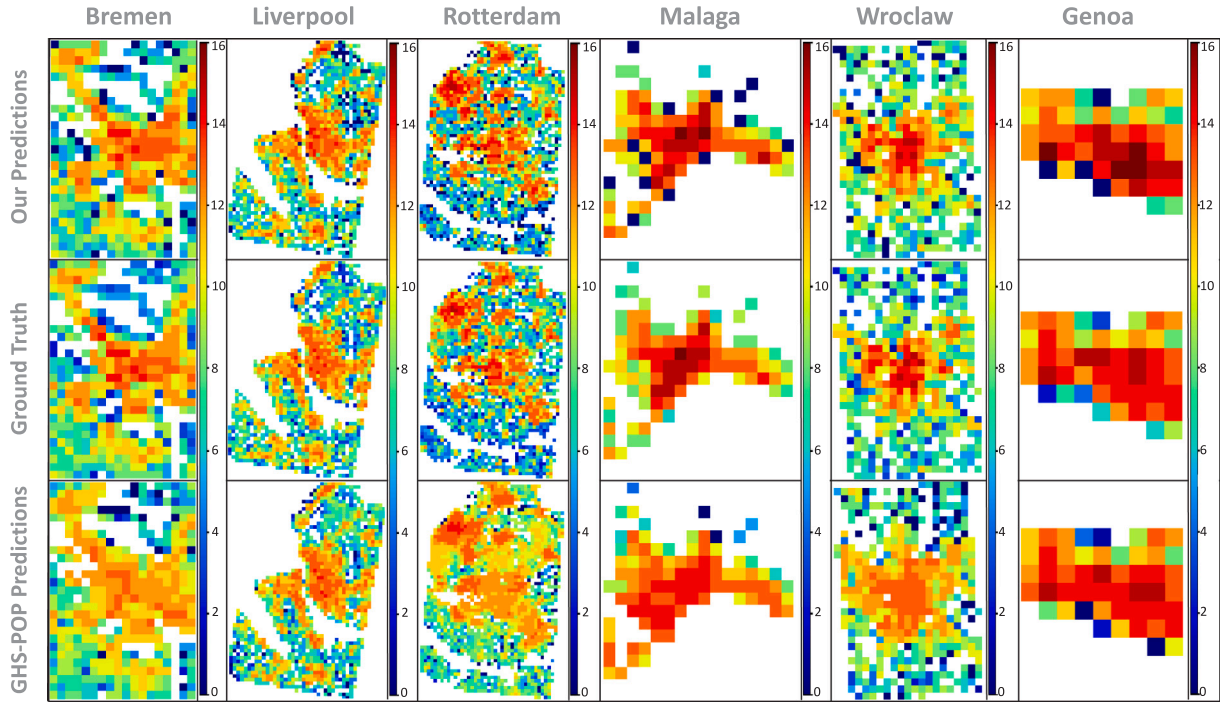


Fig. 7. Comparison of two top-performing cities (Bremen, Liverpool), two average (Rotterdam, Malaga), and two bottom-performing cities (Wroclaw, Genoa) with GHS-POP for classification.

urban areas. It underestimates the population count for densely populated central regions of the city and overall is unable to distinguish between dense and sparsely populated regions of the city.

4.4. Evaluation on inter-regional cities

Since our model is trained only on European cities, therefore we also evaluated our model outside the European Union to test the transferability and generalizability. Using the supplementary data set that we prepared for the three US cities and our best models, we predicted a population count and a population class at a resolution of 1 km. The Tables 7 and 8 show our results in comparison with GHS-POP in the US. Results indicate that our model, which was developed using data from European cities only, does not outperform the GHS-POP. However, our results are in line with GHS-POP and thus our model might be applied to other geographic regions. We believe that by finetuning our model with a few local micro-censuses, it would be possible to improve the performance of our model in a new region too.

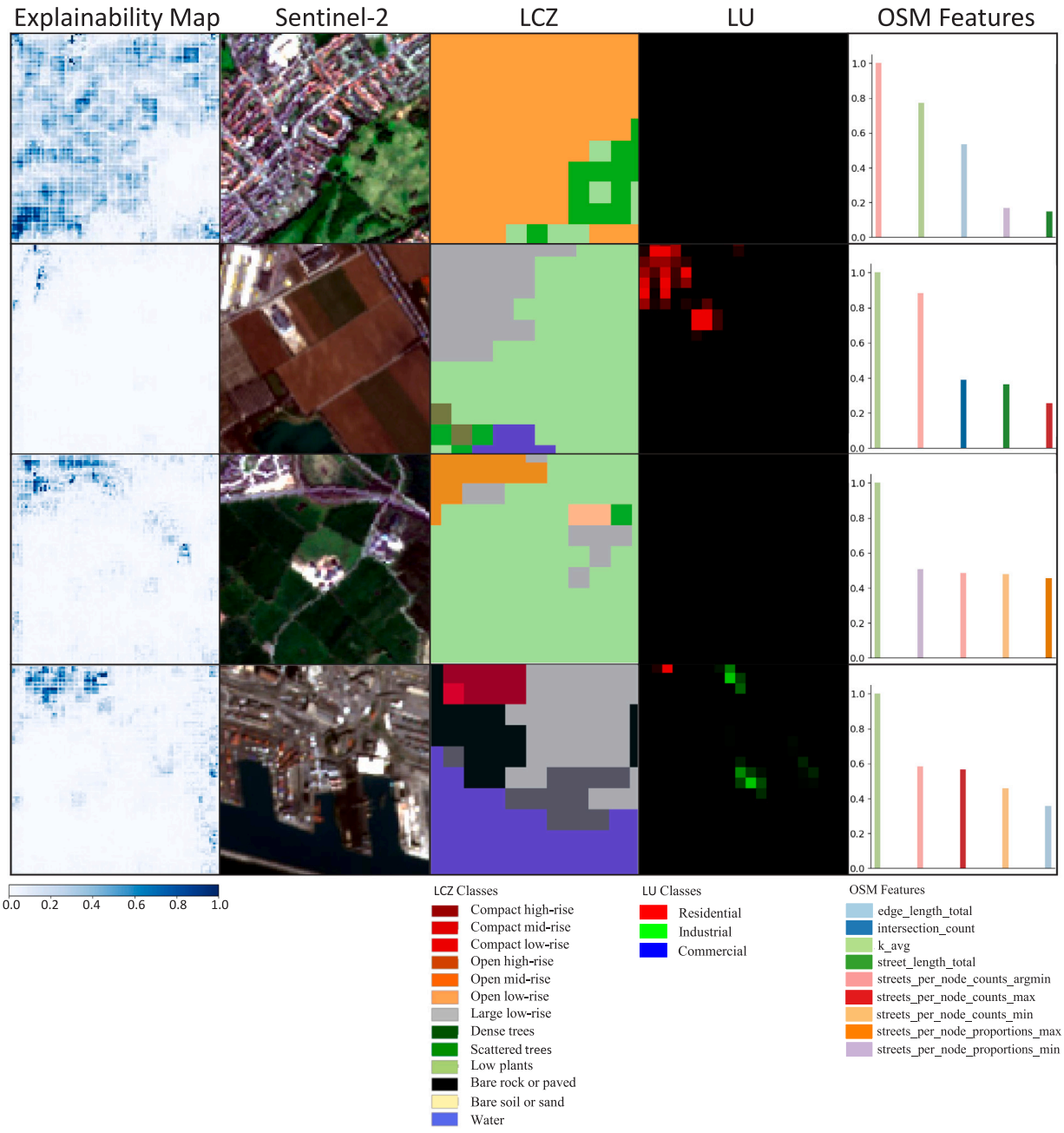


Fig. 8. Feature attribution maps for several examples from the test set. We only include Sentinel-2, LCZ, LU, and OSM patches as they allow for visual interpretation of the semantically significant features. Detailed documentation about the osm geometric and topological network features can be found at OSMnx (Boeing, 2017) user reference (https://osmnx.readthedocs.io/en/stable/osmnx.html).

Table 7
Quantitative comparison of our best Classification model (trained with European cities only) with GHS-POP on three random US test cities.

Cities	Classification					
	Ours			GHS-POP		
	Accuracy (%)	Bal. Acc. (%)	MACD	Accuracy	Balanced Accuracy	MACD
New York	18.70	21.10	2.04	12.23	8.19	3.53
San Jose	38.20	15.46	0.84	41.29	25.73	2.04
Denver	23.48	7.14	2.92	34.02	30.48	1.46

4.5. Understanding relevant features for population estimation

Our previous experimental results show that our model can reliably estimate the population and hence has the potential to support data-driven decision-making in regard to Sustainable Development Goals. Yet, to ensure its acceptance and usage by the relevant stakeholders, it is necessary to improve its transparency by unveiling its inner workings beyond the predictive performance. As stated in Section 3, we use the IG method to reveal the relevant features for population estimation.

Table 8
Quantitative comparison of our best Regression model (trained with European cities only) with GHS-POP on three random US test cities.

Cities	Regression					
	Ours			GHS-POP		
	RMSE	MAE	R^2	RMSE	MAE	R^2
New York	1615.96	674.10	0.60	2042.18	718.14	0.38
San Jose	1550.16	611.40	0.16	761.54	338.71	0.35
Denver	420.69	264.32	0.21	447.03	174.62	0.17

While assessing the importance of the different input modalities in the Section 4.2, we discovered that land use data followed by LCZ data had the greatest impact on the quality of the results. That means data that helps to identify the different types of built-up regions and differentiates them with natural classes is crucial for the model. Therefore, we try to visualize this in the feature attribution maps computed by applying the IG method on four examples from our test data set, displayed in Fig. 8. For each example, we visualize the corresponding Sentinel-2 image, the LCZ, LU, and OSM features that allow an understanding of the semantics of the relevant features. In the first and second instances, the predicted population counts are 4747 and 3, respectively, the same as the reference population count. The feature attribution maps for these examples show that the model focuses on the built-up areas and distinguishes them from the natural environments, such as water or soil for population estimation. This visually confirms the importance of land use and LCZ data, as the derived estimates are based on the settlement areas seen in the Sentinel-2, LCZ, and LU modalities. Further, street-related statistics such as length, count, and proportions rank among the most relevant OSM features. In the third instance, the reference population count is 11, while the predicted count is 216. In this case, though the information is missing from the land use data, a built-up region is clearly visible in the associated Sentinel-2 and LCZ patches. Despite the predicted population count does not match the reference population count, the corresponding feature attribution map indicates that the model has correctly identified the settlements in the areas, and there may be some disparities in the reference data. We suspect that the time lag between the collection of reference population data and other corresponding input data contributed to data noise, resulting in evaluation bias. The patch in the fourth example represents the ports of Genoa having a reference population count of 126, while our model estimates the population count to be 1400. The IG method reveals that the model focuses on the upper left part of the image, which is suggested as a residential area by the Sentinel-2, LCZ, and LU modalities. However, the model identified additional features in the top left as relevant. Furthermore, the model found relevant features along the right side of the image, which supports the over-prediction by the model. We investigated the region and discovered that it is a dock container terminal. As a result, it is full of large containers that could be easily misinterpreted as house roofs in satellite imagery. Therefore, it is mistakenly interpreted as a residential built-up by the model and it overpredicts the population in this patch. This example demonstrates some of the drawbacks of using satellite imagery to estimate population density because in certain cases, the physical characteristics of built-up areas in satellite images are not distinguishable even to humans. Thus, improvements in fine-grained information that would allow the model to differentiate between residential built-up areas and other types of built-up areas, such as heavily packed industrial areas, are required. Therefore, more detailed information about building functions would be quite beneficial in such a scenario (Hoffmann et al., 2019b).

5. Conclusion

We proposed an adaptable and interpretable deep learning framework to estimate the population at a consistent resolution of 1 km using only publicly available data sources. We aim to calculate the population as accurately and transparently as possible so that it can be utilized in real-world applications such as urban planning, developing infrastructure or risk analysis, etc. Our method is trained using the So2Sat-POP data set and tested on 18 unseen European cities, which shows promising results. In most European cities, we observe a better performance than what the standard GHS-POP product offers. Due to the lack of non-European training data, our predictions did not clearly outperform the GHS-POP estimates everywhere in the US and are also likely to happen in other non-European cities. Nevertheless, our approach still performs in a comparable manner. Whereas the GHS-POP and the other commonly employed population estimation methods disaggregate the known population count from census or administrative units to grid cells based on the external settlement layer, our method does not solve the problem of population disaggregation, but rather infer the population estimates from the publicly available remote sensing data. Of course, census data is required for the training of the deep learning model.

To promote the trustworthiness of our model's decisions, we used a popular explainable AI method to assess the most relevant features considered by our model for population estimation. We expect that the interpretation of our model decisions can serve as a reference for comparing the functioning of the deep learning methods in population estimation beyond the predictive performance. To the best of our knowledge, this is the first application of the integrated gradient explainable AI method to the problem of population estimation. The explainability analysis shows that our model is capable of locating the discriminative regions that correspond to built-up areas, which accords with the intuition that land use data is a key predictor of population. However, we have also seen that in certain cases, the model cannot distinguish between different built-up areas. We plan to incorporate these findings in our future work by integrating information about building functions into our model to obtain more accurate and reliable predictions.

Our model has only been trained in European cities. As a result, we expect that our model's predictions have a higher bias in densely or sparsely populated regions such as India, China, and Mongolia, regions with very different climate, architectural, or cultural peculiarities compared to Europe, such as modern US cities without a historic city center, and in desert regions with foreign building materials. We plan to extend our training data to include such missing regions and apply transfer learning methods to fine-tune our pre-trained model using some micro-census data. Nevertheless, even in the absence of census data, our framework could be utilized to generate more accurate, up-to-date, and interpretable population estimation maps at a large scale.

CRediT authorship contribution statement

Sugandha Doda: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Matthias Kahl:** Conceptualization, Data curation, Supervision, Validation, Writing – review & editing. **Kim Ouan:** Data curation, Methodology, Software, Writing – original draft. **Ivica Obadic:** Methodology, Software, Writing – original draft. **Yuanyuan Wang:** Conceptualization, Project administration, Validation, Writing – review & editing. **Hannes Taubenböck:** Supervision, Writing – review & editing. **Xiao Xiang Zhu:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The work is jointly supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. [ERC-2016-StG-714087], Acronym: *So2Sat*), by the Helmholtz Association, Germany through the Framework of the Munich School for Data Science (MUDS), by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001), by German Federal Ministry for Economic Affairs and Climate Action in the framework of the "National Center of Excellence ML4Earth" (grant number: 50EE2201C) and by the Munich Center for Machine Learning.

Code availability

Python is used for all the analyses and implementations. The code is available as a GitHub project at (https://github.com/zhu-xlab/So2Sat-POP-DL.git).

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160.

Arik, S.Ö., Pfister, T., 2021. Tabnet: Attentive interpretable tabular learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 6679–6687.

Bhaduri, B., Bright, E., Coleman, P., Dobson, J., 2002. LandScan. Geoinformatics 5 (2), 34–37.

Boeing, G., 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Comput. Environ. Urban Syst. 65, 126–139.

Calka, B., Bielecka, E., 2020. GHS-POP accuracy assessment: Poland and Portugal case study. Remote Sens. 12 (7), 1105.

CIESIN, 2010. Center for international earth science information network - CIESIN - columbia university. 2017. U.S. census grids (summary file 1), 2010. palisades, new york: NASA socioeconomic data and applications center (SEDAC). URL: https://doi.org/10.7927/H40Z716C. accessed on: 2022-11-03.

Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., Worley, B.A., 2000. LandScan: a global population database for estimating populations at risk. Photogramm. Eng. Remote Sens. 66 (7), 849–857.

Doda, S., Wang, Y., Kahl, M., Hoffmann, E.J., Ouan, K., Taubenböck, H., Zhu, X.X., 2022. So2Sat POP - a curated benchmark data set for population estimation from space on a continental scale. Sci. Data 9 (1), http://dx.doi.org/10.1038/s41597-022-01780-x.

Doupe, P., Bruzelius, E., Faghmous, J., Ruchman, S.G., 2016. Equitable development through deep learning: The case of sub-national population density estimation. In: Proceedings of the 7th Annual Symposium on Computing for Development. pp. 1–10.

Ebel, P., Meraner, A., Schmitt, M., Zhu, X.X., 2021. Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery. IEEE Trans. Geosci. Remote Sens. 59 (7), 5866–5878. http://dx.doi.org/10.1109/TGRS.2020.3024744.

EFGS, 2011. ESSnet project GEOSTAT 1B - final report. URL: https://www.efgs.info/wp-content/uploads/geostat/1b/GEOSTAT1B-final-technical-report.pdf. accessed on: 2022-12-25.

Eurostat, 2011. Gisco geostat 1 km^2 population grid. URL: https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat. accessed on: 2022-10-05.

Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E., Mills, J., 2016. Development of new open and free multi-temporal global population grids at 250 m resolution. Population 250.

Gallego, F.J., 2010. A population density grid of the European union. Popul. Environ. 31 (6), 460–473.

Georganos, S., Hafner, S., Kuffer, M., Linard, C., Ban, Y., 2022. A census from heaven: Unraveling the potential of deep learning and earth observation for intra-urban population mapping in data scarce environments. Int. J. Appl. Earth Obs. Geoinf. 114, 103013.

Gervasoni, L., Fenet, S., Perrier, R., Sturm, P., 2018. Convolutional neural networks for disaggregated population mapping using open data. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics. DSAA, IEEE, pp. 594–603.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, pp. 249–256.

Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A., 2021. Revisiting deep learning models for tabular data. Adv. Neural Inf. Process. Syst. 34.

Hay, S.I., Noor, A.M., Nelson, A., Tatem, A.J., 2005. The accuracy of human population maps for public health application. Trop. Med. Int. Health 10 (10), 1073–1086.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.. pp. 770–778.

Hierink, F., Boo, G., Macharia, P.M., Ouma, P.O., Timoner, P., Levy, M., Tschirhart, K., Leyk, S., Oliphant, N., Tatem, A.J., et al., 2022. Differences between gridded population data impact measures of geographic access to healthcare in sub-saharan africa. Commun. Med. 2 (1), 117.

Hoffmann, E.J., Wang, Y., Werner, M., Kang, J., Zhu, X.X., 2019a. Model fusion for building type classification from aerial and street view images. Remote Sens. 11 (11), 1259.

Hoffmann, E.J., Werner, M., Zhu, X.X., 2019b. Building instance classification using social media images. In: 2019 Jt. Urban Remote Sens. Event.. JURSE, IEEE, pp. 1–4.

Hu, W., Patel, J.H., Robert, Z.-A., Novosad, P., Asher, S., Tang, Z., Burke, M., Lobell, D., Ermon, S., 2019. Mapping missing population in rural India: A deep learning approach with satellite imagery. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 353–359.

Huang, X., Zhu, D., Zhang, F., Liu, T., Li, X., Zou, L., 2021. Sensing population distribution from satellite imagery via deep learning: Model selection, neighboring effects, and systematic biases. IEEE J. Sel. Top. Appl. Earth Obs. 14, 5137–5151.

Jacobs, N., Kraft, A., Rafique, M.U., Sharma, R.D., 2018. A weakly supervised approach for estimating spatial density functions from high-resolution satellite imagery. In: Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 33–42.

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

Klemmer, K., Yeboah, G., de Albuquerque, J.P., Jarvis, S.A., 2020. Population mapping in informal settlements with high-resolution satellite imagery and equitable ground-truth. arXiv preprint arXiv:2009.08410.

Layer, H.R.S., 2016. Facebook connectivity lab and center for international earth science information network-CIESIN-columbia university. Source imagery for hrsl© 2016 DigitalGlobe. Accessed on: 2022-06-21.

Li, F.-F., Krishna, R., Xu, D., 2021. CS231n: Convolutional Neural Networks for Visual Recognition - Lecture 7: Training Neural Networks, Part 1. Stanford University.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proc. IEEE. Int. Conf. Comput. Vis.. pp. 2980–2988.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Metzger, N., Vargas-Muñoz, J.E., Daudt, R.C., Kellenberger, B., Whelan, T.T.-T., Ofli, F., Imran, M., Schindler, K., Tuia, D., 2022. Fine-grained population mapping from coarse census counts and open geodata. Sci. Rep. 12 (1), 20085.

Neal, I., Seth, S., Watmough, G., Diallo, M.S., 2022. Census-independent population estimation using representation learning. Sci. Rep. 12 (1), 5185.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Qiu, C., Mou, L., Schmitt, M., Zhu, X.X., 2019. Local climate zone-based urban land cover classification from multi-seasonal sentinel-2 images with a recurrent residual network. ISPRS J. Photogramm. Remote Sens. 154, 151–162.

Ras, G., Xie, N., Van Gerven, M., Doran, D., 2022. Explainable deep learning: A field guide for the uninitiated. J. Artificial Intelligence Res. 73, 329–396.

Robinson, C., Hohman, F., Dilkina, B., 2017. A deep learning approach for population estimation from satellite imagery. In: Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities. pp. 47–54.

Sapena, M., Kühnl, M., Wurm, M., Patino, J.E., Duque, J.C., Taubenböck, H., 2022. Empiric recommendations for population disaggregation under different data scenarios. Plos one 17 (9), e0274504.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proc. IEEE Int. Conf. Comput. Vis.. pp. 618–626.

Simonyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR, arXiv:1312.6034.

Sirazitdinov, I., Kholiavchenko, M., Kuleev, R., Ibragimov, B., 2019. Data augmentation for chest pathologies classification. In: 2019 IEEE 16th Int. Symp. Biomed. Imaging (ISBI 2019). IEEE, pp. 1216–1219.

Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C.B., Goldstein, T., 2021. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. arXiv preprint arXiv:2106.01342.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.

Stathama, T., Foxa, S., Wolfa, L.J., 2021. Identifying urban areas: A new approach and comparison of national urban metrics with gridded population data. Comput. Environ. Urban Syst..

Stewart, I.D., Oke, T.R., 2012. Local climate zones for urban temperature studies. Bull. Am. Meteorol. Soc. 93 (12), 1879–1900.

Sun, X., Fang, H., Yang, Y., Zhu, D., Wang, L., Liu, J., Xu, Y., 2021. Robust retinal vessel segmentation from a data augmentation perspective. In: International Workshop on Ophthalmic Medical Image Analysis. Springer, pp. 189–198.

Sun, S., Iyyer, M., 2021. Revisiting simple neural probabilistic language models. arXiv preprint arXiv:2104.03474.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In: International Conference on Machine Learning. PMLR, pp. 3319–3328.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tatem, A.J., Campiz, N., Gething, P.W., Snow, R.W., Linard, C., 2011. The effects of spatial population dataset choice on estimates of population at risk of disease. Popul. Health Metr. 9 (1), 1–14.

Tian, Y., Yue, T., Zhu, L., Clinton, N., 2005. Modeling population density using land cover data. Ecol. Model. 189 (1–2), 72–88.

TReNDS, 2020. Leaving no one off the map: A guide for gridded population data for sustainable development.

Tuia, D., Roscher, R., Wegner, J.D., Jacobs, N., Zhu, X., Camps-Valls, G., 2021. Toward a collective agenda on AI for earth science data analysis. IEEE Geosci. Remote Sens. Mag. 9 (2), 88–104. http://dx.doi.org/10.1109/MGRS.2020.3043504.

UN, 2015. Sustainable development goals. URL: http://www.undp.org/content/undp/en/home/sustainable-development-goals.html. accessed on: 2022-03-09.

UN, 2022. United nations population fund. Census. URL: https://www.unfpa.org/census. accessed on: 2023-10-11.

Wardrop, N., Jochem, W., Bird, T., Chamberlain, H., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V., Tatem, A., 2018. Spatially disaggregated population estimates in the absence of national population and housing census data. Proc. Natl. Acad. Sci. 115 (14), 3529–3537.

Weber, E.M., Seaman, V.Y., Stewart, R.N., Bird, T.J., Tatem, A.J., McKee, J.J., Bhaduri, B.L., Moehl, J.J., Reith, A.E., 2018. Census-independent population mapping in northern Nigeria. Remote Sens. Environ. 204, 786–798.

WorldPop, 2018. School of Geography And Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University. Global High Resolution Population Denominators Project-Funded by the Bill and Melinda Gates Foundation (OPP1134076).