

Geographic transferability of building function models

Nikolai Skuppin, Yilei Shi, Xiao Xiang Zhu

Abstract—We study the geographic transferability of U-Nets trained for building type classification. The models assign functional classes, such as residential or non-residential to individual buildings. This extends the common building extraction task by providing more in depth insight. Since training and testing data is generally scarce, it is of fundamental importance to study transferability of models between different areas. To address this, we collected a dataset covering six countries across the globe. We paired Planet Basemap images with OpenStreetMap (OSM) labels and manually cleaned the dataset. We trained semantic segmentation models on various data splits to predict residential and non-residential buildings. The presented results suggest there is no clear benefit from training or fine-tuning models locally, compared to combining the available data and training a global model. Furthermore, we find stacking footprints on top of the input images improves both, the models' F1 scores and across-region transferability.

Index Terms—Geographic Transferability, Building Function, OpenStreetMap, Semantic Segmentation

I. INTRODUCTION

INFORMATION about building functions is vital for urban studies and planning. However, this information is not readily available in many places. A comprehensive work about European building stock found building type information only available for approximately 46% of all buildings with a significant variance between countries [1]. The information was not available for some countries, whereas others had almost full coverage. While Europe can be considered as a data rich area this coverage gap can be expected to be even more pronounced in other parts of the world. Furthermore, the inter-country variability highlights the necessity of checking the model transferability, where models trained in data rich areas should be transferred to data scarce areas. Having increased coverage of building functions allows for large-scale intra- and inter-urban studies. Urban land-use can be mapped from various sources, such as social media data [2]–[4], geospatial data [5], [6], remote sensing images [7], or a combination [8]–[12]. Ground-based Points of Interest (POI) can provide very accurate and fine-grained information, however, they are not readily available [5]. While building function can rarely be observed directly, building form can be obtained from geospatial or remote sensing data. Building function can be inferred by mapping urban form and function [13], [14]. Geospatial data availability and quality varies globally, whereas remote sensing data is readily available. Therefore, this work studies how available geospatial data from OpenStreetMap (OSM)

can be used to train models for deriving building functions from remote sensing imagery. Specifically, the geographic transferability of the U-Nets is studied. Furthermore, we investigate if adding building footprints to the input improves the transferability.

In this work, we train U-Nets for semantic segmentation on different data splits and report their performance across various cities in six countries. We study how models trained only in single countries compare to models trained on data from all countries. Finally, we compare how providing building footprints as a prior input affects the models' transferability.

II. STATE OF THE ART

A. Building Footprints & Attributes

Large-scale building and settlement extraction have been actively researched. The world settlement footprint and its extension, for example, provide a global insight into the presence of human settlements [15], [16]. Similarly, the Global Human Settlement Layer (GHSL) provides information on the presence of buildings and recently also a distinction into residential and non-residential buildings [17]. However, the resolution is too coarse for identifying individual smaller buildings. Both Microsoft and Google released building footprint datasets, which were obtained from very high-resolution satellite images [18], [19]. While they cover large parts of the world, they mainly provide information on the presence of buildings and no further attributes, such as type or height. OpenStreetMap (OSM) [20] partially provides such information but is not complete across the globe. Administrative data released by governmental institutions can often provide the most complete information. However, it is not globally available and is often laborious to obtain. Furthermore, data structuring and annotation differences can hamper the fusion of different data sources. For the EUBUCCO dataset, the authors tackled this challenge and homogenized multiple data sources, including governmental and OpenStreetMap, into one common representation [1]. They find only approximately half of the buildings to have information on their type, with some regions having full coverage and others not. The literature indicates that there is already a good availability of data on where to find buildings, however, variable availability of further attributes, such as height or type.

There is extensive literature on covering this gap by mapping urban land use from multiple data sources [2]–[12], [21]–[24] and a complete discussion is beyond the scope of this paper. Many works combine a multitude of data, such as Points of Interest [5], [25], street view images [22], social media images [26], or general social sensed data [2]–[4], [21]. Yet, many of these studies focus on single cities and transferability of their methods to other cities and regions remains open [3], [5], [8],

Nikolai Skuppin is with the German Aerospace Center (DLR), Weßling, Germany (e-mail: Nikolai.Skuppin@dlr.de).

Yilei Shi is with the School of Engineering and Design, Technical University of Munich, Munich, Germany.

Xiao Xiang Zhu is with the Chair of Data Science in Earth Observation, Technical University of Munich, Munich, Germany and the Munich Center for Machine Learning, Munich, Germany.

[21]–[24]. This work tries to bridge this gap by using globally available satellite images, building footprint data and labels from OpenStreetMap.

B. OpenStreetMap

As described in the previous paragraph OpenStreetMap can provide information on building function and its quality and completeness of has been thoroughly investigated in the literature [27], [28]. [27] performed a comprehensive analysis of the global OSM quality. They studied internal and external quality metrics. For example, the completeness of tagged buildings and adherence to OSM rules can be analyzed using just the data. The completeness of building polygons and the correctness of tags need an external ground truth. They find 19.5% of the buildings to be tagged with their type, with the completeness varying significantly between different areas. They compare OSM building types with street-level images for a small global subset of buildings and find an agreement of 84.4%. [28] use OSM to train decision trees to predict building functions in three areas in the US. Comparing OSM labels to ground truth data from official sources, they find that less than one percent of the buildings are mislabeled. [29] conduct a rule-based classification of buildings in the United States. They assign building functions based on a heuristic, which uses tags, land-use information and class distribution. Their approach results in a high recall for residential and a high precision for non-residential buildings. OpenSentinelMap [30] provides a dataset of Sentinel-2 images paired with OSM-based labels from 15 different classes. The sampling was done disregarding the quantity and quality of the OSM labels. They find a geographic bias in the completeness of annotations with higher coverage in Western Europe. The literature suggests that quality of annotations in OSM tends to be of good quality, which motivates using it for training and testing models. However, it also finds differences in completeness, which motivate training models in annotation rich areas and applying them in data scarce areas.

C. Geographic Transferability

Recently, increased interest has been in studies targeting this geographic transferability. The global coverage of satellite imagery, paired with the often only local availability of labeled data, makes transferability a vital field of study. One common approach is to split the world into multiple regions and to train separate models for each region [16], [31]. [31], for example, worked on globally mapping human settlements from Sentinel-2 data. They distinguish close-range and far-range transfer learning. Given sufficient data they train a model on a subset of an UTM zone. They then apply the model to the complete UTM zone and transfer it to other similar zones where no training data is available, where similarity can be given spatially or by similar landscape. Another approach is to combine the data from various regions into one data set and to train one single model from that. [32], for example, conducted a comprehensive analysis of urban land use classification from Sentinel-2 data. They use data from 200 cities in 78 different countries. They use a multi level classification scheme, first

identifying built-up areas and background, secondly splitting built-up into residential and non-residential and then further subdividing the residential areas into informal settlements. For the residential/non-residential classification they reach an overall accuracy of 79%. They find lower mean accuracy and wider spread for finer classification schemes. Furthermore, they find varying mean and spread for different geographic regions. Lastly, they do not find a strong correlation between city size, measured by population count and training data size, represented by the count of locales per city. Several other works study the across-region [9], [12], [28], across-country [33] or across-continent transferability [34], [35] of land-use and building function models. [35] study the across-continent and across-season generalization of models trained on subsets of SEN12MS [36]. They find models to transfer poorly across continents and to show significant variability, highlighting the necessity of in domain training data and assessing the model's performance in the application region. [33] study model transferability of CNNs for urban land-use classification among six European cities from different countries. They find models trained on the respective test city to perform best, compared to models trained on other cities. Furthermore, models trained on data from all cities do not reach top accuracy on single cities but reach competitive performance on the collection of all cities. [34] study the transferability of semantic segmentation models for informal settlement detection. They either train and test on one city, train on all but one city, or fine-tune on a given city. They further study the effect of adding OSM street network as an input feature. They find fine-tuned models often to be performing the best. Adding the OSM street network generally improves performance and reduces the gaps between non-local and fine-tuned, or local models. [9] study the transferability of random forest for urban land use classification among five cities in the United States. They assemble a multitude of input features, including Sentinel-1 and -2 band statistics, nighttime light intensity and binned tweet counts from twitter to single street blocks and predict the land use. They find poor model transferability between the five tested cities. In contrast to that, [28] study the transferability of decision trees for building classification between three areas in the US. They derived features from map data as input to their random forests. While transferring models from training to testing areas results in slight performance decrease the reduction is much less pronounced than in [9]. They find their models to transfer well, especially for the residential class. Furthermore, they suggest a correlation between training dataset size and transferability, i.e., a model trained on fewer data to transfer worse. The presented literature suggests that transferability of a model is influenced by a multitude of factors, such as input data, classification model and scheme and definition of training and testing domains.

D. Data Fusion

[37] studied unsupervised domain adaptation for built-up area extraction from Sentinel-1 and Sentinel-2 data. They train an optical, a Synthetic Aperture Radar (SAR) and a fusion

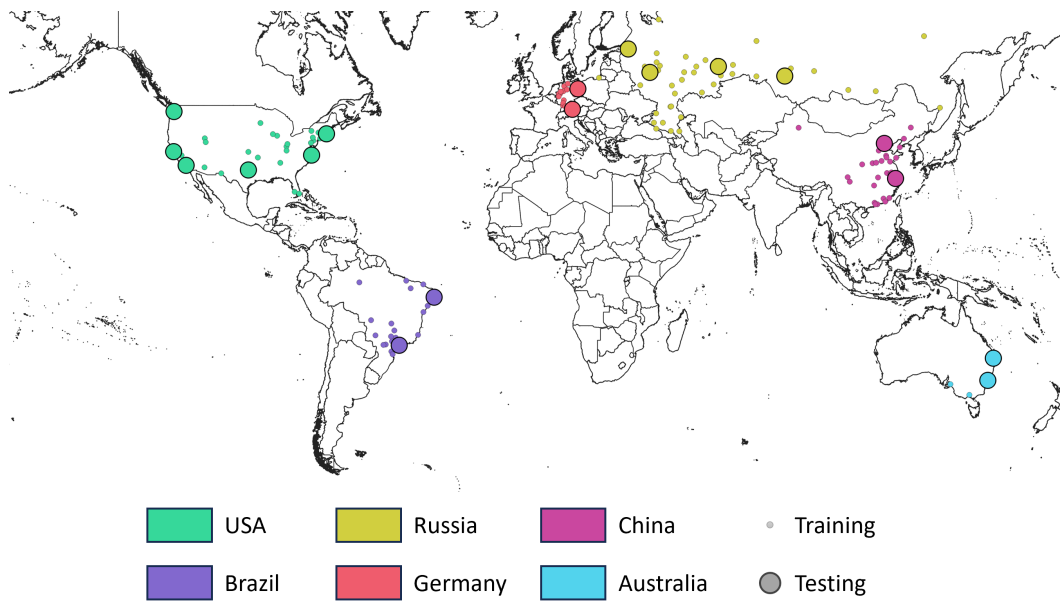


Fig. 1. Distribution of training and test cities.

sub-network to predict the correct labels. Furthermore, for unlabeled samples from the unsupervised domain adaptation they add a consistency loss between the SAR and optical sub-networks. They find data fusion to already improve model generalization to unseen test areas. This is improved further by the unsupervised domain adaptation where consistency between SAR and optical predictions is enforced. [34] study how adding OSM road network as an additional input layer affects the performance and transferability of an informal settlement classification model. They find improved transferability from adding auxiliary input. [12] combine remote sensing images and socioeconomic data, such as Points of Interest (POIs) and road network information to predict urban land-use classes in the Netherlands. They find adding socioeconomic data improves model transferability between different regions. The presented literature indicates adding further data modalities improves model performance and transferability. However, they do not study the effect of availability of the auxiliary data. This motivates for this study to use the building footprints as additional input and probe if this improves transferability.

III. METHODOLOGY

The following subsections provide a detailed introduction into the used dataset and the training and testing procedures.

A. Dataset

As presented in the previous section II-B OSM building tags can be regarded to be of good quality, which motivates for using them as training and testing data in the present work. We collected data from six different countries on five different continents, namely the United States of America, Brazil, Germany, Russia, People's Republic of China and Australia. The countries have been selected for their size and data availability. We selected two to six cities for testing for

each country and a variable number of cities for training (to reduce the imbalance of labeled pixels per country). The train and test cities are a subset of the 1692 cities used for a global study of the urban morphology [38]. Figure 1 shows the locations of the training and test cities, with the latter marked by bold circles.

The OSM building tags were aggregated to higher level classes residential/non-residential. Buildings without any functional label are treated as unlabeled. The aggregation scheme follows the one described in [26] where the keys *building*, *amenity* and *shop* are queried and mapped to a functional class. The mapping from OSM tags to functional classes was manually curated and Table III in the Appendix provides a full list of the mapping used throughout this work. Figure 2 shows exemplary data for four test cities with residential buildings in green, non-residential buildings in red, and unlabeled buildings in yellow. The differences in label completeness and city morphology can be well observed. Munich, for example, has many unlabeled buildings, whereas almost all buildings in Los Angeles have an associated type. The OSM data was rasterized to three meter resolution, paired with Planet Basemap data, and cut into non-overlapping patches of 450x450 pixels. During training, random crops of 320x320 pixels are cut from the patches. During validation and testing, a center crop is used. Slight global offsets between labels and data are corrected by edge detection followed by template matching, providing a correction offset. For this we used the algorithm used by [39]. The data has been visually assessed and grouped into four quality-level categories: *complete footprints*; *incomplete footprints*; *cloudy*; *inaccurate*. The *inaccurate* class includes annotations of whole street blocks or underground facilities. The *cloudy* class includes areas partially or fully covered by clouds. The *cloudy* and *inaccurate* patches are omitted for training and evaluation. For the *complete* and *incomplete* patches, only the labeled buildings *residential* or *non-*

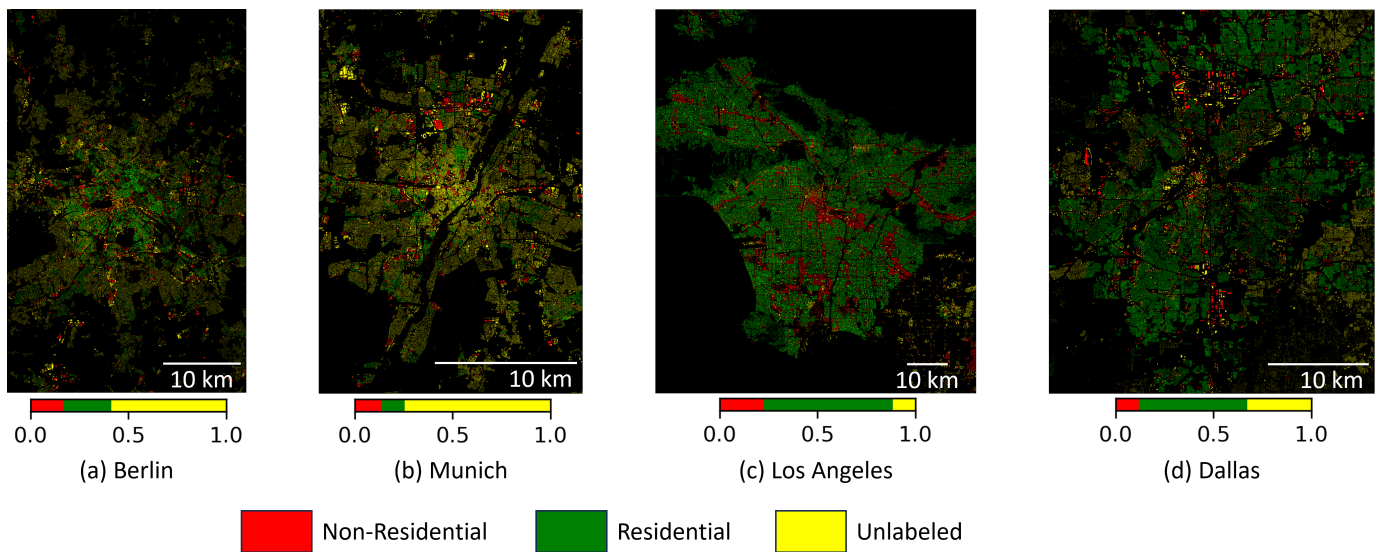


Fig. 2. Exemplary OSM rasters for different test cities. The scale varies between the subfigures.

residential are considered for training and testing. The rest is treated as unknown, mitigating the omission error due to missing buildings. Figure 3 illustrates samples of different quality. Figure 3(a) shows a completely annotated scene of a residential area with only a few buildings missing function annotation (yellow). Figure 3(b) shows another residential area where some buildings in the center and the upper part are not annotated. Figure 3(c) shows a cloudy scene with scarce building footprint and building type annotation. Figure 3(d) shows an inaccurately annotated non-residential area in the upper left. Removing cloud-occluded or inaccurately annotated areas from training and testing reduces noise. A model could only guess what is below a cloud or underground, and predicting a building function on inaccurately annotated grassland or streets would be meaningless. Completeness in the context of our data quality only refers to the completeness of the building footprints. It does not refer to the completeness or correctness of the building functions. As shown in Figure 2 and Figure 3 the completeness of function annotation varies between and within cities. This is in agreement with the literature on OSM quality [27], [28]. Figure 4 provides the statistical distribution of the residential and non-residential class for the six countries and the eighteen test cities. The numbers above the figure show the sum of residential and non-residential pixels. It can be seen that there is a distinct variation between the class distribution and pixel count, with the USA providing the the most training data and Brazil the least. For the test cities Los Angeles has most labeled pixels, whereas Novosibirsk has the least.

B. Training

We train an ImageNet pre-trained U-Net from the *Segmentation Models* library (version 0.3.0) [40] with a ResNet-34 backbone in PyTorch with Adam optimizer. We use the default values of the library with a depth of five and batchnorm. We start with a learning rate of $1e-4$ and reduce the learning rate by ten if the validation loss does not improve for five epochs. We use early stopping with a patience of twenty

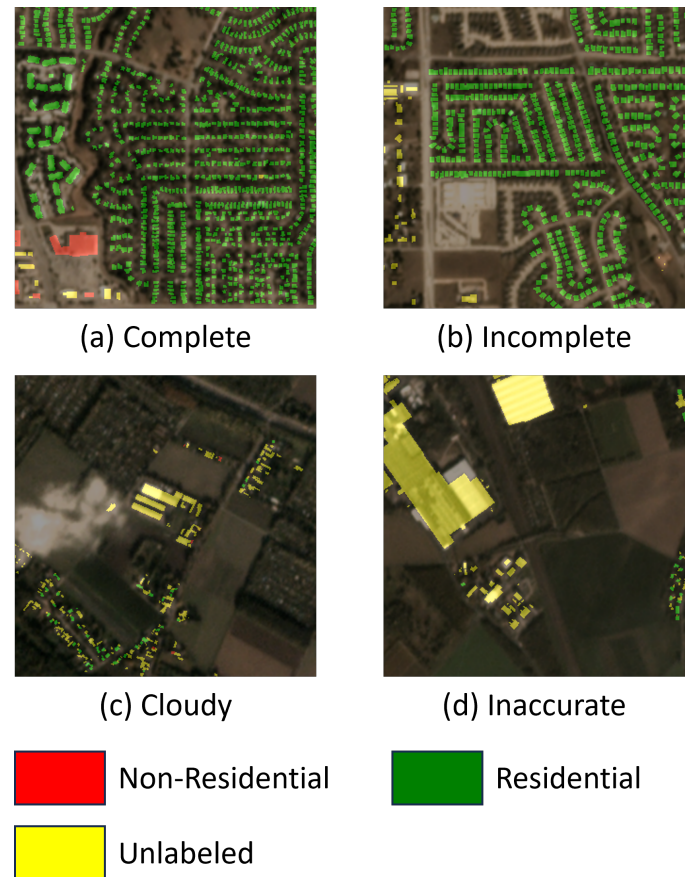


Fig. 3. Exemplary data patches of different quality. Overlay of Planet Basemap image and OSM rasters. Patch (a) and (b) are from the USA and (c) and (d) from Germany. Patch (b) has some unannotated buildings in the middle and on the upper part.



Fig. 4. Class distributions of the training and test datasets. The numbers above the figure show the total pixel count for residential and non-residential classes.

epochs and choose the model with the lowest validation loss. The training is terminated if the learning rate becomes lower than $1e-7$ or if the early stopping patience is reached. For each experiment we train five models, using a random five-fold training/validation split. We only train on the labeled (residential/non-residential) pixels and omit background and unlabeled pixels. We use normalized class-frequency reweighting to balance the two classes. Finally, we use Focal Loss for training to automatically adjust for variable difficulty within classes. For the ablation with PSPNet and DeepLabV3Plus in B the same settings as detailed above are used.

Figure 5a illustrates the approach. A U-Net is fed with a three channel satellite image and gives a dense prediction of building function (residential: green, non-residential: red). The prediction is masked with the labeled buildings and the loss is calculated only for these. Background (black) and unlabeled buildings (yellow) are ignored. This is the baseline for our generalization experiments in section IV-A1.

Figure 5b illustrates the training process for our data fusion experiment (section IV-A2). We stack building footprints as an additional input layer on top of the remote sensing images. We train the model to predict the building functions with an empty building footprint mask, an augmented mask (dropped buildings), and only the footprints. We aim to make the model invariant to omission errors in the building footprints, i.e., to correct for missing footprints. This is inspired by the work of

[37] who found SAR, optical data fusion to improve model generalization between regions. It can be interpreted as a multitask problem where the total loss is given by the losses of the three sub-tasks (see Equation 1). To cope with possibility of different task difficulties we applied dynamic weight averaging as introduced by [41] and shown in Equation 2. We used $T = 2$, which [41] found to be the optimum, and initialized the weights to one.

$$\mathcal{L}_{tot} = \mathcal{L}_{RS} + \mathcal{L}_{RS+AugFT} + \mathcal{L}_{FT} \quad (1)$$

$$\lambda_k(t) = \frac{\exp(w_k(t-1)/T)}{\sum_i \exp(w_i(t-1)/T)}, w_k(t-1) = \frac{\mathcal{L}_k(t-1)}{\mathcal{L}_k(t-2)} \quad (2)$$

Finally, we evaluate the model without and with the footprints. For comparison, we train one model only on the footprints (one input band).

C. Metrics

For evaluation, we use the F1 score on the building classes (after masking the unknown classes). Previous work [42] suggested building detection and classification can be evaluated separately, allowing the use of the OSM data of variable completeness, i.e., missing buildings would significantly impact precision and recall of the detection task but can be omitted for

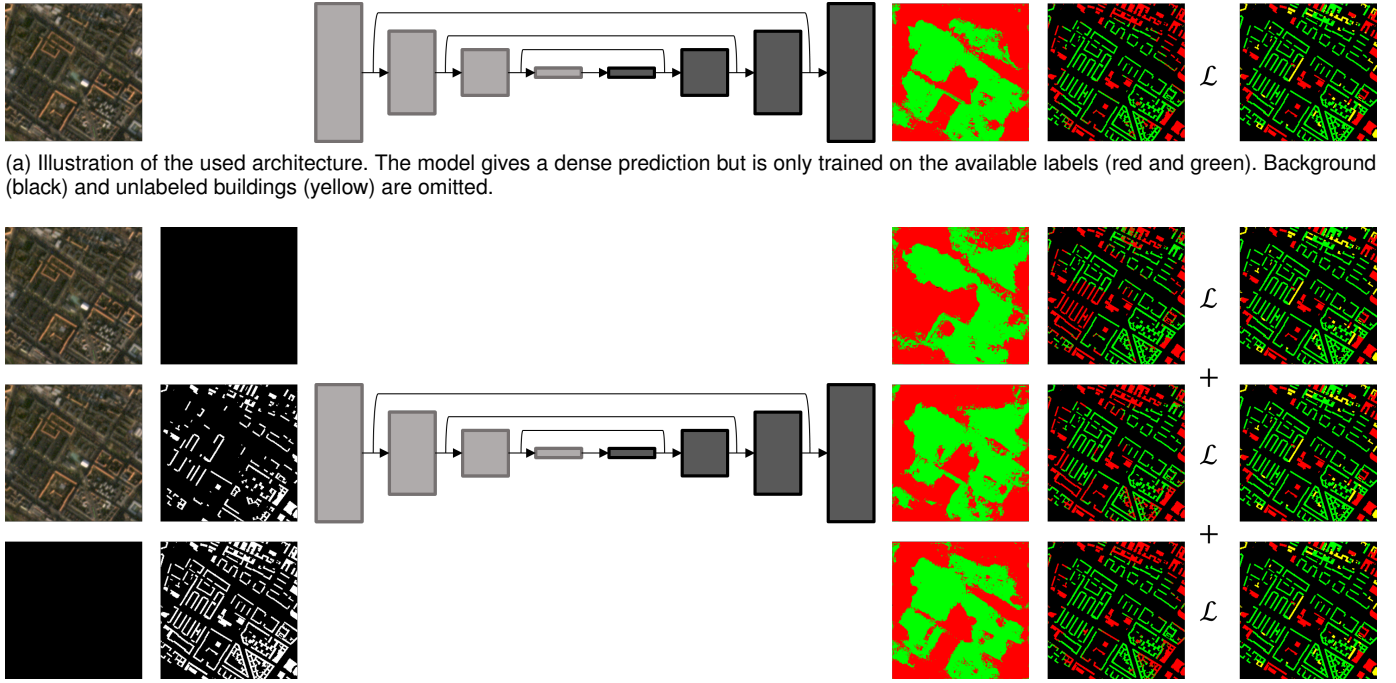


Fig. 5. Illustration of training approach. The model is given a stack of remote-sensing images and footprint masks. The model learns to predict with empty footprints, augmented footprints (dropped), and footprints only (zeroed satellite image).

the classification evaluation. We report the mean and standard deviation of the F1 score averaged over the five models. For Figure 7 we report median, minimum and maximum F1 score of the five models.

IV. EXPERIMENT & RESULT

The following section introduces the experiments and provides their results.

A. Experiments

1) *Generalization*: We conduct four experiments to compare how well a model generalizes to its training region, a model generalizes to an unseen region, pre-training affects performance and how a general global model compares against specialized local ones. This means:

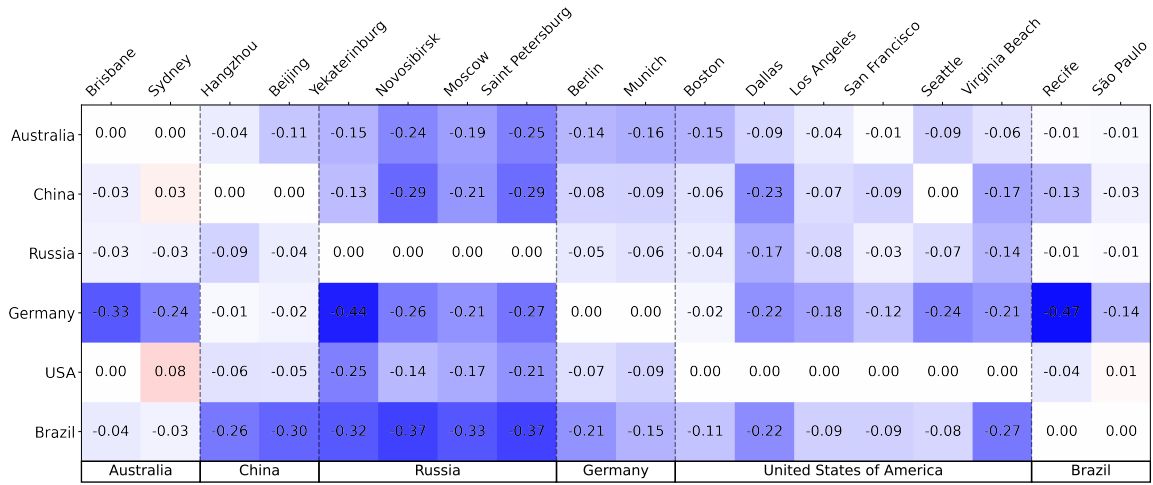
- 1) Train models on *local* data; i.e., data from the same country.
- 2) Train models *non-local* data; i.e., omit data from the same country.
- 3) Fine-tune *non-local* models on *local* data.
- 4) Train models on *global* data; i.e., combined *local* data.

2) *Footprint Fusion*: In a further experiment, we stack building footprints as an additional input layer on top of the remote sensing images. As presented in section III-B, we train the model to predict the building functions with an empty building footprint mask, an augmented mask (dropped buildings), and only the footprints. We aim to make the model invariant to omission errors in the building footprints, i.e., to correct for missing footprints. We study if this improves F1 score and model transferability. Therefore, we compare

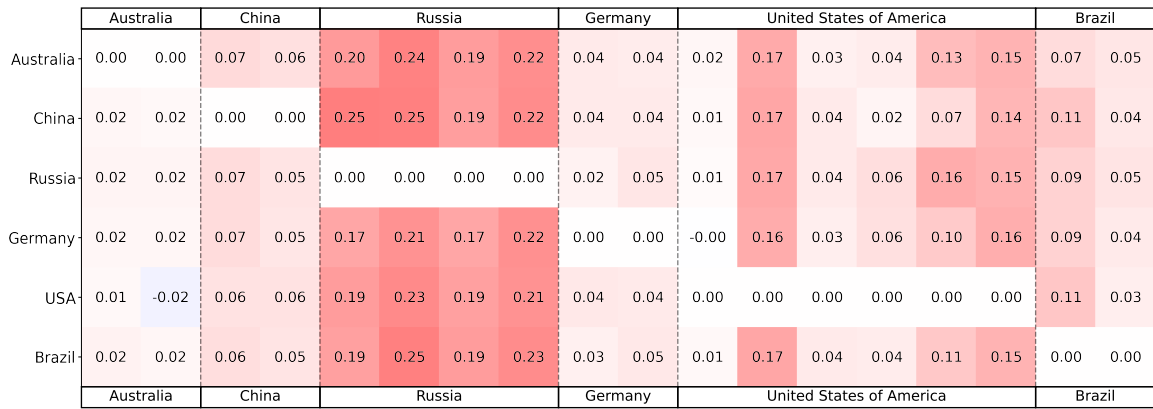
a *global* model against *non-local* models when providing empty building footprints or all available footprints. We further compare these to a baseline model trained only on building footprints.

B. Results

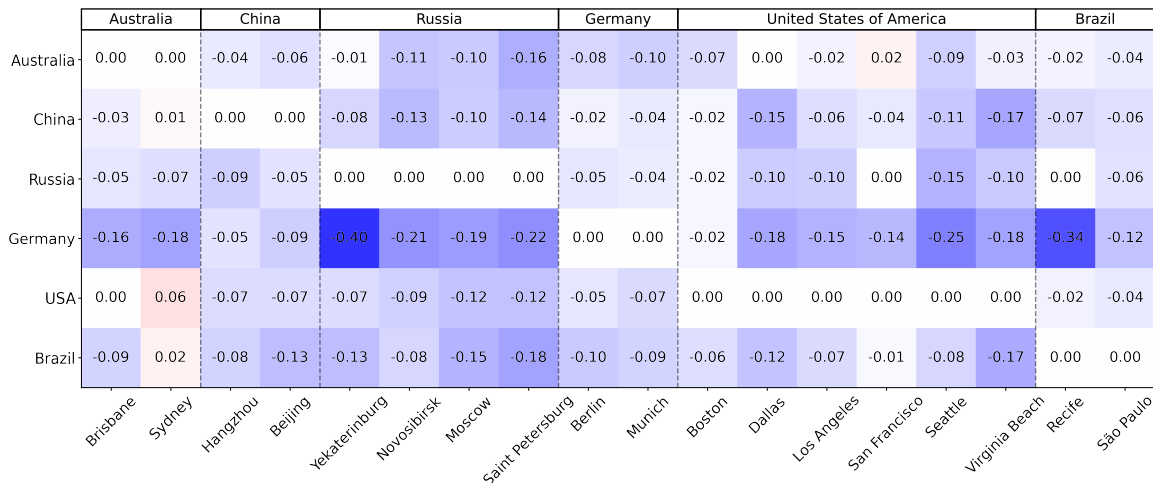
1) *Generalization*: Table I shows the results for our four training strategies. The columns represent the different training strategies, and the rows correspond to the different test cities. The horizontal bars separate test cities from different countries. For the first three columns, the horizontal bars mark different models, i.e., we have different *local*, *fine-tune*, and *non-local* models for each country. The last column refers to the performance of the *global* models evaluated on all countries, as shown by the low F1 scores of the *non-local* models in comparison with the other models. We generally find models to perform best in the country they were trained in and often to not generalize well to other countries. Figure 6a shows the relative change of F1 score from models trained in the respective country compared to models trained in other countries. This change is negative for most combinations, i.e., the *local* model performs best (blue shades in the figure). There are some combinations where models trained on local data from another country can achieve comparable performances to models trained only on data from the test country (light blue to white shades in the figure). For Sydney, the model trained on Chinese or US data even outperforms the model trained in Australia (red shades in the figure). For the *non-local* training, we observe that models omitting a country's training data generally perform worse compared to models including the data (see Figure 6b). However, the severity of



(a) $(F1_{local_column} - F1_{local_row})/F1_{local_column}$ Local models compared to local models from test city's country. Row indicates the training country. The overall blue color indicates models to not transfer well to other countries.



(b) $(F1_{non-local_column} - F1_{non-local_row})/F1_{non-local_column}$ Non-local models compared to non-local models from test city's country. Row indicates the omitted training country. The overall red color indicates models perform worse when data from test country is missing.



(c) $(F1_{fine-tune_column} - F1_{fine-tune_row})/F1_{fine-tune_column}$ Fine-tuned models compared to fine-tuned models from test city's country. Row indicates the fine-tuning country. The overall blue color indicates models perform worse when fine-tuned on a different country.

Fig. 6. Relative change of F1 score of models trained on different training data. Blue indicates a reduction, and red indicates an increase.

TABLE I
MEAN AND STANDARD DEVIATION OF BUILDING F1 SCORE OF LOCAL, FINE-TUNED, NON-LOCAL AND GLOBAL MODELS AVERAGED OVER FIVE REPEATED TRAININGS. MODELS TRAINED ONLY ON REMOTE SENSING IMAGES.

	RS Only			
	Local	Fine-Tune	Non-Local	Global
Brisbane	0.89±0.01	0.90±0.01	0.89±0.01	0.91±0.00
Sydney	0.74±0.01	0.75±0.02	0.79±0.01	0.80±0.00
Hangzhou	0.78±0.01	0.81±0.02	0.77±0.01	0.82±0.01
Beijing	0.80±0.01	0.83±0.01	0.79±0.00	0.83±0.00
Yekaterinburg	0.78±0.04	0.78±0.03	0.69±0.03	0.81±0.02
Novosibirsk	0.84±0.01	0.84±0.01	0.68±0.04	0.84±0.00
Moscow	0.84±0.01	0.85±0.00	0.71±0.01	0.84±0.00
Saint Petersburg	0.83±0.00	0.83±0.00	0.68±0.01	0.83±0.01
Berlin	0.83±0.01	0.83±0.01	0.81±0.02	0.84±0.01
Munich	0.84±0.01	0.84±0.00	0.80±0.01	0.84±0.00
Boston	0.75±0.01	0.75±0.01	0.75±0.00	0.77±0.01
Dallas	0.93±0.00	0.93±0.01	0.80±0.06	0.94±0.00
Los Angeles	0.87±0.01	0.87±0.01	0.83±0.02	0.86±0.01
San Francisco	0.77±0.02	0.76±0.02	0.75±0.03	0.78±0.02
Seattle	0.66±0.03	0.68±0.02	0.58±0.01	0.64±0.02
Virginia Beach	0.93±0.01	0.94±0.01	0.80±0.05	0.91±0.01
Recife	0.79±0.02	0.78±0.01	0.74±0.02	0.80±0.01
São Paulo	0.79±0.01	0.84±0.01	0.80±0.01	0.85±0.01

this penalty varies between different cities and countries. We observe a substantial effect for Russia but an almost negligible impact for Australia. Furthermore, for many cases, the effect stays nearly constant over different dataset combinations, i.e., it appears essential that data from the test country is in the training data, but it is of lesser importance what other data is included in the training data. There are some exceptions, such as Yekaterinburg, San Francisco or Seattle, which show more variability across different data splits. Only for Sydney we observe the *non-local* model to be better than the *local* one.

For the *fine-tuned* models we observe that they tend to be as good or slightly better than the *local* models (see Table I). For Hangzhou, Beijing, and São Paulo they are distinctly better. For Sydney, the *fine-tuned* models have a lower F1 score than the *non-local* models. Figure 6c shows the performance penalty of *fine-tuned* models evaluated outside the country they were tuned on. For many test cities outside the fine-tuning country a reduced performance is observed. Comparing with Figure 6a from the *local* training a less pronounced penalty is observed.

Training on the combination of all countries, we observe that the models are often on par or even slightly outperform the *local*, or *fine-tuned* models (see Table I). For all Brazilian, Chinese, German, and Brazilian test cities, the *global* models achieve the best average performance. For Moscow, Los Angeles, Seattle, and Virginia Beach, the *local* or *fine-tuned* models outperform the *global* models. However, the differences between *local*, *fine-tune* and *global* models are often well within the standard deviation.

2) *Footprint Fusion*: Figure 7 and Table II show the results from the experiment on footprint fusion. Figure 7a shows the results from probing the model with empty footprints and Figure 7b shows the results from probing with footprints. In

Figure 7a we omit the performance of the model purely trained with footprints, since it behaves as a random guesser and the F1 score is outside the depicted bounds (the model receives no informative input signal and can only randomly predict the building function). Table II only shows the results for global models and reproduces the global models' performance of the remote sensing only experiments (see table I).

Figure 7 shows a distinct variance among different cities and a clear gap between the non-local and the global models for many cities. For almost all models, the median F1 score when probing with footprints is higher than probing the models with empty footprints. Only the *global* models evaluated on Recife and São Paulo perform worse when presented with footprints compared to no footprints. Furthermore, the gap between *non-local* and *global* models is reduced for most cities, and for some, it appears to be closed. When probing with empty footprints the models' performances are approximately equal to the models trained only on remote sensing images (see Table II). The *footprint-only* model is often on par with the *global* model. For Yekaterinburg, it is even outperforming the *global* model, whereas, for San Francisco, Virginia Beach, and São Paulo, it is performing worse than the *global* model.

Figure 8 and Figure 9 in B show the results from repeating the experiments with a PSPNet and DeepLabV3Plus. The results appear to be similar to the ones described above for the U-Net.

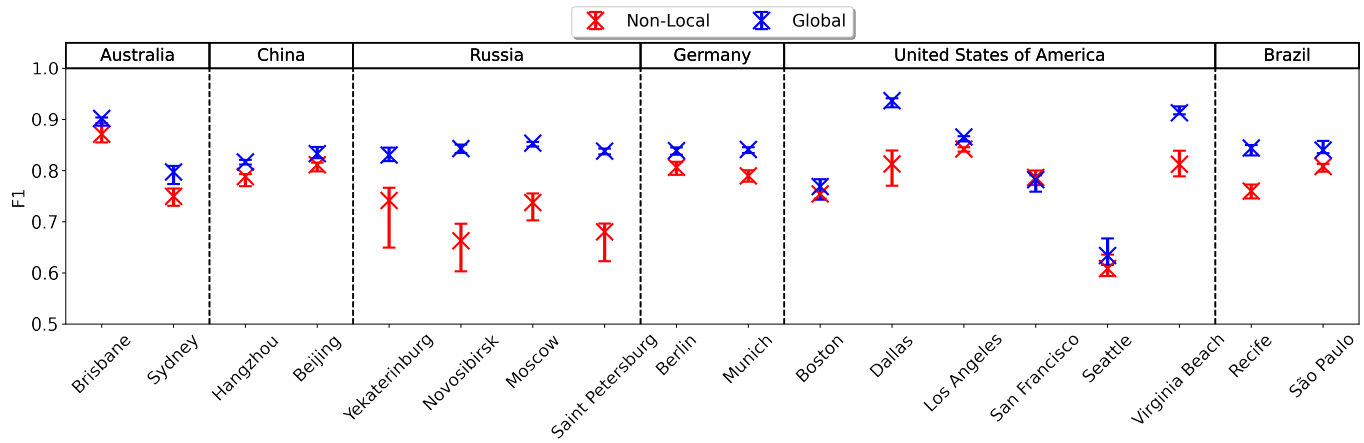
TABLE II
MEAN AND STANDARD DEVIATION OF BUILDING F1 SCORE GLOBAL MODELS AVERAGED OVER FIVE REPEATED TRAININGS. MODELS TRAINED ONLY ON REMOTE SENSING IMAGES (RS), ON BOTH, REMOTE SENSING IMAGES AND BUILDING FOOTPRINTS (RS + FOOTPRINT), OR ONLY ON FOOTPRINTS (FOOTPRINT). THE MODELS TRAINED WITH BOTH (IMAGES AND FOOTPRINTS) ARE EVALUATED WITH EMPTY FOOTPRINTS (MASKED TO ZERO) OR ALL AVAILABLE FOOTPRINTS.

	RS	RS + Footprint		Footprint
		Empty FT	FT	
Brisbane	0.91±0.00	0.90±0.01	0.93±0.00	0.93±0.00
Sydney	0.80±0.00	0.79±0.01	0.83±0.01	0.81±0.01
Hangzhou	0.82±0.01	0.82±0.00	0.87±0.01	0.85±0.01
Beijing	0.83±0.00	0.83±0.01	0.86±0.01	0.85±0.01
Yekaterinburg	0.81±0.02	0.83±0.01	0.89±0.00	0.91±0.00
Novosibirsk	0.84±0.00	0.84±0.01	0.86±0.01	0.87±0.01
Moscow	0.84±0.00	0.85±0.00	0.88±0.00	0.87±0.00
Saint Petersburg	0.83±0.01	0.84±0.00	0.87±0.00	0.86±0.00
Berlin	0.84±0.01	0.84±0.01	0.87±0.01	0.87±0.01
Munich	0.84±0.00	0.84±0.00	0.89±0.00	0.88±0.00
Boston	0.77±0.01	0.77±0.01	0.80±0.01	0.79±0.00
Dallas	0.94±0.00	0.93±0.01	0.97±0.00	0.97±0.00
Los Angeles	0.86±0.01	0.86±0.00	0.94±0.00	0.93±0.00
San Francisco	0.78±0.02	0.78±0.01	0.84±0.01	0.77±0.01
Seattle	0.64±0.02	0.64±0.02	0.71±0.03	0.73±0.00
Virginia Beach	0.91±0.01	0.92±0.01	0.97±0.00	0.94±0.00
Recife	0.80±0.01	0.84±0.01	0.81±0.01	0.79±0.01
São Paulo	0.85±0.01	0.84±0.01	0.81±0.00	0.77±0.00

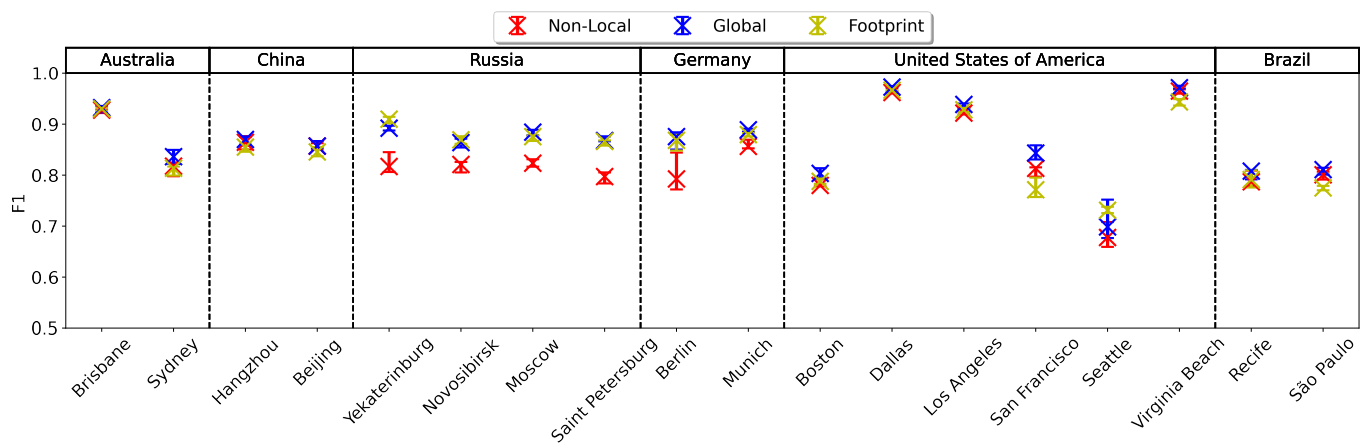
V. DISCUSSION

A. Generalizability

The experiment results indicate a poor model generalizability between countries. When evaluated outside their train-



(a) Models tested with empty (zero mask) building footprints. *Non-Local* models are trained omitting data from the test country. *Global* models are trained with data from all countries.



(b) Models tested with all available building footprints. *Non-Local* models are trained omitting data from the test country. *Global* models are trained with data from all countries. *Footprint* models are trained with data from all countries, but only using building footprints as input.

Fig. 7. F1 score of models evaluated on the different test cities. The cross marks the median value of five models and ticks the best and worst-performing models. Dashed vertical lines mark different test countries.

ing areas, the *local* and *fine-tuned* models show a distinct performance drop. This is generally confirmed by the *non-local* models, which show a performance drop in the left-out country. Only for a few combinations (e.g., USA and Sydney) a positive correlation between different countries is observed. The models trained on data from the USA outperform models trained in Australia when evaluated on Sydney. This could be due to similarities in housing patterns between the USA and Australia but could also be an artifact of data availability. As shown in Figure 4 the dataset contains more training data from the USA than from Australia. We observe the most substantial impact of *non-local* versus *local* training for the four Russian cities, Dallas and Virginia Beach. The reason for this can be manifold. One possible reason could be some distinct housing patterns in the respective cities and countries. Another reason could be biases in OSM annotations, where individuals or groups label only buildings of a certain type or use the OSM tags in a different way.

We choose to treat single countries as separate domains. This is motivated by the fact that many countries have national building regulations and styles. We acknowledge that there

is variability within countries, and there may be similarities between countries. The variability is addressed by sampling diverse training and test cities. Probing for positive correlations between single countries and test cities was one of the study's objectives. We found only a weak correlation between different countries, with the most distinct between the USA and Sydney. However, secondary effects, such as dataset size, may affect the results. While observing a poor generalizability, we observe that the specialized (*local* and *fine-tuned*) models often do not outperform the models trained on the combined dataset. This indicates that instead of training multiple local models on separate data, it may be as good to combine the data and train only one single model. We put great effort into sampling the dataset, yet, due to data limitations we miss an African country in the dataset. Furthermore, we focus on large countries and it is not clear how well the findings transfer to smaller countries and countries not covered in the dataset.

B. Footprint Fusion

We studied the effect of providing building footprints as an additional input layer on the prediction performance and

model generalization. This is motivated by footprints being a high-level abstraction of the built-up environment. A common approach is to map building functions by first extracting morphological features, such as size, shape, and distance to other buildings or streets, and feeding these features to classifiers, such as random forests [10], [28]. We hypothesize that a deep neural network can implicitly learn several of these features. Providing the known building footprints directly to the model may allow it to focus more on these morphological descriptors of the buildings. We found providing the building footprints as inputs improves the F1 score in all cases except for the two Brazilian test cities (see Table II). The reason for this is unknown but may be related to poor footprint quality, as these are also the only cities where the footprint-only model performs worse than the pure remote sensing model. Interestingly, [34] found a similar anomaly when adding OSM street network as an additional input layer to study informal settlement mapping. They found performance improvements except for Rio de Janeiro and São Paulo, which they attributed to ordered morphology of informal settlements, significantly deviating from structures in other cities of their dataset. Furthermore, the gap between *non-local* and *global* models appears to be reduced for most cities when providing footprints. For some cities, the models performances become even indistinguishable. This suggests the provision of additional input data can help the models to transfer between different areas, as was described by others [12], [34], [37]

The models trained and tested just on building footprints achieve top performance for many cities. In Yekaterinburg they perform better than the stacked models, whereas in Sydney, Boston, San Francisco, Virginia Beach, and São Paulo they perform slightly worse. This suggests that building footprints alone provide a strong input feature to derive high-level functional classes, which are used in this study. This is in line with other works, which solely used OSM features to predict building functions [28]. The reduced performance of the footprint-only models in some cities may be associated with the varying quality and completeness of the footprints. While we performed manual data cleaning to remove obvious cloudy and coarsely annotated patches from training and testing data, we used both complete and incomplete building footprints as inputs.

C. OSM quality

This study uses OSM data for model training and testing and assumes labeled buildings are likely to be annotated correctly. This assumption is supported by related literature [27], [28], [42], which found significant variance in the completeness of building function annotation in OSM, but the generally high quality of available annotations. We manually cleaned the data to remove obvious annotation errors and improve the dataset quality (see section III-A). Nevertheless, comparison with other datasets, such as the Global Human Settlement Layer non-residential dataset [17] could provide further insights into the quality of the OSM labels and help identify regions with lower labelling accuracy. The incompleteness of OSM labels is dealt with by training and testing only on the labeled

buildings. Possible inaccuracies in the OSM labels may lead to over- or underestimating the model performance in individual cities. Nevertheless, the study focuses on a relative comparison between models trained on different data combinations and not on absolute performance, which should be less sensitive to the data quality. The findings in Table I suggest a *global* model to be often on par with a *local* model, i.e., both models to be equally good. The pronounced differences between different cities may be associated to different difficulty of the cities, but could also be related to annotation errors in the OSM labels.

D. Varying difficulty

We observe a relatively high variability of the model performance between different cities. This can be partly explained by the variable complexity of the given task. Classifying a homogeneous settlement structure or unambiguous buildings is easier than classifying a heterogeneous structure. Furthermore, volunteers in a given region may be more likely to annotate buildings of a specific type. We only selected test areas with a certain number of annotated buildings and pixels to mitigate the effects of annotation bias. Furthermore, focusing on relative comparison instead of absolute performance should mitigate possible selection effects.

E. Limitations

Building functions are not directly observable from remote sensing images. Instead, they can be inferred from auxiliary information, such as building morphology and the surrounding buildings [13], [14]. Yet, this information may not be sufficient in an ambiguous heterogeneous area, such as dense city centers. Furthermore, the used classification scheme's granularity may influence the feasibility of inferring functions. In this study, we focused on a coarse *residential/non-residential* classification scheme and used the F1 score for evaluation. The binary classification scheme adds already valuable information to the widely available building footprint products but avoids ambiguities of finer-grained classification schemes [1], [28], [32]. The observed findings are also observed when considering precision and recall instead of the F1 score (see appendix C). Nevertheless, more granular evaluation metrics and classifications may be desirable, but may not be feasible with the data used throughout this study.

VI. CONCLUSION

This work studied the geographic transferability U-Nets trained for building type classification. It was found that models do not transfer well to countries, not included in the training data. Training models on combined data from multiple countries yields no penalty compared to training models for each country separately. This suggests it is advantageous to train only one model on the combination of the data, compared to training multiple local models. Furthermore, it was found that stacking building footprints on-top of the remote sensing images improves both, the model's performance and the geographic transferability. It was found that for many test cities this could reduce or even close the gap between non-local and global models. While a model using only building

footprints often performs as good as a model using footprints and remote sensing images, it collapses to a random guesser in the case of missing building footprints. The combined one, however, only collapses to a model solely trained on remote sensing images. Therefore, fusing footprints and remote sensing images improves the model's transferability, while making it more robust to noise in the footprint data. This suggests the approach may help predicting building functions in label scarce areas. Further studies could investigate how a finer classification scheme affects the results and how models fine-tuned from foundation models behave in the given setting.

FUNDING

The work was funded through European Union's Horizon 2020 research and innovation programme within the DAPHNE project under grant agreement No. 957407. This work was supported by the Helmholtz Association's Initiative and Networking Fund on the HAICORE@FZJ partition.

DATA AVAILABILITY

The satellite imagery is commercial and cannot be shared. The OSM labels and code can be shared upon reasonable request.

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.

CREDIT AUTHOR STATEMENT

Nikolai Skuppim: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing - Original Draft, Writing - Review & Editing **Yilei Shi:** Conceptualization, Supervision, Writing - Review & Editing **Xiao Xiang Zhu:** Conceptualization, Supervision, Writing - Review & Editing, Funding acquisition

ACKNOWLEDGMENT

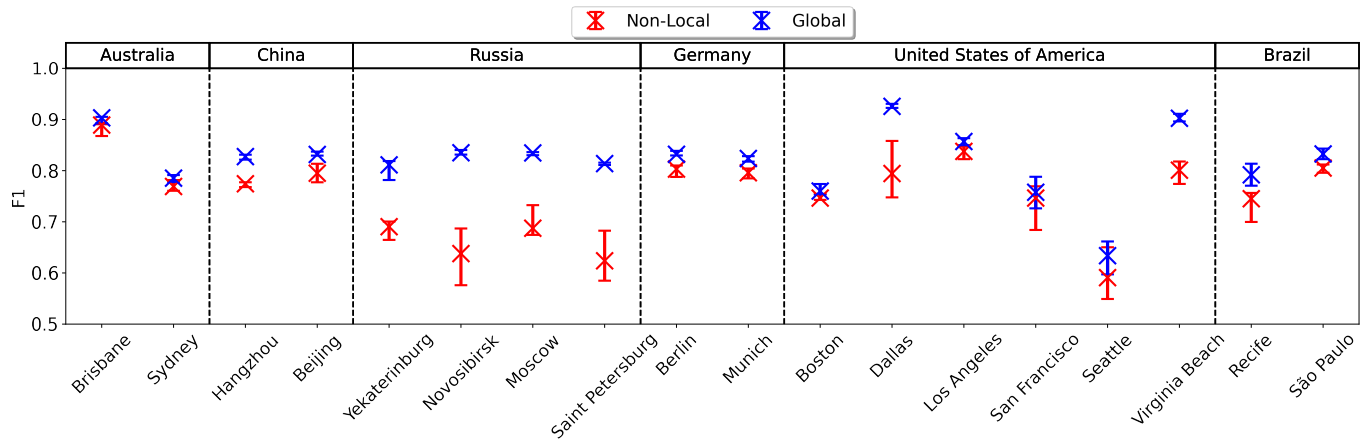
The authors thank Eike Jens Hoffmann, Conrad Albrecht and Yuanyuan Wang for their help in collecting and processing the raw data.

APPENDIX A OPENSTREETMAP MAPPING

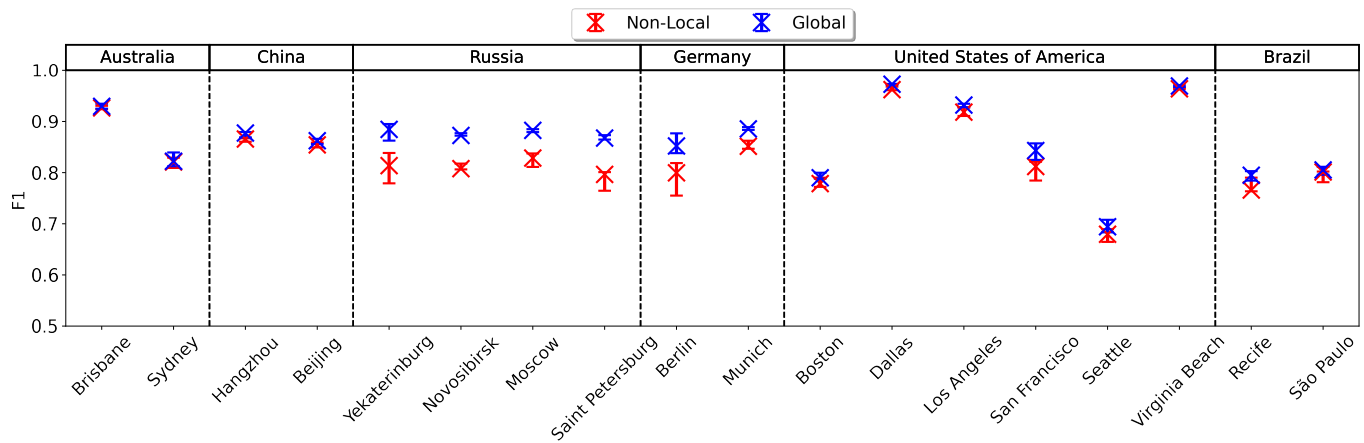
Table III provides the mappings of OpenStreetMap entries to the functional classes.

TABLE III
MAPPING FROM OPENSTREETMAP KEY, VALUE PAIRS TO BUILDING FUNCTION CLASSES, USED IN THE STUDY.

class	key	value
residential	building	<i>apartments, bungalow, cabin, detached, dormitory, farm, house, houseboat, residential, static_caravan</i>
non-residential	building	<i>bakehouse, barn, cathedral, chapel, church, civic, commercial, conservatory, cowshed, farm_auxiliary, government, grandstand, greenhouse, hangar, hospital, hotel, industrial, kindergarten, kiosk, mosque, office, pavilion, public, religious, retail, riding_hall, school, shrine, sports_hall, stable, stadium, sty, supermarket, synagogue, temple, toilets, transportation, university, warehouse</i>
non-residential	amenity	<i>animal_boarding, animal_shelter, arts_centre, bank, bicycle_rental, biergarten, boat_rental, brothel, bureau_de_change, cafe, car_rental, car_wash, casino, charging_station, childcare, cinema, clinic, college, community_centre, courthouse, crematorium, dentist, dive_centre, doctors, driving_school, embassy, fast_food, ferry_terminal, fire_station, food_court, fuel, gambling, gym, hospital, ice_cream, internet_cafe, kindergarten, language_school, library, marketplace, monastery, music_school, nightclub, nursing_home, pharmacy, place_of_worship, planetarium, police, post_depot, post_office, prison, pub, public_building, ranger_station, restaurant, sanitary_dump_station, school, shop, social_centre, social_facility, stripclub, studio, swingerclub, theatre, townhall, university, vehicle_inspection, veterinary</i>
non-residential	shop	<i>agrarian, anime, antiques, appliance, art, atm, baby_goods, bag, bakery, bathroom_furnishing, beauty, bed, beverages, bicycle, boat, bookmaker, books, boutique, brewing_supplies, butcher, camera, candles, cannabis, car, car_parts, car_repair, caravan, carpet, charity, cheese, chemist, chocolate, clothes, coffee, collector, computer, confectionery, convenience, copyshop, cosmetics, craft, curtain, dairy, deli, department_store, doityourself, doors, drugstore, dry_cleaning, e-cigarette, electrical, electronics, energy, erotic, fabric, farm, fashion, fashion_accessories, fireplace, fishing, flooring, florist, frame, free_flying, frozen_food, fuel, funeral_directors, furniture, games, garden_centre, garden_furniture, gas, general, gift, glazery, golf, greengrocer, hairdresser, hairdresser_supply, hardware, health_food, hearing_aids, herbalist, hifi, houseware, hunting, ice_cream, interior_decoration, jetski, jewelry, kiosk, kitchen, lamps, laundry, leather, locksmith, lottery, mall, massage, medical_supply, mobile_phone, model, money_lender, motorcycle, music, musical_instrument, newsagent, nutrition_supplements, optician, organic, outdoor, paint, party, pasta, pastry, pawnbroker, perfumery, pest_control, pet, pet_grooming, photo, pyrotechnics, radiotechnics, religion, robot, scuba_diving, seafood, second_hand, security, sewing, shoes, ski, snowmobile, spices, sports, stationery, storage_rental, street_vendor, supermarket, swimming_pool, tailor, tattoo, tea, ticket, tiles, tobacco, toys, trade, trailer, travel_agency, trophy, tyres, vacant, vacuum_cleaner, variety_store, video, video_games, watches, water, weapons, wholesale, window_blind, windows, wine</i>
unlabeled	building	<i>bridge, bunker, carport, construction, digester, garage, garages, hut, parking, roof, ruins, service, shed, terrace, train_station, transformer_tower, user_defined, water_tower, yes</i>
unlabeled	amenity	<i>atm, baby_hatch, baking_oven, bbq, bench, bicycle_parking, bicycle_repair_station, boat_sharing, bus_station, car_sharing, clock, drinking_water, firepit, fountain, grave_yard, grit_bin, hunting_stand, kitchen, kneipp_water_cure, motorcycle_parking, parking, parking_entrance, parking_space, photo_booth, post_box, public_bath, public_bookcase, recycling, sauna, shelter, shower, taxi, telephone, toilets, user_defined, vending_machine, waste_basket, waste_disposal, waste_transfer_station, water_point, watering_place</i>



(a) Models tested with empty (zero mask) building footprints. *Non-Local* models are trained omitting data from the test country. *Global* models are trained with data from all countries.

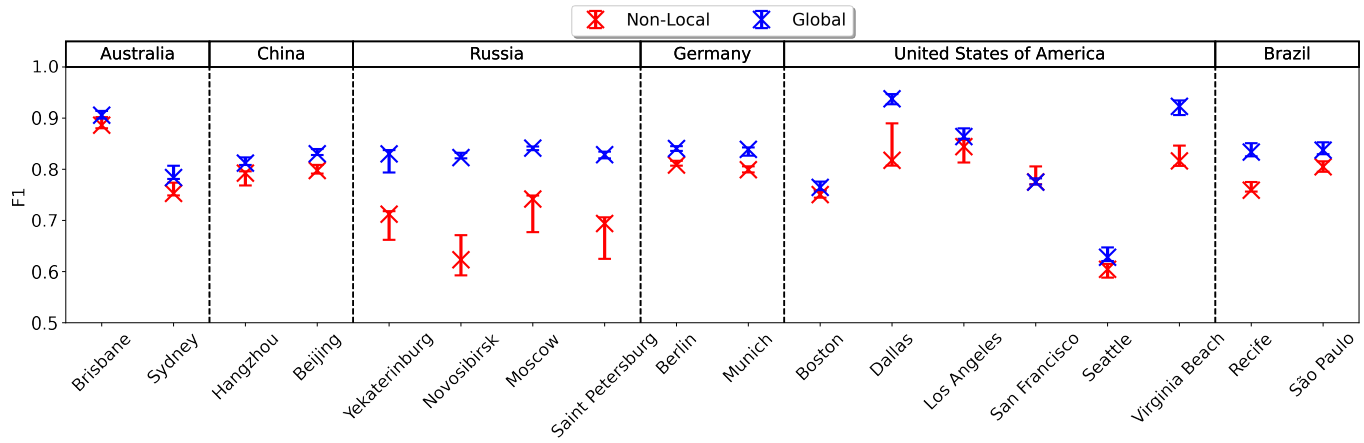


(b) Models tested with all available building footprints. *Non-Local* models are trained omitting data from the test country. *Global* models are trained with data from all countries.

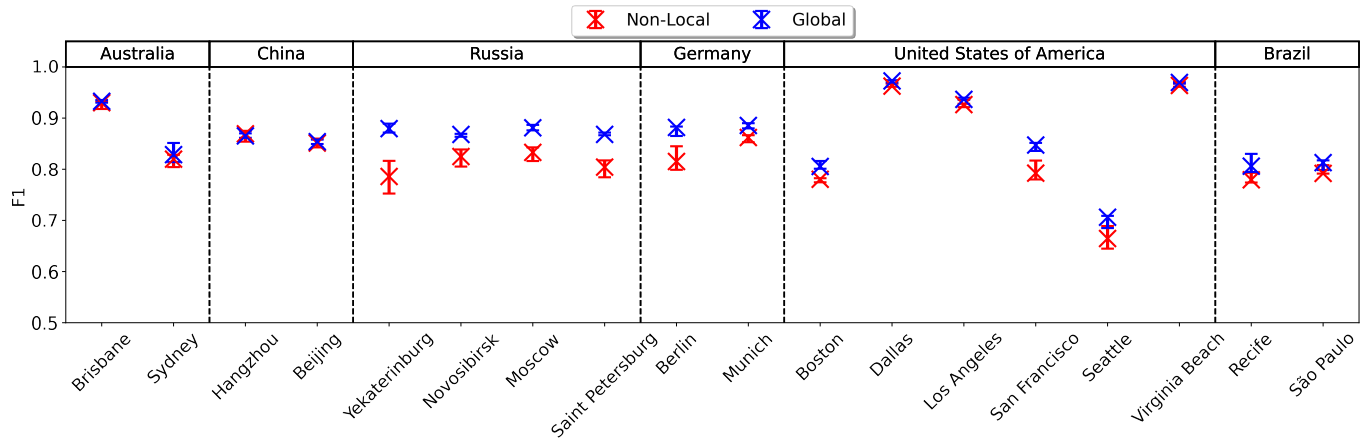
Fig. 8. F1 score of models evaluated on the different test cities for PSPNet. The cross marks the median value of five models and ticks the best and worst-performing models. Dashed vertical lines mark different test countries.

APPENDIX B MODEL ABLATION

Figure 8 shows results for PSPNet and figure 9 for DeepLabV3Plus.

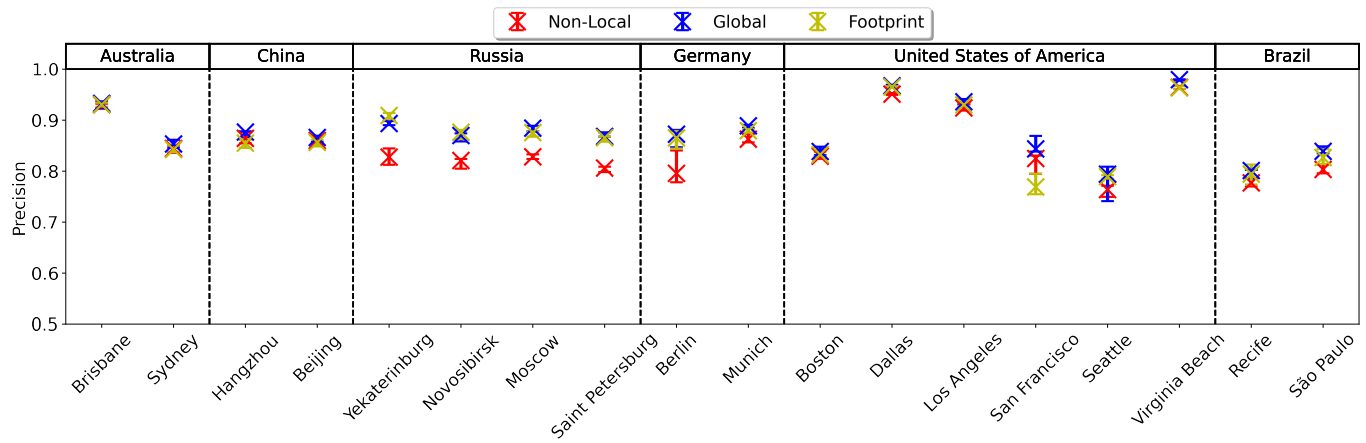


(a) Models tested with empty (zero mask) building footprints. *Non-Local* models are trained omitting data from the test country. *Global* models are trained with data from all countries.

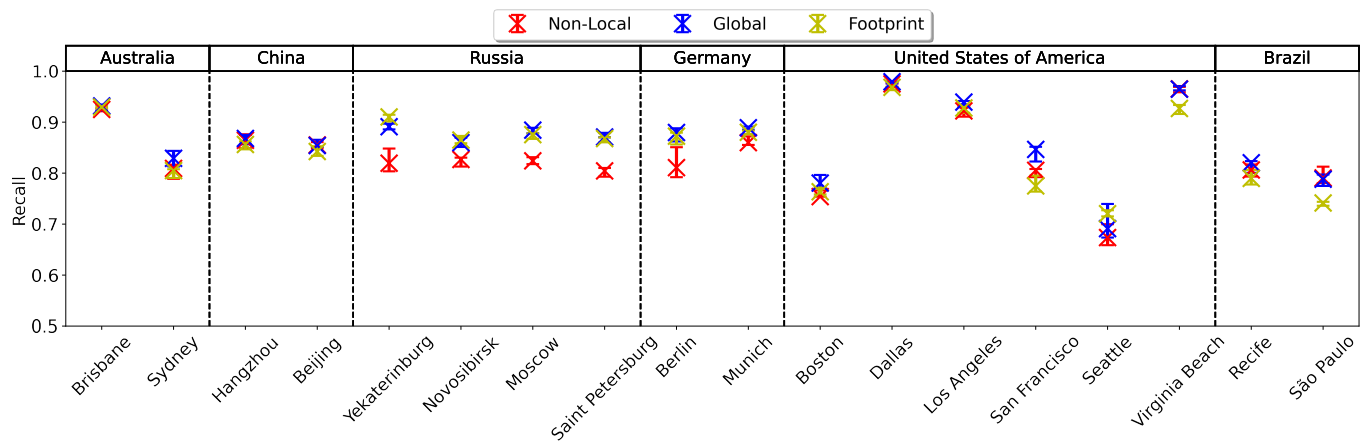


(b) Models tested with all available building footprints. *Non-Local* models are trained omitting data from the test country. *Global* models are trained with data from all countries.

Fig. 9. F1 score of models evaluated on the different test cities for DeepLabV3Plus. The cross marks the median value of five models and ticks the best and worst-performing models. Dashed vertical lines mark different test countries.



(a) Models tested with all available building footprints. *Non-Local* models are trained omitting data from the test country. *Global* models are trained with data from all countries.



(b) Models tested with all available building footprints. *Non-Local* models are trained omitting data from the test country. *Global* models are trained with data from all countries.

Fig. 10. Precision and Recall of models evaluated on the different test cities. The cross marks the median value of five models and ticks the best and worst-performing models. Dashed vertical lines mark different test countries.

APPENDIX C METRIC ABLATION

Figure 10a shows results for Precision and figure 10b for Recall.

REFERENCES

- [1] N. Milojevic-Dupont, F. Wagner, F. Nachtigall, J. Hu, G. B. Brüser, M. Zumwald, F. Biljecki, N. Heeren, L. H. Kaack, P.-P. Pichler, and F. Creutzig, "EUBUCCO v0.1: European building stock characteristics in a common and open database for 200+ million individual buildings," *Sci Data*, vol. 10, no. 1, p. 147, Mar. 2023. [Online]. Available: <https://www.nature.com/articles/s41597-023-02040-2>
- [2] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. Beijing China: ACM, Aug. 2012, pp. 1–8. [Online]. Available: <https://dl.acm.org/doi/10.1145/2346496.2346498>
- [3] V. Frias-Martinez and E. Frias-Martinez, "Spectral clustering for sensing urban land use using Twitter activity," *Eng. Appl. Artif. Intell.*, vol. 35, pp. 237–245, Oct. 2014. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0952197614001419>
- [4] M. Häberle, E. J. Hoffmann, and X. X. Zhu, "Can linguistic features extracted from geo-referenced tweets help building function classification in remote sensing?" *ISPRS J. Photogramm. Remote Sens.*, vol. 188, pp. 255–268, Jun. 2022. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0924271622001058>
- [5] C. C. Fonte, M. Minghini, V. Antoniou, J. Patriarca, and L. See, "CLASSIFICATION OF BUILDING FUNCTION USING AVAILABLE SOURCES OF VGI," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XLII-4, pp. 209–215, Sep. 2018. [Online]. Available: <https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-4/209/2018/>
- [6] A. Bandam, E. Busari, C. Syranidou, J. Linssen, and D. Stolten, "Classification of Building Types in Germany: A Data-Driven Modeling Approach," *Data*, vol. 7, no. 4, p. 45, Apr. 2022, number: 4 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: <https://www.mdpi.com/2306-5729/7/4/45>
- [7] X. Huang, L. Ren, C. Liu, Y. Wang, H. Yu, M. Schmitt, R. Hänsch, X. Sun, H. Huang, and H. Mayer, "Urban Building Classification (UBC) - A Dataset for Individual Building Detection and Classification From Satellite Imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Jun. 2022, pp. 1413–1421.
- [8] N. Niu, X. Liu, H. Jin, X. Ye, Y. Liu, X. Li, Y. Chen, and S. Li, "Integrating multi-source big data to infer building functions," *Int. J. Geogr. Inf. Sci.*, vol. 31, no. 9, pp. 1871–1890, Sep. 2017, publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/13658816.2017.1325489>. [Online]. Available: <https://doi.org/10.1080/13658816.2017.1325489>
- [9] B. Chen, B. Xu, and P. Gong, "Mapping essential urban land use categories (EULUC) using geospatial big data: Progress, challenges, and opportunities," *Big Earth Data*, vol. 5, no. 3, pp. 410–441, Jul. 2021, publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/20964471.2021.1939243>. [Online]. Available: <https://doi.org/10.1080/20964471.2021.1939243>
- [10] A. Lin, X. Sun, H. Wu, W. Luo, D. Wang, D. Zhong, Z. Wang, L. Zhao, and J. Zhu, "Identifying Urban Building Function by Integrating Remote Sensing Imagery and POI Data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 8864–8875, 2021, conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.
- [11] M. Yang, B. Kong, R. Dang, and X. Yan, "Classifying urban functional regions by integrating buildings and points-of-interest using a stacking ensemble method," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 108, p. 102753, Apr. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0303243422000794>
- [12] J. F. Rosier, H. Taubenböck, P. H. Verburg, and J. van Vliet, "Fusing Earth observation and socioeconomic data to increase the transferability of large-scale urban land use classification," *Remote Sens. Environ.*, vol. 278, p. 113076, Sep. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425722001900>
- [13] S. L. Barr, M. J. Barnsley, and A. Steel, "On the Separability of Urban Land-Use Categories in Fine Spatial Scale Land-Cover Data Using Structural Pattern Recognition," *Environ. Plan. B: Plan. Des.*, vol. 31, no. 3, pp. 397–418, Jun. 2004, publisher: SAGE Publications Ltd STM. [Online]. Available: <https://journals.sagepub.com/doi/abs/10.1068/b3016>
- [14] S. Steiniger, T. Lange, D. Burghardt, and R. Weibel, "An Approach for the Classification of Urban Building Structures Based on Discriminant Analysis Techniques," *Trans. GIS*, vol. 12, no. 1, pp. 31–59, Feb. 2008. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9671.2008.01085.x>
- [15] M. Marconcini, A. Metz-Marconcini, S. Üreyen, D. Palacios-Lopez, W. Hanke, F. Bachofer, J. Zeidler, T. Esch, N. Gorelick, A. Kakarla, M. Paganini, and E. Strano, "Outlining where humans live, the World Settlement Footprint 2015," *Sci. Data*, vol. 7, no. 1, p. 242, Jul. 2020, number: 1 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41597-020-00580-5>
- [16] T. Esch, E. Brzoska, S. Dech, B. Leutner, D. Palacios-Lopez, A. Metz-Marconcini, M. Marconcini, A. Roth, and J. Zeidler, "World Settlement Footprint 3D - A first three-dimensional survey of the global building stock," *Remote Sens. Environ.*, vol. 270, p. 112877, Mar. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425721005976>
- [17] M. Pesaresi, M. Schiavina, P. Politis, S. Freire, K. Krasnodębska, J. H. Uhl, A. Carioli, C. Corbane, L. Dijkstra, P. Florio, H. K. Friedrich, J. Gao, S. Leyk, L. Lu, L. Maffneni, I. Mari-Rivero, M. Melchiorri, V. Syrris, J. Van Den Hoek, and T. Kemper, "Advances on the Global Human Settlement Layer by joint assessment of Earth Observation and population survey data," *International Journal of Digital Earth*, vol. 17, no. 1, p. 2390454, Dec. 2024. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/17538947.2024.2390454>
- [18] W. Sirko, S. Kashubin, M. Ritter, A. Annkah, Y. S. E. Bouchareb, Y. Dauphin, D. Keysers, M. Neumann, M. Cisse, and J. Quinn, "Continental-Scale Building Detection from High Resolution Satellite Imagery," *arXiv:2107.12283 [cs]*, Jul. 2021, arXiv: 2107.12283. [Online]. Available: <http://arxiv.org/abs/2107.12283>
- [19] Microsoft, "GlobalMLBuildingFootprints," 2023. [Online]. Available: <https://github.com/microsoft/GlobalMLBuildingFootprints/>
- [20] O. Contributors, "Openstreetmap," 2024. [Online]. Available: <https://www.openstreetmap.org>
- [21] Y. Zhang, Q. Li, H. Huang, W. Wu, X. Du, and H. Wang, "The Combined Use of Remote Sensing and Social Sensing Data in Fine-Grained Urban Land Use Mapping: A Case Study in Beijing, China," *Remote Sensing*, vol. 9, no. 9, p. 865, Aug. 2017. [Online]. Available: <https://www.mdpi.com/2072-4292/9/9/865>
- [22] X. Li, C. Zhang, and W. Li, "Building block level urban land-use information retrieval based on Google Street View images," *GIScience & Remote Sensing*, vol. 54, no. 6, pp. 819–835, Nov. 2017. [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/15481603.2017.1338389>
- [23] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou, "A new insight into land use classification based on aggregated mobile phone data," *International Journal of Geographical Information Science*, vol. 28, no. 9, pp. 1988–2007, Sep. 2014. [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1080/13658816.2014.913794>
- [24] D. Duarte and C. C. Fonte, "Combining readily available population and land cover maps to generate non-residential built-up labels to train Sentinel-2 image segmentation models," *International Journal of Applied Earth Observation and Geoinformation*, vol. 135, p. 104272, Dec. 2024. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1569843224006289>
- [25] S. Gao, K. Janowicz, and H. Couclelis, "Extracting urban functional regions from points of interest and human activities on location-based social networks," *Transactions in GIS*, vol. 21, no. 3, pp. 446–467, Jun. 2017. [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1111/tgis.12289>
- [26] E. J. Hoffmann, K. Abdulahad, and X. X. Zhu, "Using social media images for building function classification," *Cities*, vol. 133, p. 104107, Feb. 2023. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0264275122005467>
- [27] F. Biljecki, Y. S. Chow, and K. Lee, "Quality of crowdsourced geospatial building information: A global assessment of OpenStreetMap attributes," *Build. Environ.*, vol. 237, p. 110295, Jun. 2023. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0360132323003220>
- [28] K. S. Atwal, T. Anderson, D. Pfoser, and A. Züfle, "Predicting building types using OpenStreetMap," *Sci. Rep.*, vol. 12, no. 1, p. 19976, Nov. 2022, number: 1 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41598-022-24263-w>
- [29] H. F. De Arruda, S. M. Reia, S. Ruan, K. S. Atwal, H. Kavak, T. Anderson, and D. Pfoser, "An OpenStreetMap derived building classification dataset for the United States," *Sci Data*, vol. 11, no. 1, p. 1210, Nov. 2024. [Online]. Available: <https://www.nature.com/articles/s41597-024-04046-w>
- [30] N. Johnson, W. Treible, and D. Crispell, "OpenSentinelMap: A Large-Scale Land Use Dataset using OpenStreetMap and Sentinel-2 Imagery," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. New Orleans,

- LA, USA: IEEE, Jun. 2022, pp. 1332–1340. [Online]. Available: <https://ieeexplore.ieee.org/document/9856983/>
- [31] C. Corbane, V. Syrris, F. Sabo, P. Politis, M. Melchiorri, M. Pesaresi, P. Soille, and T. Kemper, “Convolutional neural networks for global human settlements mapping from Sentinel-2 satellite imagery,” *Neural Comput. Appl.*, Oct. 2020. [Online]. Available: <http://link.springer.com/10.1007/s00521-020-05449-7>
- [32] B. Gunder-Williams, E. Mackres, S. Angel, A. M. Blei, and P. Lamson-Hall, “Intra-urban land use maps for a global sample of cities from Sentinel-2 satellite imagery and computer vision,” *Comput. Environ. Urban Syst.*, vol. 100, p. 101917, Mar. 2023. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0198971522001612>
- [33] A. Albert, J. Kaur, and M. C. Gonzalez, “Using Convolutional Networks and Satellite Imagery to Identify Patterns in Urban Environments at a Large Scale,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax NS Canada: ACM, Aug. 2017, pp. 1357–1366. [Online]. Available: <https://dl.acm.org/doi/10.1145/3097983.3098070>
- [34] T. Stark, M. Wurm, X. X. Zhu, and H. Taubenböck, “Satellite-Based Mapping of Urban Poverty With Transfer-Learned Slum Morphologies,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 5251–5263, 2020. [Online]. Available: <https://ieeexplore.ieee.org/document/9174807/>
- [35] L. Hu, C. Robinson, and B. Dilkina, “Model Generalization in Deep Learning Applications for Land Cover Mapping,” Jun. 2021, arXiv:2008.10351 [cs, stat]. [Online]. Available: <http://arxiv.org/abs/2008.10351>
- [36] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, “SEN12MS – A CURATED DATASET OF GEOREFERENCED MULTI-SPECTRAL SENTINEL-1/2 IMAGERY FOR DEEP LEARNING AND DATA FUSION,” in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7, 2019, pp. 153–160. [Online]. Available: <https://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/IV-2-W7/153/2019/>
- [37] S. Hafner, Y. Ban, and A. Nascetti, “Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data,” *Remote Sens. Environ.*, vol. 280, p. 113192, Oct. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425722003029>
- [38] X. X. Zhu, C. Qiu, J. Hu, Y. Shi, Y. Wang, M. Schmitt, and H. Taubenböck, “The urban morphology on our planet – Global perspectives from space,” *Remote Sens. Environ.*, vol. 269, p. 112794, Feb. 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0034425721005149>
- [39] X. X. Zhu, Q. Li, Y. Shi, Y. Wang, A. Stewart, and J. Prexl, “GlobalBuildingMap – Unveiling the Mystery of Global Buildings,” 2024, eprint: 2404.13911. [Online]. Available: <https://arxiv.org/abs/2404.13911>
- [40] P. Iakubovskii, “Segmentation Models Pytorch,” 2019, publication Title: GitHub repository. [Online]. Available: https://github.com/qubvel/segmentation_models.pytorch
- [41] S. Liu, E. Johns, and A. J. Davison, “End-To-End Multi-Task Learning With Attention,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 1871–1880. [Online]. Available: <https://ieeexplore.ieee.org/document/8954221/>
- [42] N. Skuppin, E. J. Hoffmann, Y. Shi, and X. X. Zhu, “Building Type Classification with Incomplete Labels,” in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. Kuala Lumpur, Malaysia: IEEE, Jul. 2022, pp. 5844–5847. [Online]. Available: <https://ieeexplore.ieee.org/document/9884076/>

APPENDIX D BIOGRAPHY SECTION



Nikolai Skuppin received the B.Sc. degree in aerospace engineering from the University of Stuttgart, Stuttgart, Germany, in 2016 and the M.Sc. degree in aerospace engineering from the Technical University of Munich (TUM), Germany, in 2019. He spent two years as a German National Trainee at the Mission Analysis section of the European Space Agency (ESA) in Darmstadt, Germany. Since 2021 he works as a research assistant at the Remote Sensing Technology Institute (IMF) of the German Aerospace Center (DLR) in Oberpfaffenhofen, Germany. His research interests include deep learning for Earth observation with a focus on the urban environment.



Yilei Shi (Member, IEEE) received the Dipl.-Ing. degree in mechanical engineering and the Dr.-Ing. degree in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2010 and 2019, respectively. He is a Senior Scientist with the Chair of Remote Sensing Technology, TUM. His research interests include fast solver and parallel computing for large-scale problems, high-performance computing and computational intelligence, advanced methods on synthetic aperture radar (SAR) and InSAR processing, machine learning, and deep learning for variety of data sources, such as SAR, optical images, and medical images, and partial differential equation (PDE)-related numerical modeling and computing.



Xiao Xiang Zhu (Fellow, IEEE) is the Chair Professor for Data Science in Earth Observation at the Technical University of Munich (TUM), Munich, Germany, where she also serves as Director of the Munich Data Science Institute (MDSI). She is currently a Visiting AI Professor at the European Space Agency’s ϕ -lab, Frascati, Italy. Since 2022, she leads the national center of excellence ML4Earth (ml4earth.de), funded by the German Space Agency. Her research interests include Earth observation and remote sensing, signal processing, machine learning, and data science, with applications to global urbanization, the UN Sustainable Development Goals, and climate change. More information is available at <https://www.asg.ed.tum.de/sipeo>