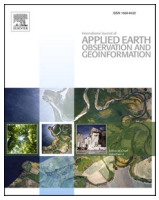
International Journal of Applied Earth Observations and Geoinformation 111 (2022) 102824

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



CrossGeoNet: A Framework for Building Footprint Generation of Label-Scarce Geographical Regions

Qingyu Li [a,b], Lichao Mou [a,b], Yuansheng Hua [a,b], Yilei Shi [c], Xiao Xiang Zhu [a,b,*]

[a] Data Science in Earth Observation, Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany
[b] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, 82234 Wessling, Germany
[c] Remote Sensing Technology (LMF), Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany

ARTICLE INFO

Keywords:
Building footprint
Semantic segmentation
Convolutional neural network
Co-segmentation
Planet satellite

ABSTRACT

Building footprints are essential for understanding urban dynamics. Planet satellite imagery with daily repetition frequency and high resolution has opened new opportunities for building mapping at large scales. However, suitable building mapping methods are scarce for less developed regions, as these regions lack massive annotated samples to provide strong supervisory information. To address this problem, we propose to learn cross-geolocation attention maps in a co-segmentation network, which is able to improve the discriminability of buildings within the target city and provide a more general building representation in different cities. In this way, the limited supervisory information resulting from insufficient training examples in target cities can be compensated. Our method is termed as CrossGeoNet, and consists of three elemental modules: a Siamese encoder, a cross-geolocation attention module, and a Siamese decoder. More specifically, the encoder learns feature maps from a pair of images from two different geo-locations. The cross-location attention module aims at learning similarity based on these two feature maps and can provide a global overview of common objects (e.g., buildings) in different cities. The decoder predicts segmentation masks of buildings using the learned cross-location attention maps and the original convolved images. The proposed method is evaluated on two datasets with different spatial resolutions, i.e., Planet dataset (3 m/pixel) and Inria dataset (0.3 m/pixel), which are collected from various locations around the world. Experimental results show that CrossGeoNet can well extract buildings of different sizes and alleviate false detections, which significantly outperforms other competitors.

1. Introduction

Building footprint maps offer insights for the comprehensive understanding of urban development. In less developed regions (e.g., Africa), significant changes occur in urban areas annually due to rapid urban expansion and city renewal (Huang et al., 2020), resulting in environmental and ecological problems (Guo et al., 2021a). Therefore, acquiring up-to-date building footprint maps for these regions is essential to the urban-related analysis.

In recent decades, high spatial resolution satellite images are capable of deriving spatial details of individual buildings. However, there are some weaknesses in high-resolution commercial satellites, e.g., high cost and low revisit frequency. This limits the regional or global building footprint generation. Planet is a new micro-satellite constellation, which consists of more than 120 satellites in orbit and is able to collect meter-level spatial resolution imagery on a daily basis at low-cost (Houborg and McCabe, 2016). Its high revisit capability also helps to acquire low cloud cover observations for the regions with above-average cloud cover (Asner et al., 2017). To date, most high-resolution building footprint generation studies are limited to aerial imagery (Bischke et al., 2019, Bischke et al., 2019; Maggiori et al., 2017,Maggiori et al., 2017; Li et al., 2020,Li et al., 2020) or WorldView satellite imagery (Pan et al., 2020b; Pan et al., 2020b; Tonbul and Kavzoglu, 2020; Tonbul and Kavzoglu, 2020), and the investigation on Planet satellite imagery is lacking.

Although some approaches (Ivanovsky et al., 2019,Ivanovsky et al., 2019;Li et al., 2020, Li et al., 2020;Li et al., 2021, Li et al., 2021;Shi et al., 2020,Shi et al., 2020) are capable of delivering very promising results on Planet satellite imagery, they are mostly developed for Europe. To the best of our knowledge, few are dedicated to the cities in less developed regions represented by Africa, South America, and Asia,

* Corresponding author.
E-mail addresses: qingyu.li@tum.de (Q. Li), lichao.mou@dlr.de (L. Mou), yuansheng.hua@dlr.de (Y. Hua), yilei.shi@tum.de (Y. Shi), xiaoxiang.zhu@dlr.de (X.X. Zhu).

https://doi.org/10.1016/j.jag.2022.102824
Received 18 February 2022; Received in revised form 11 May 2022; Accepted 13 May 2022
Available online 11 June 2022

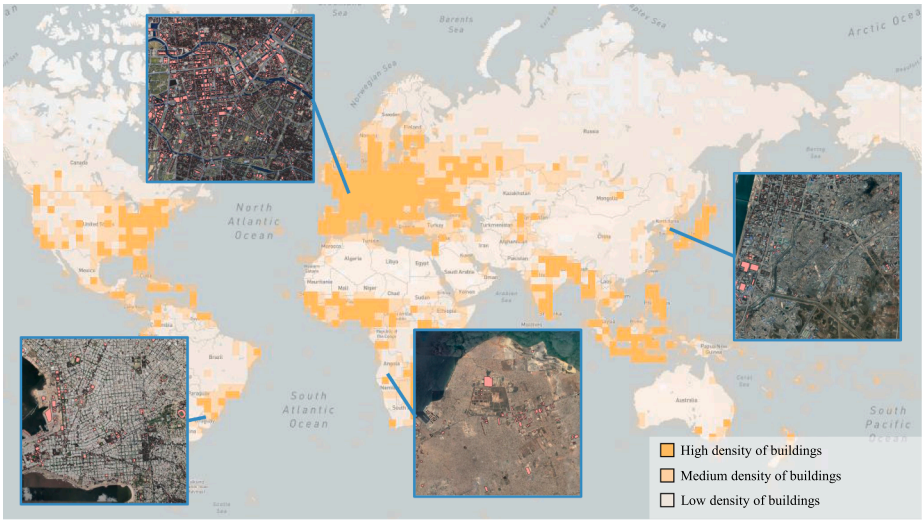


Fig. 1. The annotated building footprints in OpenStreetMap (counted by continents), and four examples of cities in Europe, Africa, South America, and Asia. The base map about building densities on OpenStreetMap is obtained from OpenStreetMap Analytics (osm, 2021-08-24.).

Fig. 2. Illustration of geographic peculiarities across different geolocations. The Planet satellite images are collected from (a) Munich (Germany), (b) Yaounde (Cameroon), (c) Lisbon (Portugal), and (d) Niamey (Niger), respectively. We can see that appearances of buildings in different cities are noticeably different.

where buildings differ substantially in size and type from those in Europe.

To generate building footprint maps from Planet satellite imagery, existing studies use convolutional neural networks (CNNs) that can effectively learn high-level features from raw data without heuristic feature design. Nevertheless, there remains a challenge for extracting building footprints on target cities — massive data need to be collected to promote the performance of CNNs. Considering that the manual annotation of reference data is a very time-consuming and costly process, OpenStreetMap (OSM) could be considered as a good source for acquiring manually annotated building footprints for training networks (Kaiser et al., 2017). By analyzing available building annotation data in OSM, we observe that they have an extremely uneven distribution across cities in different continents (see Fig. 1). For example, there are abundant labeled samples in European cities, while for cities in Africa, South America, and Asia, annotated data are quite limited. The lack of annotated data usually restricts the performance of existing methods in these regions, as they require a lot of strong supervisory information for network learning.

In this paper, we aim to generate building footprint maps using Planet satellite imagery for target cities that suffer from data deficit of labeled samples. In order to improve the performance of a network trained on the target city with scarce labeled data, a straightforward idea is to take advantage of the cities with massive annotated data (hereafter called auxiliary set). Nonetheless, geographic peculiarities across different geolocations raise several challenges. As shown in Fig. 2, appearances of buildings in different continents are noticeably different. This induces CNNs to produce unsatisfactory results when we directly apply a network trained on the auxiliary set to target cities. In this regard, some works (Maggiori et al., 2016) utilize transfer learning that fine-tunes a pre-trained model with a few labeled instances in target cities. Domain adaptation methods (Vu et al., 2019) aim to transfer the knowledge learned from a domain to improve performance on target cities. Other works (He et al., 2020) utilize a new learning strategy, where the model is first pre-trained with a large number of unlabeled images in a self-supervised way and then transferred to the task of semantic segmentation with very few labeled samples.

Recently, co-segmentation is proposed for the object segmentation in computer vision, aiming at jointly segmenting semantically similar objects in video frames (Papoutsakis et al., 2017; Papoutsakis et al., 2017; Wang et al., 2019; Wang et al., 2019) or multiple images (Li et al., 2018). The success of these works suggests that co-segmentation can fully harness the sequential or pair-wise relations among consecutive frames to discover common objects, which helps to alleviate the dependency of strong supervisory information. This gives us an incentive that the co-segmentation framework may benefit our cross-city building extraction task. Therefore, we propose an end-to-end trainable network—CrossGeoNet, which consists of three modules: a Siamese encoder, a cross-geolocation attention module, and a Siamese decoder. The encoder takes as input a pair of images from two different geolocations and is responsible for learning feature representations for both images. The cross-geolocation attention module learns to explicitly encode correlations between the features of the two images, enabling the network to attend more to common objects (i.e., building in our case). The decoder combines convolved images with the cross-geolocation attention maps to generate segmentation masks through a series of

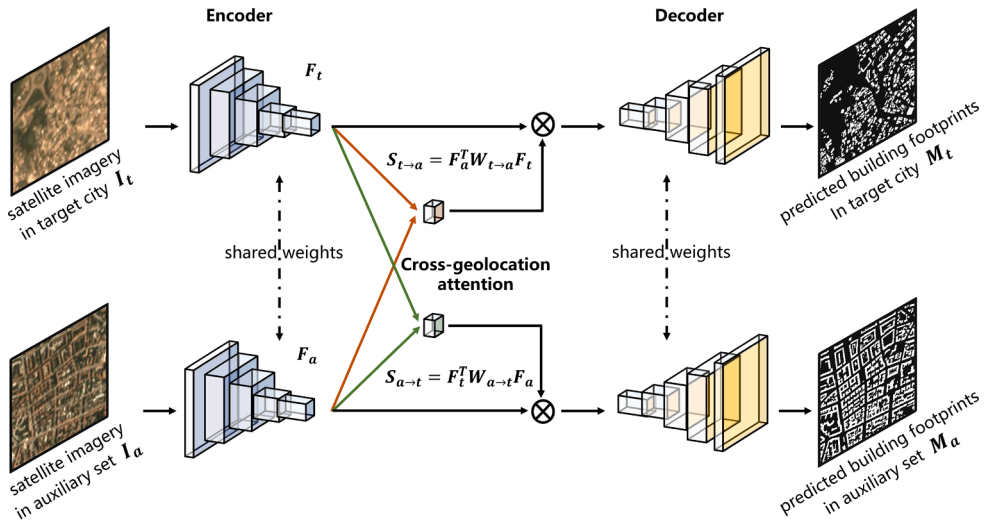


Fig. 3. Overview of the proposed CrossGeoNet framework.

deconvolutional layers. Note that the three components are jointly optimized in our method. This work's contributions are threefold.

(1) The proposed CrossGeoNet examines the potential of Planet satellite imagery for building mapping in less developed regions (e.g., cities in Africa, South America, and Asia).
(2) To tackle the problem of insufficient labeled samples in target cities, we propose to use a co-segmentation learning framework that can leverage a large amount of labeled data in other cities to improve the performance of a model in the target cities. To the best of our knowledge, our work is the first one that exploits co-segmentation learning to generate building footprint maps.
(3) Since capturing the relationship between the two inputs is the key element in our CrossGeoNet, we propose a cross-geolocation attention module to effectively learn the underlying similarity between different geolocations, which is superior to other existing methods (e.g. mutual correlation (Li et al., 2018) and Fourier domain correlation (Danelljan et al., 2014)). Compared with other competitors, our approach gains significantly better results. The codes of CrossGeoNet will be made publicly available in https://github.com/lqycrystal/coseg_building.

This article is organized as follows. Section 2 presents the framework of CrossGeoNet for building footprint generation. The experiments are described in Section 3. Results are provided in Section 4. The performance of CrossGeoNet on another data source is investigated in Section 5. Eventually, Section 6 summarizes this work.

2. Methodology

In this section, the co-segmentation pipeline of CrossGeoNet is first presented. Afterward, we present the proposed cross-geolocation attention module in detail. Finally, the end-to-end network learning procedure is described.

2.1. Co-segmentation Pipeline

When objects of the same class vary in pose, shape, or color, the idea of co-learning can exploit the synergistic relationship between video frames or multiple images to provide generic features, improving model performance. In this work, our motivation is that by jointly viewing common objects (i.e., building in our case) in different geolocations, networks can learn underlying similarities for extracting more generic representations for buildings. In this regard, we propose to integrate co-segmentation learning into the framework of building footprint generation, which is capable of fully harnessing information from various locations and further enhancing the generalizability of the model. Specifically, we propose a cross-geolocation attention module in the co-segmentation pipeline that learns to enhance latent features by encoding relations between the target city and cities from the auxiliary set. As a consequence, our co-segmentation network is able to not only improve building discriminability within target cities but also learn generic features of buildings across different cities. By doing so, the limited supervisory information in target cities can be compensated.

As shown in Fig. 3, a Siamese encoder-decoder architecture is adopted in CrossGeoNet. The Siamese encoder is composed of two identical CNNs with shared weights for the purpose of feature extraction. The input of the encoder is an image pair, where one image $\boldsymbol{I}_t$ is from a target city and the other image $\boldsymbol{I}_a$ is from the auxiliary set, and their feature representations are denoted as $\boldsymbol{F}_t \in \mathbb{R}^{C\times W\times H}$ and $\boldsymbol{F}_a \in \mathbb{R}^{C\times W\times H}$, respectively. H and W represent the height and width, and C denotes the channel dimension. Unlike conventional semantic segmentation networks, where high-level features are directly decoded for inferring building masks, here we enhance the learned feature maps through the proposed cross-geolocation attention module. Specifically, this module takes two feature maps as input and outputs two attention maps $\boldsymbol{S}_{t\to a}$ and $\boldsymbol{S}_{a\to t}$. Afterward, they are fused with the corresponding convolved images and fed into the decoder. The Siamese decoder is comprised of a set of transposed convolutional layers that upsample the convolved images to generate two building segmentation masks $\mathbf{M}_t$ and $\mathbf{M}_a$. Note that all modules are integrated into one framework and optimized in an end-to-end manner.

2.2. Cross-geolocation Attention

The feature maps learned from the Siamese encoder contain abstract semantic information, and when the input images contain the common object (e.g., building), their features should also include similar information. The key component of co-segmentation learning is to find similarities in feature vectors among various images. In the literature, there have been several commonly used similarity measures, e.g., mutual correlation (Li et al., 2018) and Fourier domain correlation (Danelljan et al., 2014).

Inspired by the success of self-attention (Hu et al., 2018) in capturing long-range interactions among input signals, we propose a cross-geolocation attention module that can adaptively learn the similarity between target cities and the auxiliary set. By doing so, semantic information of the common object (e.g., building) can be enhanced. More

Table 1
Statistics of the datasets utilized in this research.

	Continent	Name	The number of patches		
			train	validation	test
Target city	Africa	Yaounde	100	100	300
	South America	Porto Alegre	100	100	300
	Asia	Kyoto	100	100	300
Auxiliary set	Europe	Madrid	2971	743	0
		London	2256	565	0
		Rome	2303	576	0
		Lisbon	2043	511	0
		Munich	2271	568	0
		Zurich	1849	463	0

specifically, we calculate the cross-geolocation attention map $\boldsymbol{S}_{t\to a} \in \mathbb{R}^{(WH)\times(WH)}$ between $\boldsymbol{F}_t$ and $\boldsymbol{F}_a$ as:

$$\boldsymbol{S}_{t\to a} = \boldsymbol{F}_a^T \boldsymbol{W}_{t\to a} \boldsymbol{F}_t, \tag{1}$$

where $\boldsymbol{W}_{t\to a} \in \mathbb{R}^{C\times C}$ is a weight matrix. Here $\boldsymbol{F}_t$ and $\boldsymbol{F}_a$ are flattened into vectors with the size of $C \times WH$ and can be represented as:

$$\boldsymbol{F}_t = [\boldsymbol{f}_t^1, \boldsymbol{f}_t^2, \ldots, \boldsymbol{f}_t^p, \ldots, \boldsymbol{f}_t^{WH}], \tag{2}$$

$$\boldsymbol{F}_a = [\boldsymbol{f}_a^1, \boldsymbol{f}_a^2, \ldots, \boldsymbol{f}_a^q, \ldots, \boldsymbol{f}_a^{WH}], \tag{3}$$

where $\boldsymbol{f}_t^p$ is a C-dimiensional feature vector at position $p \in \{1, 2, \ldots, WH\}$ in $\boldsymbol{F}_t$, and $\boldsymbol{f}_a^q$ is a C-dimiensional feature vector at position $q \in \{1, 2, \ldots, WH\}$ in $\boldsymbol{F}_a$. Thus, the entry (q, p) of $\boldsymbol{S}_{t\to a}$ reflects the similarity between $\boldsymbol{f}_a^q$ and $\boldsymbol{f}_t^p$, and can be learned automatically. $\boldsymbol{S}_{t\to a}$ is capable of capturing the dependencies between any two positions of feature maps without regard for their distance in the spatial dimension. Therefore, our cross-geolocation module can model rich contextual dependencies, which is superior to other similarity measures that only consider local features.

Since the weight matrix $\boldsymbol{W}_{t\to a}$ is a square matrix, the diagonalization of $\boldsymbol{W}_{t\to a}$ can be represented as follows:

$$\boldsymbol{W}_{t\to a} = \boldsymbol{P}_{t\to a}^{-1} \boldsymbol{D}_{t\to a} \boldsymbol{P}_{t\to a}, \tag{4}$$

where $\boldsymbol{P}_{t\to a}$ is an invertible matrix and $\boldsymbol{D}_{t\to a}$ is a diagonal matrix. Then, Eq. (1) can be rewritten as:

$$\boldsymbol{S}_{t\to a} = \boldsymbol{F}_a^T \boldsymbol{P}_{t\to a}^{-1} \boldsymbol{D}_{t\to a} \boldsymbol{P}_{t\to a} \mathbf{F}_t. \tag{5}$$

According to Eq. (5), a learnable linear transformation is first applied to the feature representation of each image, and then the similarity between these two feature representations is dynamically captured by the dot product. Similarly, the cross-geolocation attention map $\boldsymbol{S}_{a\to t}$ between $\boldsymbol{F}_a$ and $\boldsymbol{F}_t$ is computed as:

$$\boldsymbol{S}_{a\to t} = \boldsymbol{F}_t^T \boldsymbol{P}_{a\to t}^{-1} \boldsymbol{D}_{a\to t} \boldsymbol{P}_{a\to t} \mathbf{F}_a, \tag{6}$$

where $\boldsymbol{P}_{a\to t}$ is an invertible matrix, and $\boldsymbol{D}_{a\to t}$ is a diagonal matrix.

Note that $\boldsymbol{S}_{t\to a}^q$ is the q-th row of $\boldsymbol{S}_{t\to a}$, which is a vector with length WH and represents the similarity between each feature vector in $\boldsymbol{F}_t$ and $\boldsymbol{f}_a^q$. If the p-th element in $\boldsymbol{S}_{t\to a}^q$ has a larger value than others, $\boldsymbol{f}_t^p$ is more similar to $\boldsymbol{f}_a^q$ than other feature vectors in $\boldsymbol{F}_t$, which indicates a very high probability of having the common object in $\boldsymbol{f}_t^p$ and $\boldsymbol{f}_a^q$.

Afterward, we obtain the cross-geolocation attention-enhanced features $\boldsymbol{Z}_t$ by allocating the learned cross-geolocation attention map to $\boldsymbol{F}_t$, which is computed with the following equations:

$$\boldsymbol{Z}_t = \boldsymbol{S}_{t\to a} \boldsymbol{F}_t^T. \tag{7}$$

And $\boldsymbol{Z}_a$ is calculated in the same manner:

$$\boldsymbol{Z}_a = \boldsymbol{S}_{a\to t} \boldsymbol{F}_a^T. \tag{8}$$

Finally, $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_a$ are reshaped into the size of $C \times H \times W$ and fed into the Siamese decoder to produce final segmentation masks $\boldsymbol{M}_t$ and $\boldsymbol{M}_a$, respectively.

In what follows, we discuss in detail why the proposed approach can improve the performance of a model in target cities. It is well known that contextual information is able to offer important cues for semantic segmentation tasks. In conventional CNNs, convolutions are used to extract such information. However, the performance might be limited due to their local receptive fields. Also, inadequate samples affect the learning of CNNs. On the contrary, the proposed cross-geolocation module explores global contextual information by learning cross-geolocation attention maps. Specifically, for a pixel in a sample from the target city, the cross-geolocation attention map can effectively capture relations between it and not only all other pixels in the same sample but also all pixels in a sample from the auxiliary set. Afterward, CrossGeoNet selectively aggregates global contextual information to provide a global view of common objects (i.e., building), alleviating the influence of background. In other words, we leverage the auxiliary set to provide additional supervisory information to enhance the discriminability of building, which improves building extraction results on the target city.

2.3. Network Learning

We propose an end-to-end training pipeline for the supervised learning of CrossGeoNet. The whole network is trained by the following loss function:

$$L = L_t + \lambda \cdot L_a, \tag{9}$$

where L_t and L_a are two cross-entropy loss functions for measuring the difference between segmentation masks and their corresponding ground-truth masks. λ is a hyperparameter to control the importance of the second loss.

3. Experiments

3.1. Dataset

In this work, we collect Planet satellite images and their corresponding OSM building footprints from different cities all over the globe. Planet satellite images have 3 bands (i.e., red, green, blue), and their spatial resolution is 3 m/pixel. In the pre-processing step, all images and ground-truth masks are cropped into small patches with the size of 256 × 256 pixels. To thoroughly investigate the performance of CrossGeoNet, we select three target cities from different continents: Yaounde (Cameroon), Porto Alegre (Brazil), and Kyoto (Japan). As to the auxiliary set, 6 European cities, Madrid (Spain), London (UK), Rome (Italy), Lisbon (Portugal), Munich (Germany), and Zurich (Switzerland), are selected due to their massive building footprint annotations. The numbers of patches collected from each city for network training, validation, and test are reported in Table 1.

3.2. Experimental Setup

To verify the effectiveness of CrossGeoNet for building footprint generation, we compare it with several commonly-used network learning methods, i.e., Baseline-t, Baseline-a, Baseline-a+t, fine-tuning, ADVENT (Vu et al., 2019) IntraDA (Pan et al., 2020a), MetaCorrection (Guo et al., 2021b), MoCo (He et al., 2020), DenseCL (Wang et al., 2021), U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), and CSGANet (Chen et al., 2021). Note that experiments are independently conducted in three target cities. That is to say, for the experiment in one target city, training samples consist of only patches from that target city and the auxiliary set. For the evaluation of our cross-geolocation attention module, we conduct comparisons with the aforementioned two similarity measures, i.e., mutual correlation (Li

Table 2
Accuracies (%) of different learning methods for building footprint generation on tagert cities.

Method	Yaounde		Porto Alegre		Kyoto	
	F1 score	IoU	F1 score	IoU	F1 score	IoU
Baseline-t	63.85	46.90	58.57	41.41	59.80	42.65
Baseline-a	1.90	0.96	27.41	15.88	36.35	22.21
Baseline-a+t	64.95	48.10	60.44	43.31	62.76	45.72
Fine-tuning	63.35	46.36	60.12	42.98	59.31	42.16
ADVENT(Vu et al., 2019)	55.26	38.18	31.13	18.43	46.89	30.63
IntraDA (Pan et al., 2020a)	56.59	39.46	40.86	25.67	53.05	36.10
MetaCorrection (Guo et al., 2021b)	55.44	38.35	51.68	34.84	49.27	32.69
MoCo(He et al., 2020)	60.98	43.87	57.59	40.44	58.22	41.06
DenseCL(Wang et al., 2021)	60.99	43.88	59.12	39.00	58.10	40.94
U-Net-AFM (Li et al., 2021)	61.32	44.19	53.64	36.72	52.86	36.01
CBRNet (Guo et al., 2022)	63.52	46.54	59.98	42.84	61.78	44.70
EPU-Net (Guo et al., 2021a)	52.45	35.55	45.72	29.64	50.04	33.37
CGSANet (Chen et al., 2021)	61.51	44.42	56.69	39.55	58.07	40.92
CrossGeoNet	**67.77**	**51.26**	**62.12**	**45.05**	**65.28**	**48.46**

et al., 2018) and Fourier domain correlation (Danelljan et al., 2014).

3.3. Training Details

CrossGeoNet is implemented on PyTorch framework and trained on an NVIDIA Quadro P4000 GPU with 8 GB memory. The training epochs of all models are set as 100 epochs, and stochastic gradient descent (SGD) with a learning rate of 0.001 is set as the optimizer. The size of the training batch for all models is 4. Detailed configurations of all methods in our experiments are presented as follows:

(1) CrossGeoNet: Since our model is trained for each target city independently, we select $\boldsymbol{I}_t$ and $\boldsymbol{I}_a$ from one target city and the auxiliary set, respectively, in the training phase. To enlarge the number of training pairs, for each patch in the target city, we create 100 duplicates and pair them with 100 samples randomly selected from one city in the auxiliary set. In the inference stage, $\boldsymbol{I}_t$ and $\boldsymbol{I}_a$ are both selected from test patches of the target city. The loss term weighting parameter λ in Eq. (9) is set as 0.00001 empirically.
(2) Baseline-t : An Efficient-UNet is trained and tested with training and test sets of the target city.
(3) Baseline-a: An Efficient-UNet is trained with samples collected from the auxiliary set and tested on test instances in the target city.
(4) Baseline-a+t: An Efficient-UNet is trained using samples from training sets of the target city and the auxiliary set, and tested on test samples from the target city.
(5) Fine-tuning: It consists of two steps. Firstly, all samples from the auxiliary set are used to pre-train an Efficient-UNet. Secondly, the pre-trained network is fine-tuned with the training set of the target city.
(6) ADVENT (Vu et al., 2019), IntraDA (Pan et al., 2020a), and MetaCorrection (Guo et al., 2021b): They aim at addressing the task of domain adaptation in semantic segmentation. The auxiliary set is regarded as the source domain, and the target city is the target domain.
(7) MoCo (He et al., 2020) and DenseCL (Wang et al., 2021): They first learn knowledge from a large number of unlabeled images in a self-supervised way. Afterward, the weights are transferred to the task of semantic segmentation. In our research, MoCo (He et al., 2020) learns from the auxiliary set, while for DenseCL (Wang et al., 2021), we use its pre-trained weights (Deng et al., 2009).
(8) U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), and CSGANet (Chen et al., 2021): They are semantic segmentation networks for the task of building footprint generation.

Note that for MoCo (He et al., 2020), DenseCL (Wang et al., 2021), U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), and CSGANet (Chen et al., 2021), we have separately organized the training set according to three experiment procedures (i.e., Baseline-t, Baseline-a+t, and Fine-tuning), and the best result among three cases is reported.

We evaluate the performance of all models using two metrics: F1 score and intersection over union (IoU).

4. Results

4.1. Comparison of Different Learning Methods

This section presents the comparisons among CrossGeoNet, Baseline-t, Baseline-a, Baseline-a+t, fine-tuning, ADVENT (Vu et al., 2019), IntraDA (Pan et al., 2020a), MetaCorrection (Guo et al., 2021b), Moco (He et al., 2020), DenseCL (Wang et al., 2021), U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), and CSGANet (Chen et al., 2021). Their performance is evaluated from quantitative (cf. Tables 2) and and qualitative (see Figs. 4–6) perspectives in three target cities.

Compared with Baseline-t, the proposed method has largely improved the accuracy. It can be seen from numerical results in three target cities that CrossGeoNet reaches improvements of above 3% in both F1 score and IoU. Especially for the target city of Kyoto, our method obtains increments of 5.48% in F1 score and 5.81% in IoU, respectively. As shown in Fig. 4, Baseline-t fails to recover complete masks of large buildings. This is due to the fact that limited training samples can not represent the true class distribution comprehensively (Hou et al., 2019). Although Baseline-a exploits massive annotated samples of the auxiliary set, it still performs worse than CrossGeoNet. For instance, in the target city of Yaounde (see Table 2), Baseline-a only achieves 1.90% in F1 score and 0.96% in IoU. Moreover, these results are worse than those of Baseline-t. This is caused by significant differences between the target cities and the auxiliary set, e.g., variant morphological appearance of human settlements and material available for building construction (Li et al., 2020).

Afterward, we select another seven competitors (Baseline-a+t, fine-tuning, ADVENT (Vu et al., 2019), IntraDA (Pan et al., 2020a), MetaCorrection (Guo et al., 2021b), MoCo (He et al., 2020), and DenseCL (Wang et al., 2021)) to make a further comparison, as these methods also jointly utilize training samples of both the target city and the auxiliary set. Fine-tuning is a commonly used method to handle the issue of scarce training data in target datasets (Maggiori et al., 2016). Nevertheless, compared with Baseline-t, fine-tuning even leads to decreases in accuracy metrics for Yaounde and Kyoto. A possible explanation is that the gap between target cities and auxiliary set is quite large, making it difficult to transfer the knowledge learned from the auxiliary set to target cities. Domain adaptation methods are also capable of transferring the knowledge from the auxiliary set to the target city. From the results in Table 2, it can be seen that ADVENT (Vu et al., 2019), IntraDA (Pan et al., 2020a), and MetaCorrection (Guo et al., 2021b) perform worse than fine-tuning in knowledge transfer. One important reason is that the labels in the target domain are not utilized by domain adaptation methods. It can be observed from statistical results that MoCo (He et al., 2020) and DenseCL (Wang et al., 2021) are even inferior to Baseline-t on all three cities. This might be attributed to two factors. On

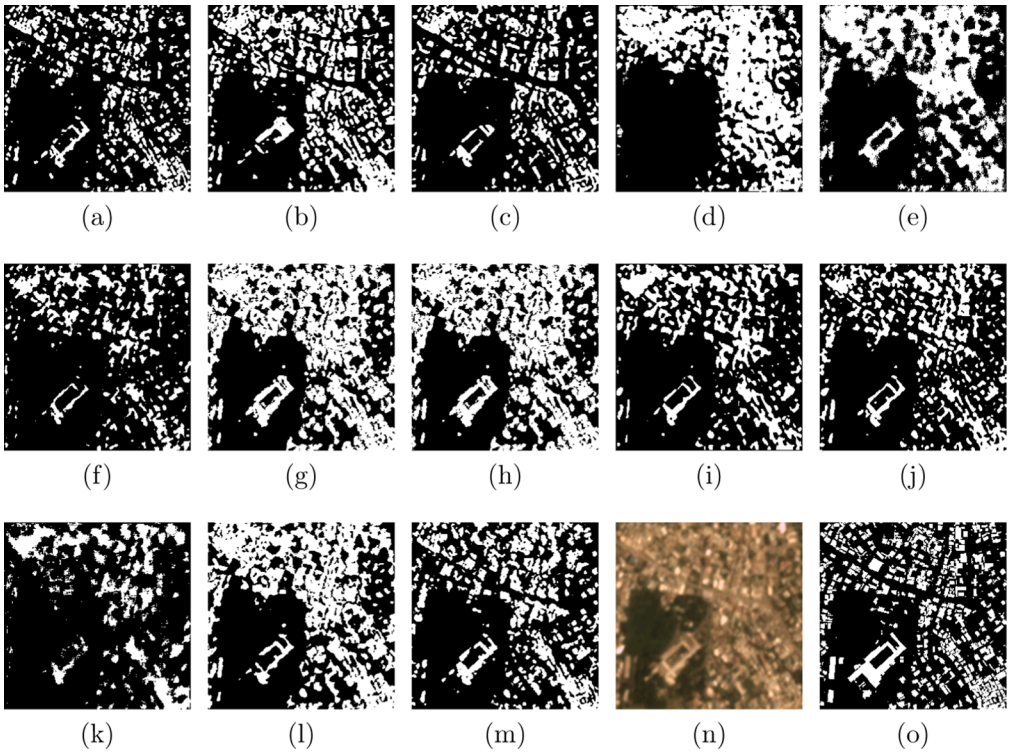


Fig. 4. Examples of building extraction results obtained by different learning methods. (a) Baseline-t. (b) Baseline-a+t. (c) Fine-tuning. (d) ADVENT (Vu et al., 2019). (e) IntraDA (Pan et al., 2020a). (f) MetaCorrection (Guo et al., 2021b). (g) MoCo (He et al., 2020). (h) DenseCL (Wang et al., 2021). (i) U-Net-AFM (Li et al., 2021). (j) CBRNet (Guo et al., 2022), (k) EPU-Net (Guo et al., 2021a). (l) CSGANet (Chen et al., 2021). (m) Cross-GeoNet. (n) and (o) are Planet satellite imagery and ground reference from Yaounde.

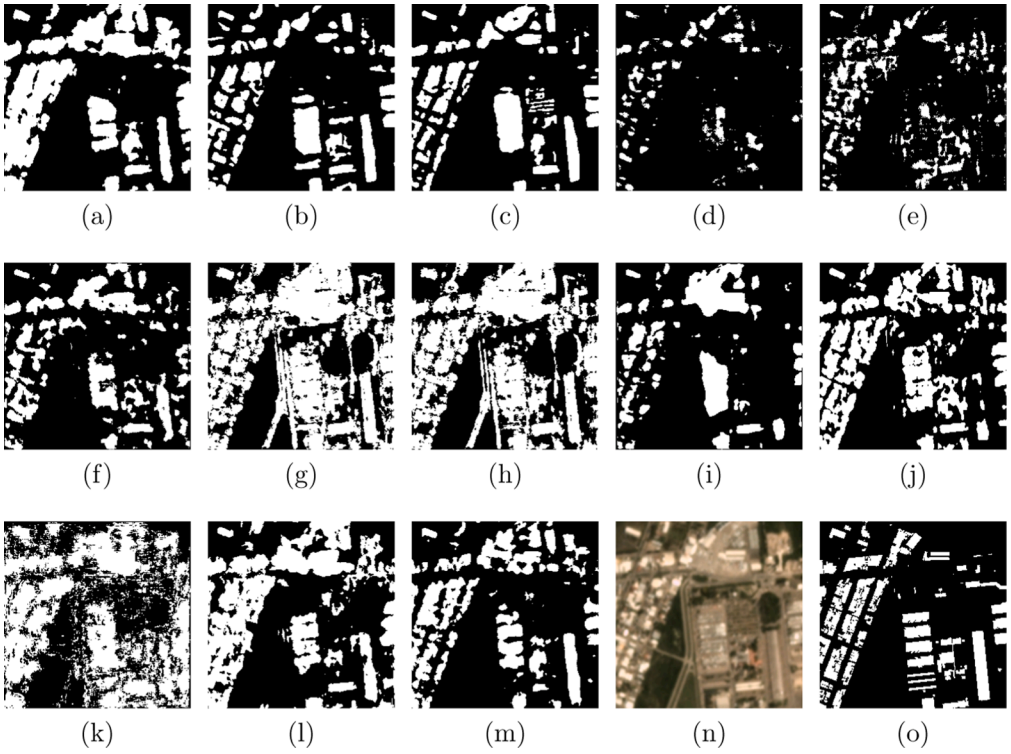


Fig. 5. Examples of building extraction results obtained by different learning methods. (a) Baseline-t. (b) Baseline-a+t. (c) Fine-tuning. (d) ADVENT (Vu et al., 2019). (e) IntraDA (Pan et al., 2020a). (f) MetaCorrection (Guo et al., 2021b). (g) MoCo (He et al., 2020). (h) DenseCL (Wang et al., 2021). (i) U-Net-AFM (Li et al., 2021). (j) CBRNet (Guo et al., 2022), (k) EPU-Net (Guo et al., 2021a). (l) CSGANet (Chen et al., 2021). (m) Cross-GeoNet. (n) and (o) are Planet satellite imagery and ground reference from Porto Alegre.

the one hand, the annotated information of the auxiliary set has not been leveraged in self-supervised learning. On the other hand, large differences existing between the auxiliary set and target cities might impair the model performance when migrated to target cities.

CrossGeoNet has achieved the highest accuracies among all methods, and it shows nearly 2% improvements of F1 score and IoU on all target cities compared to Baseline-a+t. From qualitative results, we can observe that Baseline-a+t fails to detect some small buildings (cf. Fig. 6). This can be explained by the imbalanced number of training samples collected from target cities and the auxiliary set. When training samples of the auxiliary set dominate the learning procedure, the network fails to guarantee accurate segmentation in target cities. On the contrary, our method is able to avoid these omission errors and reconstruct complete building structures to a large extent. These observations suggest that

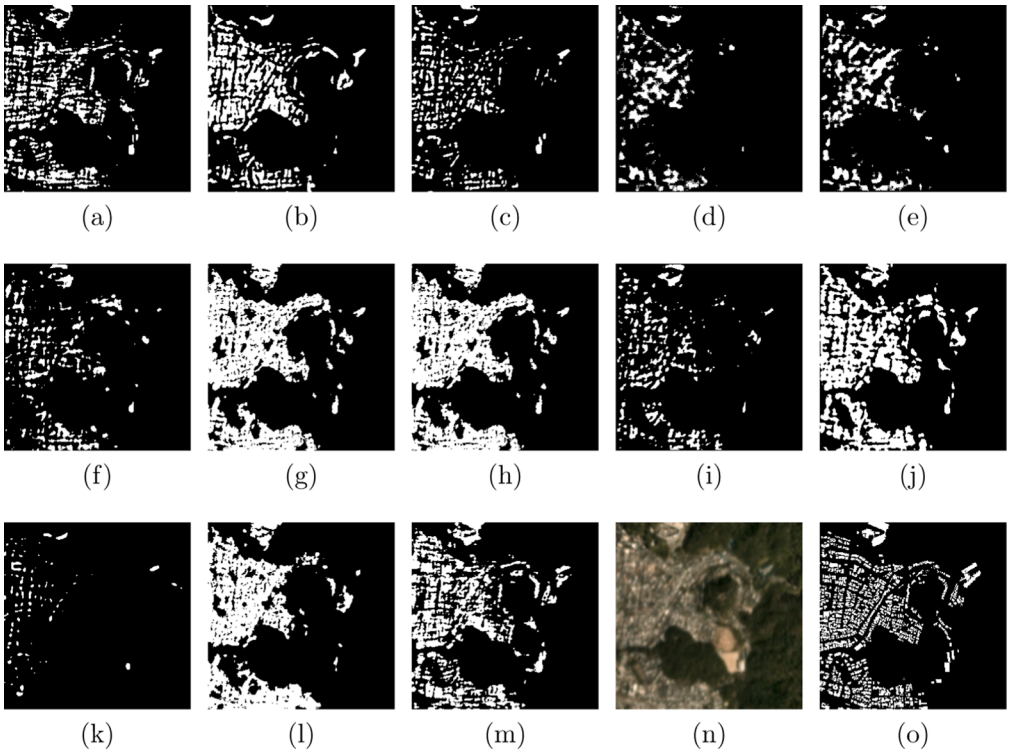


Fig. 6. Examples of building extraction results obtained by different learning methods. (a) Baseline-t. (b) Baseline-a+t. (c) Fine-tuning. (d) ADVENT (Vu et al., 2019). (e) IntraDA (Pan et al., 2020a). (f) MetaCorrection (Guo et al., 2021b). (g) MoCo (He et al., 2020). (h) DenseCL (Wang et al., 2021). (i) U-Net-AFM (Li et al., 2021). (j) CBRNet (Guo et al., 2022), (k) EPU-Net (Guo et al., 2021a). (l) CSGANet (Chen et al., 2021). (m) Cross-GeoNet. (n) and (o) are Planet satellite imagery and ground reference from from Kyoto.

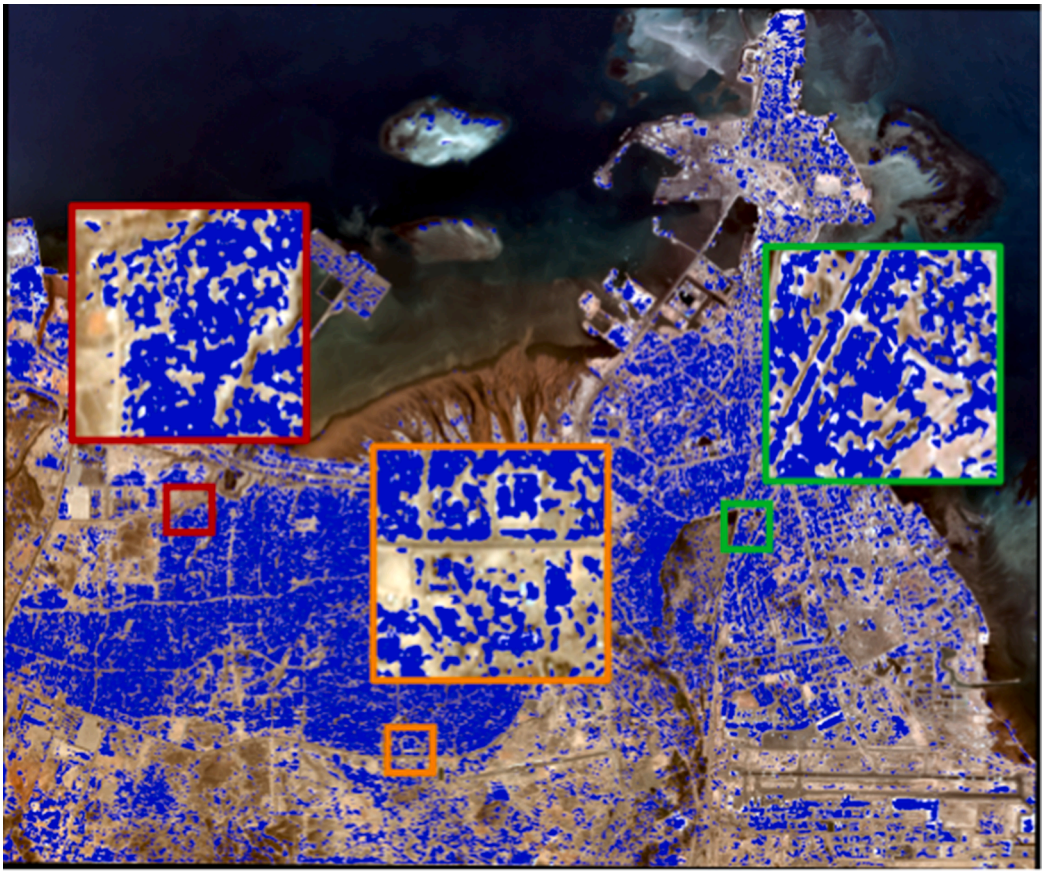


Fig. 7. Building extraction results (in blue) obtained by CrossGeoNet from Djibouti and three zoomed in areas.

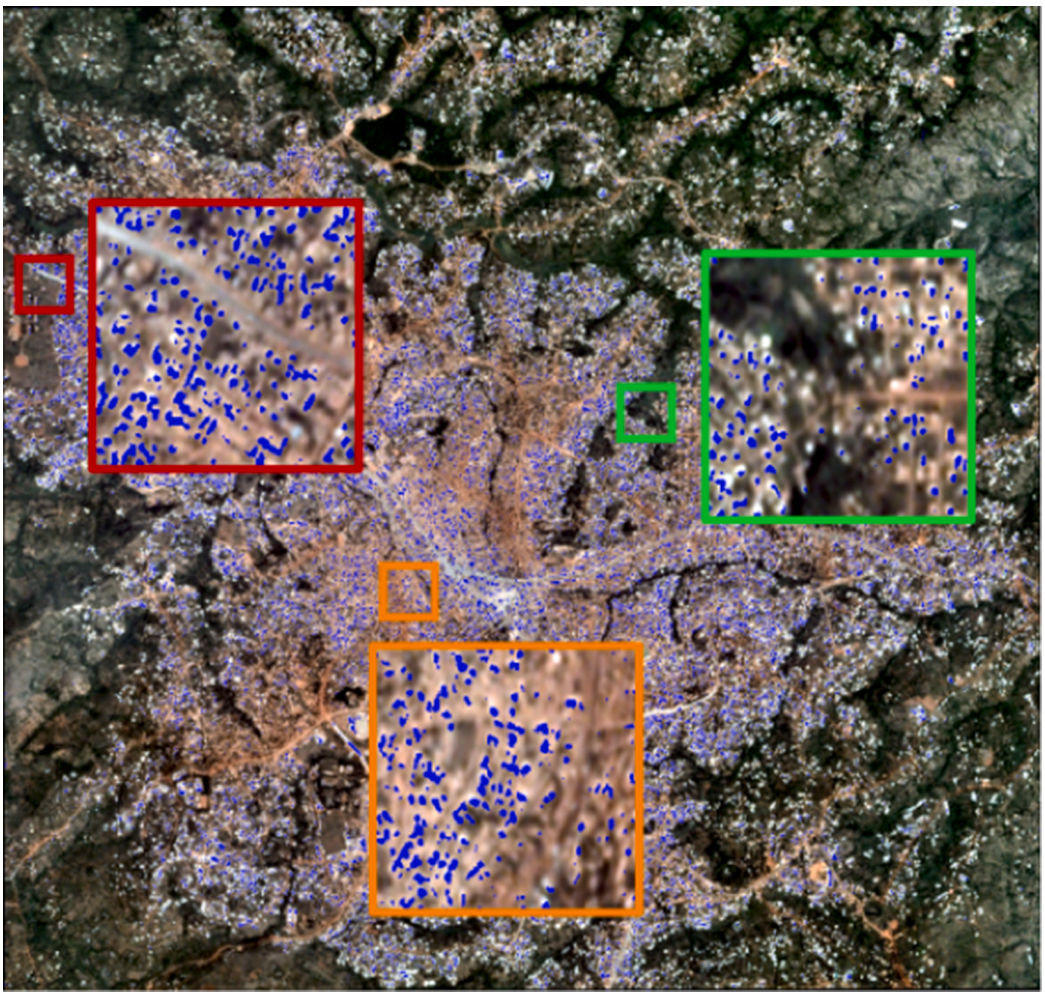


Fig. 8. Building extraction results (in blue) obtained by CrossGeoNet from Bafoussam and three zoomed in areas.

Table 3
Accuracies (%) of different similarity measures on Yaounde.

Method	F1 score	IoU
Mutual correlation (Li et al., 2018)	66.78	50.13
Fourier domain correlation (Danelljan et al., 2014)	65.76	48.99
Proposed cross-geolocation attention module	**67.77**	**51.26**

CrossGeoNet benefits from the learning of the cross-geolocation attention module, enabling the leverage of rich relationships between target cities and the auxiliary set.

We then compare CrossGeoNet with U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), and CSGANet (Chen et al., 2021), which are four state-of-the-art methods for the task of building footprint generation. It can be observed from the statistical and visual results on three cities that our method surpasses all other building extraction methods.

We further explore the generalizability of model trained by CrossGeoNet and test it on unseen cities (which are neither from the target city nor from the auxiliary set). Note that we directly apply the trained model to the unseen cities. Specifically, we select two African cities, Djibouti (Republic of Djibouti) and Bafoussam (Cameroon). In the training phase, we select Yaounde as the target city due to its high similarity with Djibouti and Bafoussam. Figs. 7 and 8 illustrate visual results on these two cities. CrossGeoNet is promising to provide building footprint maps in other unseen geographic regions.

4.2. Comparison With Different Similarity Measures

Explicitly capturing similarities among various cities is essential for

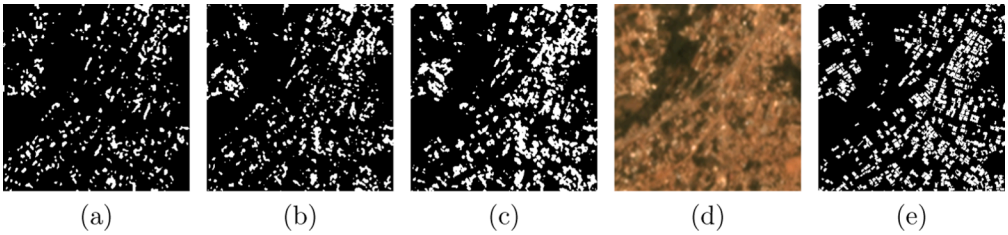


Fig. 9. Examples of building extraction results obtained by different similarity measures. (a) Mutual correlation (Li et al., 2018). (b) Fourier domain correlation (Danelljan et al., 2014). (c) Proposed cross-geolocation attention module. (d) and (e) are Planet satellite imagery and ground reference from Yaounde.

Table 4
Accuracies (%) of different learning methods for building footprint generation on Vienna.

Method	F1 score	IoU
Baseline-t	82.32	69.96
Baseline-a	78.75	64.95
Baseline-a+t	85.02	73.95
Fine-tuning	85.38	74.49
ADVENT(Vu et al., 2019)	81.07	68.17
IntraDA (Pan et al., 2020a)	82.44	70.12
MetaCorrection (Guo et al., 2021b)	83.93	72.31
MoCo (He et al., 2020)	85.66	74.91
DenseCL (Wang et al., 2021)	86.52	76.25
U-Net-AFM (Li et al., 2021)	86.64	76.42
CBRNet (Guo et al., 2022)	86.46	76.09
EPU-Net (Guo et al., 2021a)	86.04	75.50
CGSANet (Chen et al., 2021)	86.59	76.35
CrossGeoNet	**87.51**	**77.79**

co-segmentation methods. Therefore, we further investigate the aforementioned two similarity measures, i.e., mutual correlation (Li et al., 2018) and Fourier domain correlation (Danelljan et al., 2014), to make a comparison with our cross-geolocation attention module.

The statistical results on Yaounde are reported in Table 3. The proposed module outperforms the other two methods by over 1% in statistical metrics. In Fig. 9, the building masks obtained by CrossGeoNet are much closer to ground-truth masks. However, the results provided by Fourier domain correlation show many omitted detection. One reason is that mutual correlation (Li et al., 2018) and Fourier domain correlation (Danelljan et al., 2014) operate on a local neighborhood, leading to the loss of global information. In contrast, our cross-geolocation attention module can capture long-range dependencies, enabling the leverage of useful information from more remote regions in the target image and those from the auxiliary set. This is beneficial to the reduction of semantic noise and the enhancement of semantic information of buildings. Another reason is that these two methods simply concatenate correlation maps with original convolved images to generate new features, while our module updates features by selectively aggregating contexts according to the learned attention maps. By doing so, mutual gains can be achieved through similar features, providing more representative features for building footprint generation.

5. Performance Investigation on Another Data Source

In this section, we further investigate the performance of CrossGeoNet on another dataset, INRIA Aerial Image Labeling data (Maggiori et al., 2017), comprising images captured by airborne sensors. The INRIA dataset is a benchmark dataset, which consists of 360 tiles of aerial imagery. Each aerial image has 5000 × 5000 pixels at a spatial resolution of 30 cm/pixel. In this dataset, only ground reference data for five cities (Austin, Chicago, Kitsap County, Western Tyrol, and Vienna) are made publicly available, and hence we only conduct experiments on these cities. According to the setup in (Bischke et al., 2019), data are split into training and validation sets in our research. We observe that buildings in Vienna have very different structures and sizes in comparison with the other four cities. Therefore, we select Vienna as the target city and the other four cities as the auxiliary set. To verify the effectiveness of CrossGeoNet on INRIA dataset, we make a comparison of different learning methods, i.e., Baseline-t, Baseline-a, Baseline-a+t, fine-tuning, ADVENT (Vu et al., 2019) IntraDA (Pan et al., 2020a), MetaCorrection (Guo et al., 2021b), MoCo (He et al., 2020), DenseCL (Wang et al., 2021), U-Net-AFM (Li et al., 2021), CBRNet (Guo et al., 2022), EPU-Net (Guo et al., 2021a), CSGANet (Chen et al., 2021), and CrossGeoNet. Note the statistics are computed from the validation set of

Table 5
Accuracies (%) of different learning methods on Vienna. Auxiliary and target sets are chosen from Vienna for ensuring similar data distribution.

Method	F1 score	IoU
Baseline-t	78.93	65.19
Baseline-a	81.27	68.45
Baseline-a+t	82.32	69.96
CrossGeoNet	**86.38**	**76.03**

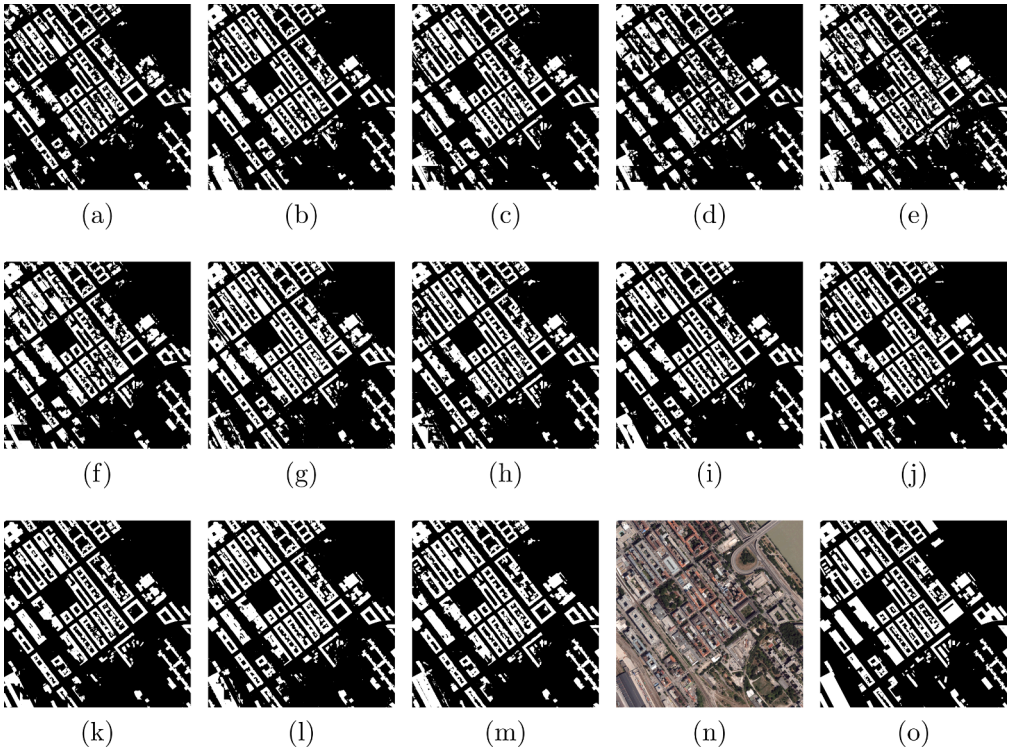


Fig. 10. Examples of building extraction results obtained by different learning methods. (a) Baseline-t. (b) Baseline-a+t. (c) Fine-tuning. (d) ADVENT (Vu et al., 2019). (e) IntraDA (Pan et al., 2020a). (f) MetaCorrection (Guo et al., 2021b). (g) MoCo (He et al., 2020). (h) DenseCL (Wang et al., 2021). (i) U-Net-AFM (Li et al., 2021). (j) CBRNet (Guo et al., 2022), (k) EPU-Net (Guo et al., 2021a). (l) CSGANet (Chen et al., 2021). (m) CrossGeoNet. (n) and (o) are INRIA aerial imagery and ground reference from Vienna.

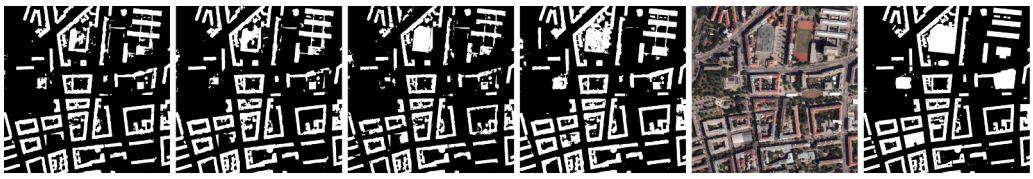


Fig. 11. Examples of building extraction results obtained by different learning methods. (a) Baseline-t. (b) Baseline-a. (c) Baseline-a+t.(d) CrossGeoNet. (e) and (f) are INRIA aerial imagery and ground reference from Vienna. Auxiliary and target sets are chosen from Vienna for ensuring similar data distribution.

Vienna.

We first compare the proposed method against the Baseline-a. It is observed from the statistical results in Table 4, our network obtains increments of 12.84% in IoU. Moreover, CrossGeoNet surpasses Baseline-t by 7.83% in IoU. This indicates that the proposed approach is able to boost the network performance by the joint use of training samples from both the target city and the auxiliary set. From accuracy metrics in Table 4, the proposed method has achieved better performance than other learning methods that aim at transferring the knowledge learned from the auxiliary set to the target city. This demonstrates the effectiveness and robustness of the proposed method for this task, as cross-geolocation co-segmentation learning is able to improve the results on different data sources. When compared with state-of-the-art building extraction methods, CrossGeoNet shows above 1.3% improvement in IoU.

Fig. 10 presents a visual comparison among different learning methods on Vienna. The building footprints generated by CrossGeoNet are more accurate and reliable, as they coincide better with the ground reference when compared with the other methods. For instance, most methods detect only a part of the large building in the bottom left area. In contrast, the proposed approach is capable of accurately capturing a more complete roof outline. Furthermore, for buildings in complex shapes, buildings masks obtained by our network contain more detailed structures, which suggests that CrossGeoNet is still promising in such challenging situations.

In order to investigate the performance of CrossGeoNet when target and auxiliary sets are similar, we have split the original training data of Vienna into two parts, i.e., auxiliary set and target set. Furthermore, we explore the performance of models trained by different learning methods. Specifically, we compare CrossGeoNet with three competitors (i.e., Baseline-t, Baseline-a, and Baseline-a+t) quantitatively and qualitatively. The quantitative results are shown in Table 5. Baseline-t performs poorly than Baseline-a. This is because the number of training patches in the target set is smaller than that in the auxiliary set, which makes it difficult for Baseline-t to achieve good results. Baseline-a+t provides better results than both Baseline-a and Baseline-t, as all training patches are jointly utilized during network learning. It should be noted that CrossGeoNet significantly outperforms Baseline-a+t, with the IoU improved by 6.07%. This demonstrates that our cross-geolocation co-segmentation learning helps to improve model performance. Moreover, this improvement is more significant than that in the case where target and auxiliary sets are less similar. This is because the similarity between target and auxiliary contributes to extracting more generic representations for buildings. Fig. 11 illustrates visual comparisons of different learning methods. Baseline-t and Baseline-a fail to detect some building footprints on the top area. On the contrary, CrossGeoNet is able to alleviate omission errors.

6. Conclusion

Planet satellite imagery holds potentials for generating high-resolution building footprint maps at a large scale. However, generating building footprint maps from Planet satellite imagery is difficult for less developed regions because of the lack of massive annotated samples. Given these issues, we have proposed a novel end-to-end building mapping method, namely CrossGeoNet, aiming at exploring the use of Planet satellite images in detecting buildings on the target city with scarce labeled samples. CrossGeoNet comprises three modules: a Siamese encoder, a cross-geolocation attention module, and a Siamese decoder. More specifically, the encoder is designed to learn features from a pair of images from different geolocations. Afterward, the cross-geolocation attention module learns to encode similarities between them, enabling the capture of a more discriminative and generic representation of the common object (i.e., building in our case). Finally, the decoder exploits the original feature maps and the learned cross-geolocation attention maps to predict building masks. We investigate the proposed approach on two datasets with different spatial resolutions, i.e., Planet dataset (3 m/pixel) and Inria dataset (0.3 m/pixel), which are collected from diverse cities across the globe. Experimental results suggest that the incorporation of the proposed cross-geolocation attention module in co-segmentation learning can offer more satisfactory building footprints than other competitors. Thus, we believe that CrossGeoNet is a robust solution for the task of building footprint generation when dealing with scarce training samples within target cities.

Acknowledgement

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 through the Framework of Helmholtz AI (grant number: ZT-I-PF-5-01) - Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation - Big Data Fusion for Urban Research"(grant number: W2-W3-100), 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) and by German Federal Ministry of Economics and Technology in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C).

References

OpenStreetMap Analytics analysis map. http://osm-analytics.org/#/. Accessed: 2021-08-24.

Asner, G.P., Martin, R.E., Mascaro, J., 2017. Coral reef atoll assessment in the south china sea using planet dove satellites. Remote Sensing in Ecology and Conservation 3, 57–65.

Bischke, B., Helber, P., Folz, J., Borth, D., Dengel, A., 2019. Multi-task learning for segmentation of building footprints with deep neural networks, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 1480–1484.

Chen, S., Shi, W., Zhou, M., Zhang, M., Xuan, Z., 2021. Cgsanet: A contour-guided and local structure-aware encoder–decoder network for accurate building extraction from very high-resolution remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 1526–1542.

Danelljan, M., Häger, G., Khan, F., Felsberg, M., 2014. Accurate scale estimation for robust visual tracking, in: British Machine Vision Conference, Nottingham, September 1-5, 2014, BMVA Press.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Guo, H., Du, B., Zhang, L., Su, X., 2022. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing 183, 240–252.
Guo, H., Shi, Q., Marinoni, A., Du, B., Zhang, L., 2021a. Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. Remote Sensing of Environment 264, 112589.
Guo, X., Yang, C., Li, B., Yuan, Y., 2021b. Metacorrection: Domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation.
He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.
Hou, R., Chang, H., Ma, B., Shan, S., Chen, X., 2019. Cross attention network for few-shot classification. arXiv preprint arXiv:1910.07677.
Houborg, R., McCabe, M.F., 2016. High-resolution ndvi from planet's constellation of earth observing nano-satellites: A new data source for precision agriculture. Remote Sensing 8, 768.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132–7141.
Huang, X., Cao, Y., Li, J., 2020. An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images. Remote Sensing of Environment 244, 111802.
Ivanovsky, L., Khryashchev, V., Pavlov, V., Ostrovskaya, A., 2019. Building detection on aerial images using u-net neural networks, in: 2019 24th Conference of Open Innovations Association (FRUCT), IEEE. pp. 116–122.
Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning aerial image segmentation from online maps. IEEE Transactions on Geoscience and Remote Sensing 55, 6054–6068.
Li, Q., Mou, L., Hua, Y., Shi, Y., Zhu, X.X., 2021. Building footprint generation through convolutional neural networks with attraction field representation. IEEE Transactions on Geoscience and Remote Sensing.
Li, Q., Shi, Y., Auer, S., Roschlaub, R., Möst, K., Schmitt, M., Glock, C., Zhu, X.X., 2020. Detection of undocumented building constructions from official geodata using a convolutional neural network. Remote Sensing 12, 3537.
Li, Q., Shi, Y., Huang, X., Zhu, X.X., 2020. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (fpcrf). IEEE Transactions on Geoscience and Remote Sensing.
Li, W., Jafari, O.H., Rother, C., 2018. Deep object co-segmentation, in: Asian Conference on Computer Vision, Springer. pp. 638–653.
Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2016. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Transactions on Geoscience and Remote Sensing 55, 645–657.
Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize benchmark to any city? the inria aerial image labeling, in: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE.
Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S., 2020a. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision, in: IEEE Conference on Computer Vision and Pattern Recoginition (CVPR).
Pan, Z., Xu, J., Guo, Y., Hu, Y., Wang, G., 2020b. Deep learning segmentation and classification for urban village using a worldview satellite image based on u-net. Remote Sensing 12, 1574.
Papoutsakis, K., Panagiotakis, C., Argyros, A.A., 2017. Temporal action co-segmentation in 3d motion capture data and videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6827–6836.
Shi, Y., Li, Q., Zhu, X.X., 2020. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. ISPRS Journal of Photogrammetry and Remote Sensing 159, 184–197.
Tonbul, H., Kavzoglu, T., 2020. Semi-automatic building extraction from worldview-2 imagery using taguchi optimization. Photogrammetric Engineering & Remote Sensing 86, 547–555.
Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P., 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2517–2526.
Wang, W., Lu, X., Shen, J., Crandall, D.J., Shao, L., 2019. Zero-shot video object segmentation via attentive graph neural networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9236–9245.
Wang, X., Zhang, R., Shen, C., Kong, T., Li, L., 2021. Dense contrastive learning for self-supervised visual pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3024–3033.