A Diverse Large-Scale Building Dataset and a Novel Plug-and-Play Domain Generalization Method for Building Extraction

Muying Luo, Shunping Ji, Senior Member, IEEE, and Shiqing Wei, Member, IEEE

Abstract—In this article, we introduce a new building dataset and propose a novel domain generalization method to facilitate the development of building extraction from high-resolution remote sensing images. The problem with the current building datasets involves that they lack diversity to train a practical learning model with good generalization ability, and the quality of the labels is unsatisfactory. To address these issues, we built a diverse, large-scale, and high-quality building dataset named the WHU-Mix building dataset, which is more practice-oriented. The WHU-Mix building dataset consists of a training/validation set containing 43,727 diverse images collected from all over the world, and a test set containing 8,402 images from five other cities on five continents. In addition, to further improve the generalization ability of a building extraction model, we propose a domain generalization method named batch style mixing (BSM), which can be embedded as an efficient plug-and-play module in the front-end of a building extraction model, providing the model with a progressively larger data distribution to learn data-invariant knowledge. The experiments conducted in this article confirmed the potential of the WHU-Mix building dataset to improve the performance of a building extraction model, resulting in a 6–36% improvement in mean intersection over union (mIoU), compared to the other existing datasets. The adverse impact of the inaccurate labels in the other datasets can cause about 20% intersection over union decrease. The experiments also confirmed the high performance of the proposed BSM module in enhancing the generalization ability and robustness of a model, exceeding the baseline model without domain generalization by 13% and the recent domain generalization methods by 4–15% in mIoU. The WHU-Mix dataset is available at http://gpcv.whu.edu.cn/data.

Index Terms—Building extraction, deep learning, domain generalization, remote sensing building dataset.

I. INTRODUCTION

AUTOMATIC building extraction from high-resolution remote sensing images can facilitate a wide range of applications, including topographic map making, population density estimation, land-use management, and urban planning [1], [2], [3]. The development of deep learning techniques, especially convolutional neural networks (CNNs), has greatly boosted the performance in various image and vision based tasks [4], [5], including building extraction [6], [7], [8], [9], [10].

Nevertheless, there is still a prominent issue that should be further addressed: the questionable generalization ability of the deep learning models when faced with versatile building structures and architectural styles from all over the world, in addition to the various remote sensing images acquired from different sensors, in diverse weather conditions and different seasons [11], [12]. This problem is recognized to be the biggest challenge currently being faced, and it far surpasses the improvement of structures and parameters in a deep learning based building extraction model.

To handle this generalization problem, large-scale building datasets with various building structures and styles, collected under different acquisition conditions, are urgently needed [13], [14]. On the one hand, the current deep learning based methods are highly data-driven, requiring large-scale and diverse training data to achieve good generalization. Early studies [15], [16] have shown a model’s performance gaps between two datasets with different data distributions. Increasing the diversity of the training dataset allows the model to better learn invariant and robust features. On the other hand, the test data should have a certain distribution difference with the training data (e.g., training and test data from different cities), to more properly and practically evaluate the performance of building extraction models. To date, several remote sensing building datasets, such as the Massachusetts buildings dataset [17], the WHU building dataset (WHU) [18], and the Inria aerial image labeling dataset [19], have been made available openly and freely. However, most of these datasets suffer from several drawbacks. The construction of these datasets, to some extent, caters to the requirement of deep learning methods for the training and test data to share the same or similar data distributions. For example, the training and test data are from the same cities in the Massachusetts buildings dataset and WHU aerial building dataset [18], and thus, lack diversity. Although the Inria dataset provides a training set and a test set from different cities, the labels of the test set are not openly available. As a result, most studies [6], [10], [18], [20] have selected a proportion of the images from each city in the training set for the training, and used the rest for the validation. Nevertheless, the well-performing deep learning models on these datasets can suffer a severe performance drop.
with target data, which have a quite different distribution to the training data [18], [21], [22], [23], i.e., out-of-distribution data. Generally speaking, these datasets lack the ability to serve as benchmarks to both train a building extraction model with strong generalization ability and properly evaluate the performance of a model in a more practical scene where the target data are from other cities and are independent of a precreated training set.

The other major drawback is many datasets, including Massachusetts, Inria, SpaceNet challenge dataset [24], and OpenAI dataset [25], suffer from some low-quality labels, which greatly impacts the correctness and fairness of evaluating a building extraction model. For example, we discovered that there are about 20% intersection over union (IoU) score gap between using the original Inria and the one repaired by us.

The generalization problem cannot be fully solved by only increasing the quantity of the datasets, because images captured over the broad earth’s surface and under ever-changing atmospheric conditions, various sensors, and seasons form a vast distribution space [26], [27]. This situation has impelled researchers to look for algorithmic solutions. A commonly used methodology is called domain generalization [28], [29], i.e., training a model using the limited training data to achieve better generalization for out-of-distribution data. One type of domain generalization methods, the image/data-level methods, which are often called data augmentation methods, attempt to simulate more data from the training data to enlarge the data distribution. The other type of methods, the feature-level methods, attempt to learn robust and invariant features in a task-specific model [30], [31] so that the model will be more effective on the unseen data. Nevertheless, most domain generalization methods, especially the feature-level methods developed in the computer vision community are designed for the situation where there are several given domains with specific domain labels. For example, a domain of real street scenarios and another of simulated street scenarios are provided in the GTAS [32] and Cityscapes dataset [33], and four domains of photo, art, cartoon, and sketch are provided in the PACS dataset [34]. However, this is rarely the case in remote sensing image processing, where images collected from optical cameras cannot be easily divided into different domains. To date, except for the commonly used data augmentation techniques, such as Gaussian blur and rotation, and few complex feature-level methods [11], there is a lack of widely recognized domain generalization baselines specially for handling complex remote sensing images.

In order to handle the abovementioned challenges in generalization, we not only built a large-scale and diverse building dataset collection, but also propose a novel domain generalization method to enhance the generalization ability of deep learning based building extraction models.

The building dataset collection, which we have named the WHU-Mix building dataset, is intended to be adaptable to practical situations, as much as possible. The dataset collection includes a training/validation (trainval) set with diverse building structures and styles from different regions of the globe, and a test set also containing diverse data, but from other global cities. The trainval set consists of 43 727 image tiles each with 512 × 512 pixels, collected from various geographical areas, times, and sensors. The test set contains 8402 512 × 512 image tiles from five cities on five different continents. The WHU-Mix building dataset is larger than the WHU aerial building dataset [18], which contains about 8000 tiles of the same size, but is smaller than the CrowdAI dataset [35], which has 340 000 tiles. Nevertheless, the new dataset features “diversity,” which was found to be much more important than quantity in this article.

We also developed a novel, stable, and plug-and-play domain generalization method, named batch style mixing (BSM). The BSM method, as a portable component, dynamically simulates and generates new diverse data to expand the data distribution in a learnable manner during the training of a building extraction model, so that the model can learn more domain-invariant information and perform better on unseen data. In fact, it could be incorporated in any deep learning based building extraction model, because BSM is decoupled from the target model.

The main contributions of this article are summarized as follows:

1) We introduce an open-source, large-scale, diverse, and highly accurate building dataset, named the WHU-Mix building dataset, which guides remote sensing building extraction research toward more practical applications. When trained with the WHU-Mix building dataset, the same building extraction model obtained an improvement in mean intersection over union (mIoU) score of between 6.43% and 36.52%, compared to the same model trained on the other existing building datasets. In addition, we found and repaired many wrong labels in some of the existing datasets, and proved that these wrong labels can significantly reduce the model performance.

2) We propose a domain generalization method—BSM—to further improve the generalization and robustness of building extraction models. BSM was developed through a combination of the classic data augmentation methods and a new style mixing method. Although style transfer has recently used in the domain adaptation of paired images in the computer vision and remote sensing communities, its potential in domain generalization has not yet been explored by the two communities. Introducing the style transfer on a batch of training samples with a mixing technology, and combining it with classic data augmentation to expand the data distribution space for domain generalization is the brand-new idea of this article. Specifically, BSM is placed at the front-end of a deep learning model to handle the complex generalization problem of remote sensing images, and can be easily imbedded in the existing building extraction models. The experiments indicated that the proposed domain generalization method can improve the performance of a deep learning model by 13.17% in mIoU on the diverse test set of the WHU-Mix building dataset.

II. RELATED WORK

A. Building Extraction Datasets of High-Resolution Remote Sensing Images

Over the past years, several building extraction datasets have been publicly released by different research groups (see Table I). These datasets are briefly reviewed as follows.
TABLE I
SEVEN PUBLICLY AVAILABLE DATASETS AND THE WHU-MIX BUILDING DATASET BUILT IN THIS ARTICLE. “∼” DENOTES “ABOUT”

| Dataset                                    | Subsets                | Number of image tiles | Image size          | Image spatial resolution (m) | Area (km²) | Image type |
|--------------------------------------------|------------------------|-----------------------|---------------------|-----------------------------|------------|------------|
| CrowdAI mapping challenge dataset (2018)   |                        | 341,058               | 300 × 300           | unknown                     | /          | satellite  |
| Massachusetts buildings dataset (2013)     |                        | 151                   | 1,500 × 1,500       | 1                           | 340        | aerial     |
| Open AI dataset (2018)                     | Aerial imagery dataset | 8,188                 | 0.3                 | 193                         | 102        | aerial     |
| WHU building dataset (2016)                | Satellite dataset I (global cities) | 204             | 512 × 512           | 0.3–2.5                     | 5          | satellite  |
|                                            | Satellite dataset II (East Asia) | 17,388             | 0.35                | 558                         |            | satellite  |
| SpaceNet challenge dataset (2017)          | Rio de Janeiro         | 6,940                 | 438 × 406           | 0.5                         | 308        | aerial     |
|                                            | Vegas                  | 3,851                 |                     | 146                         |            |            |
|                                            | Paris                  | 1,148                 | 650 × 650           | 0.3                         | 44         | satellite  |
|                                            | Shanghai               | 4,582                 |                     | 174                         |            |            |
|                                            | Khartoum               | 1,012                 |                     | 38                          |            |            |
| Learning Aerial Image Segmentation From Online Maps dataset (2017) | Berlin             | 200                   |                     | 10                          |            |            |
|                                            | Chicago                | 497                   |                     | 51                          |            |            |
|                                            | Paris                  | 625                   | −3000 × 3000        | −0.1                        | 60         | aerial     |
|                                            | Potsdam                | 24                    |                     | 2                           |            |            |
|                                            | Zurich                 | 375                   |                     | 36                          |            |            |
| Inria aerial image labeling dataset (2017)  | Training set (Austin, Chicago, Kitsap County, Western Tyrol, Vienna) | 180             | 5000 × 5000         | 0.3                         | 405        | aerial     |
| WHU-Mix building dataset                   | Trainval set           | 43,727                | 512 × 512           | 0.091–2.5                   | 1080       | aerial/satellite |
|                                            | Test set               | 7718                  |                     | 133                         |            |            |

1) **CrowdAI mapping challenge dataset**: The CrowdAI dataset [35] is a very large scale building extraction dataset consisting of 341,058 300 × 300 images collected from Google Earth. The annotation accuracy of this dataset is high. However, we found that the images in this dataset tend to be acquired in the same local region, and thus, lack diversity.

2) **Massachusetts buildings dataset**: The Massachusetts buildings dataset [17] is an early building extraction dataset consisting of 804,150 1500 × 1500 aerial images acquired in the Boston area of the U.S. The images have a relatively low resolution of 1 m, and there are some obvious annotation errors.

3) **OpenAI dataset**: The OpenAI dataset [25] was released as part of the 2018 OpenAI Tanzania Building Footprint Segmentation Challenge Competition. The images were collected in Tanzania by unmanned aerial vehicle, with a very high resolution of about 0.07 m. However, there are a large number of errors in the annotations.

4) **WHU building dataset**: The WHU building dataset [18] is a multisource building extraction dataset providing high-quality labels, which consists of three subsets. Among these subsets, the aerial imagery subset is the most widely used for building extraction in the remote sensing community.

5) **SpaceNet challenge dataset**: The SpaceNet challenge dataset [24] was designed for the task of building extraction in satellite images. The dataset consists of images collected from five cities on five continents. Unfortunately, the annotation accuracy is low, and the dataset cannot be used to properly evaluate the performance of a building extraction model.

6) **Learning Aerial Image Segmentation From Online Maps (LAIS) dataset**: The LAIS dataset [36] was designed to evaluate the performance of deep learning based models when trained with noisy data. Three land-cover classes are annotated: building, street, and background. This dataset contains images of four cities, which were downloaded from Google Maps, and their corresponding relatively inaccurate annotations obtained from OpenStreetMap, with a displacement as large as 10 pixels. This dataset also provides relabeled Potsdam images from the ISPRS two-dimensional Semantic Labeling Contest dataset [37].

7) **Inria aerial image labeling dataset**: The Inria dataset [19] was released to benchmark the generalization ability of different building extraction methods, for a similar purpose as this article. It contains a training set and a test set, and the images of the two sets are from different cities. However, as the annotations of the test set are not openly available from the provider, most studies have divided the training set for the training and testing [10], [20], resulting in the dataset having a limited ability to properly evaluate a model’s generalization capability. Moreover, there are many obvious annotation errors in the dataset, which have an adverse impact on the training and testing.

As a summary, although these datasets have greatly facilitated the development of building extraction study, the lack of diversity and the low-quality labels are the main concerns of the existing datasets.

**B. Domain Generalization**

The domain shift problem poses challenges for cross-domain recognition tasks, e.g., the land-cover classification of
multisource or multitemporal remote sensing images, leading to a significant performance drop for the data-driven deep learning models. To narrow the discrepancy across domains, there are two main approaches. Domain adaptation [38] involves using both the labeled training data in the source domain and the unlabeled test data in the target domain in the training process, to learn a target-adapted model. Differing from domain adaptation, domain generalization [39] is independent of target domains. It attempts to make the model learn domain-invariant knowledge from only the labeled source domain, and is performed directly on an arbitrary target domain that has not been seen before. Therefore, compared to domain adaptation, domain generalization is more challenging because no knowledge from the target domain can be accessed. However, domain generalization has more practical value because there is no need to repeat the training process of a model every time for different target domains, especially when the source domain is an extraordinarily large dataset (e.g., ImageNet [40]).

As an image/data-level method, data augmentation is one of the important categories of domain generalization [39]. Data augmentation involves manipulating the training data to generate diverse samples, to help the model learn domain-invariant knowledge. The classic data augmentation operation is the simplest, cheapest, and most common way to prevent overfitting and improve the generalization ability of a deep learning model. Geometric data augmentation usually includes flipping, rotation, scaling, cropping, and so on. Color augmentation usually consists of color jitter and adding random noise, to enhance the robustness of a model to image color variations.

However, the classic data augmentation operations are fundamentally linear transformations restricted by the range of the transformation parameters, leading to a limited ability to augment images. Some recent augmentation methods introduce a style transfer technique to generate more diverse and rich data. Nevertheless, most style transfer based methods require auxiliary datasets as the augmentation reference, or require known different source domain categories within the dataset to be the references for each other. For example, the domain randomization and pyramid consistency method [41] applies a generative adversarial network (GAN) based style transfer network named CycleGAN [42] to synthesize images, randomly transferring the styles of real images to have similar visual appearances to ImageNet classes. Borlino et al. [43] applied an off-the-shelf feedforward stylization based method known as Adaptive Instance Normalization (AdaIN) [44] to transfer styles between different domains. Mixup [45] involves performing linear interpolation between any two image pairs and their labels to generate a continuous data distribution. Inspired by Mixup, Mixstyle [46] mixes deep feature statistics of the training images to synthesize new domains.

In contrast to data augmentation, in which the model is expected to learn from the augmented diverse data, domain generalization at the feature level attempts to provide the domain-invariant deep features of the training data for a task model. Therefore, the feature-level domain generalization must be combined with a task model. Some earlier studies attempted to align the features of the source domains by explicitly minimizing the distance function between the source domain feature distributions. Examples of such methods are maximum mean discrepancy [31], first/second order correlation [47], and the linear-dependency domain generalization (LDDG) method [48], which uses Kullback–Leibler divergence as a linear-dependency regularization term. Other studies have adopted the adversarial learning technique [49] to align the features of the source domains implicitly. For example, Li et al. [50] proposed a conditional invariant adversarial network for domain-invariant representation learning. Matsuura and Harada [51] proposed a method named the mixture of multiple latent domains (MMLD), which performs adversarial training using data of latent domains divided by iterative clustering. Recently, the feature normalization technique has been introduced to learn domain-invariant features. IBN-Net [52] combines instance normalization and batch normalization to enhance the generalization and learning ability of a deep learning model. Following the work of IBN-Net, the RobustNet method [53] introduces an instance selective whitening loss, which separates the domain-specific and domain-invariant features by comparing the features of the images before and after color augmentation (i.e., photometric transformation), and selectively suppresses the domain-specific features. Tang et al. [11] used a vector, named a “capsule,” to store the features of a building and its constituent parts during training to improve the generalization ability of the network. However, as the feature-level domain generalization methods must be coupled with a specific task model, they lack flexibility. Furthermore, as the learned features are in fact only invariant to the source domains, their reliability in unseen domains is not fully guaranteed, and will likely result in an unstable performance.

In addition, some general machine learning strategies have also been explored for some parts of domain generalization. The main strategies include meta-learning [54], [55], ensemble learning [56], and self-supervised learning [57].

It should be noted that most of the abovementioned domain generalization methods [31], [41], [43], [47], [50], [54], [56], [57] developed in the computer vision community can only be applied when there are several defined domains in the training data. Therefore, these methods cannot be implemented for most remote sensing image datasets without domain classification. There is therefore an urgent need to develop a new domain generalization baseline that exceeds the conventional geometric and color augmentation for remote sensing image processing, including building extraction.

III. WHU-MIX BUILDING DATASET

Although several building extraction datasets have been publicly released, as reviewed in Section II-A, and have contributed to the development of building extraction methods, the diversity they represent, including the building styles and geographic distributions, is not perfect, and the quality of some datasets is deeply affected by nonignorable label errors, both of which have hindered the development of further studies of building
The WHU-Mix building dataset consists of 51,445 512 image tiles with 0.09–2.5 m spatial resolutions, captured by a wide range of aerial and satellite sensors from about 58 cities worldwide, as shown in Fig. 1, covering an area of about 1213 km². It is much larger than most of the other datasets listed in Table I. The different sources and imaging conditions, different spatial resolutions, and different buildings styles of the local regions all contribute to the diversity of the sample distribution (see Fig. 2) of the dataset, which is crucial to improve the generalization ability of deep learning based models [58]. The WHU-Mix building dataset is smaller than the CrowdAI dataset [35]; however, in this article, we demonstrate that the newly built dataset significantly surpasses the CrowdAI dataset, and prove that diversity is more important than quantity.

2) \textit{High accuracy and uniform format}: The WHU-Mix building dataset provides high-accuracy labels. For the labeled images from the existing datasets, we checked the labels and corrected the labeling errors. For the newly acquired images without labels, we labeled them manually and then double-checked the labeling. The WHU-Mix building dataset also provides a uniform format for the data collected from different sources with different formats. The image tiles and corresponding building label maps are stored in TIFF format. The building label maps are Boolean maps, where the building pixels and backgrounds are denoted as 255 and 0, respectively. The images and labels of the trainval and test sets are stored in corresponding folders and named with the uniform “city + number” rule.

\section{Construction of the WHU-Mix Building Dataset}

Although, we collected some of the samples from the existing datasets, the whole construction process for the WHU-Mix building dataset took more than six months. The WHU building dataset [18] is directly included in the WHU-Mix building dataset. Some parts of three other datasets, i.e., Inria [19], SpaceNet [24], and LAIS [36], which feature some erroneous annotations, were edited and corrected. Finally, new remote sensing data were collected and manually labeled.

1) \textit{Collection of the existing datasets}: The WHU building dataset [18] is a widely used dataset with high-accuracy building maps. We incorporated the WHU building dataset as a part of the trainval set of the WHU-Mix building dataset. Specifically, for the images from the aerial imagery dataset and satellite dataset I (global cities), we included them in the proposed dataset directly. For satellite dataset II (East Asia), as the images were captured in an area with sparse buildings, 13350 out of the 17388 512 × 512 image tiles have only backgrounds. We, therefore, removed them and kept the remaining 4038 images.

2) \textit{Revision of the existing datasets}: We revised the labels for the Kitsap County, Western Tyrol, and Vienna images from the Inria dataset [19], the Khartoum images from the SpaceNet dataset [24], and the Potsdam images from the LAIS dataset [36]. There are a great many obvious errors in the building labels of these datasets, including non-existing, missing, displaced, and inaccurate delineation (see Fig. 3). To construct

\begin{table}[h]
\centering
\caption{Components of the WHU-Mix Building Dataset, Including the Trainval Set Consisting of Data From the WHU Building Dataset [18], a Subset of the Inria Dataset [19], and New Samples, and a Test Set Consisting of a Five-City Dataset From Five Continents}
\begin{tabular}{llll}
\hline
Division & Dataset & Number of image tiles & Image spatial resolution (m) \\
\hline
Trainval set & WHU building dataset & 12,430 & 0.3–2.5 \\
& Subset of edited Inria (Western Tyrol, Vienna) & 5832 & 0.3 \\
& New samples & 25,465 & 0.15–0.64 \\
Test set & US_Kitsap & 3600 & 0.3 \\
& CN_Wuxi & 1409 & 0.17 \\
& NZ_Dunedin & 1312 & 0.2 \\
& DE_Potsdam & 1176 & 0.09 \\
& SD_Khartoum & 905 & 0.3 \\
\hline
\end{tabular}
\end{table}
Fig. 1. Geographic distribution of the images of the WHU-Mix building dataset. The green dots and red dots represent images of the trainval set and test set, respectively. The global cities in the original WHU building dataset consist of images from about 45 cities.

Fig. 2. Examples of various images and the corresponding labels (cyan mask) in the WHU-Mix building dataset. (a) Christchurch. (b) Cairo. (c) Santiago. (d) Western tyrol. (e) Vienna. (f) Chongqing. (g) Hubei. (h) Hangzhou. (i) Tianjin. (j) Kitsap county. (k) Wuxi. (l) Dunedin. (m) Potsdam. (n) Khartoum.
Fig. 3. Example images of the original datasets and the revised images. The left of each column shows the original labels, and the right shows the edited labels.

a dataset with high-quality labels, we first converted the different label formats of all the datasets into vector format. We then manually corrected the outlines of all the building samples using ArcGIS 10.2. After the revision, we converted the edited vector maps back to the uniform raster format. Finally, we added the data of Western Tyrol and Vienna into the trainval set of the WHU-Mix building dataset, and included the data of Kitsap County, Khartoum, and Potsdam in the test set.

3) New samples: A large amount of new samples was collected to enrich the trainval set. The samples were made up of 25 465 512 × 512 diverse image tiles acquired in the cities of Chongqing, Tianjin, and Hangzhou, and Hubei province in western, northern, eastern, and central China, respectively, from various remote sensing resources. Among the different images, the images of Chongqing have a resolution of 0.15 m and a total coverage area of 32 km². The buildings in Chongqing are mostly built on the mountainsides, with varied shapes. The images of Tianjin cover a 3 km² area with a 0.1 m ground resolution, and the building types are dominated by large factories and residential areas. The images of Hangzhou have a ground resolution of 0.2 m and cover an area of 156 km² in the rural area, where the buildings are sparsely distributed, and water bodies and farmland are the main land-cover types. The images of Hubei province have a resolution of 0.64 m and cover a total area of 427 km² in several rural and urban regions of Hubei province. The buildings in both the rural and urban regions are densely distributed; however, the buildings in the rural regions are basically low-rise buildings, while the buildings in the urban regions are typically multistory buildings.

We also collected data from two cities—Wuxi in China and Dunedin in New Zealand—for inclusion in the test set, so that the test set contains images of five typical cities from five continents. The data of Wuxi (named CN_Wuxi in Table II) were collected from Google Earth, and the data of Dunedin (named NZ_Dunedin in Table II) were acquired from the Land Information service of New Zealand [59].

For the abovementioned newly acquired remote sensing data, we manually delineated the vector building maps in ArcGIS 10.2. To ensure the label quality and avoid misjudgments, we relied on one additional annotator for a complete check, and three annotators for random checks of a proportion of around 20%, respectively. Finally, we converted the vector building maps to raster maps. Examples of the new samples in the WHU-Mix building dataset are shown in Fig. 2(f)–(i), (k), and (l).

IV. METHODOLOGY

In addition to providing a diverse building dataset, the other main contribution of this article is that we propose a simple but very effective and efficient plug-and-play domain generalization method for boosting the performance of a modern deep learning based building extraction model on various target remote sensing images. The proposed domain generalization method uses only the available training datasets to enable the building extraction model to learn data-invariant features. That is to say, it is not

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
necessary to train a new model every time for different target images, as is required with the domain adaptation methods. The pretrained model obtained by the proposed method can be directly applied to target images captured from different sensors, times, regions, and imaging conditions. In this section, we first introduce the basic idea of the proposed method, and then describe the implementation details.

### A. Proposed BSM Domain Generalization Method

The proposed image-level domain generalization method, which we call BSM, can be easily plugged into the front-end of a deep learning based building extraction model. BSM draws inspiration from the idea of style transfer for domain adaptation [60],[61]. However, in the domain generalization problem, style transfer is used to generate diverse new images and expand the sample distribution space, instead of aligning the styles of the source images in the training set and the target images to be processed in a domain adaptation problem. Specifically, we propose generating diverse new samples in an ever-changing batch of the training data in the training process, where the images in the batch are first processed by random data augmentation, and their styles are then mixed (and, thus, expanded) by the idea of mixing [45]. The workflow of BSM is shown in Fig. 4, which includes three main steps: geometric and color augmentation, shuffling, and a style transfer network for obtaining the mixed-style new samples.

As shown in Fig. 4, for a batch X with a given batch size (here 4), some common geometric and color augmentations are performed on the images in X to obtain a content batch Xc. The images in Xc are then shuffled along the batch dimension to obtain a style batch Xs. The shuffling operation prepares pairs of images (with the same order numbers), for which the styles will next be mixed with each other. The content and style batches Xc and Xs are both fed into the style transfer network. fc and fs are the feature outputs from the encoder of the style transfer network corresponding to Xc and Xs. We choose AdaIN [44] to process fc and fs, i.e., to re-normalize fc by replacing its channelwise mean and standard deviation with that of fs. In fact, AdaIN operates in the same way as the Wallis filter [62], but on the feature maps. As a result, the style information represented by feature fc is the same or similar to fs, but the content information remains unchanged. fc can be denoted as

\[
fc = \frac{1}{\sigma(f)} \left( \frac{f - \mu(f)}{\sigma(f)} \right) + \mu(f) \tag{1}
\]

where \(\sigma(*)\) and \(\mu(*)\) are the channelwise mean and standard deviation of the features. Then, as linear interpolation has been shown to be an effective way to mix samples [45], we also perform linear interpolation between the feature maps fc and fs for the mixing styles. Feature fccs is then obtained, as shown in

\[
f_{ccs} = \alpha f_{cs} + (1 - \alpha) f_{c} \tag{2}
\]

where the tradeoff parameter \(\alpha\) determines the degree of mixing of the styles, which was set to 0.5 in this article. Finally, the feature fccs is fed into the decoder of the style transfer network to output a mixed-style batch Xccs. The ith image in Xccs has the same semantic content as the ith image in Xc, and its style is a mixture of the ith images in Xc and Xs. The images in a batch

![Overview of batch style mixing](image-url)
are constantly changing during the training, which ensures that the combination of data augmentation and mixing creates a huge distribution space.

B. Implementation of BSM for Building Extraction

1) Building extraction framework with BSM: The overall building extraction framework with the proposed domain generalization method is shown in Fig. 5. The framework consists of two parts: the front-end BSM for domain generalization, as described in Fig. 4, and a back-end building extraction model, which can be an arbitrary semantic segmentation model. In this article, we applied a modified model named AT-MAFCN.

The training process is shown in the dashed box in Fig. 5. During the training process, the training data are first fed into the BSM module. Specifically, in each iteration, a batch \(X\) (with batch size 16) is sequentially forwarded from the training set to the geometric augmentation, color augmentation, and style mixing, to obtain an augmented batch \(X_{\text{aug}}\). The operations in the three submodules of the BSM module are each activated with a probability of 0.5. For the geometric augmentation, random horizontal flip, random vertical flip, random rotation within 30°, random scaling of the image size within 0.5–2.0, and cropping (or filling) to the original size are applied. The color augmentation includes the adjustment of the brightness, color balance, contrast and sharpness, and random Gaussian blur.

The augmented images in batch \(X_{\text{aug}}\) are then input into the building extraction model \(F\) to complete the standard training process and obtain the prediction maps \(f_F(X_{\text{aug}})\). As the number of iterations increases, BSM provides a progressively larger and larger data distribution for the building extraction model to learn data-invariant features. Cross-entropy loss is used to calculate the pixel-level classification loss between \(f_F(X_{\text{aug}})\) and the corresponding label \(Y_s\). For each image \(x_{\text{aug}}\) in \(X_{\text{aug}}\), the loss is calculated as follows:

\[
L_{\text{ce}}(x_{\text{aug}}) = - \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} y_s^{(h,w,c)} \log (f_F(x_{\text{aug}}))^{(h,w,c)}
\]

where \(y_s\) is the corresponding label of \(x_{\text{aug}}\), \(H\) and \(W\) are the height and width of the prediction map \(f_F(x_{\text{aug}})\), and \(C\) is the number of categories (for building extraction, \(C = 2\)).

The building extraction model trained along with BSM, thus, has a strong generalization capability for images that do not match the distribution of the training set. The test process is depicted in the solid box in Fig. 5. The various test images are fed into the pretrained building extraction model without BSM, to output the prediction maps.

2) The implemented building extraction model: The implemented building extraction model—AT-MAFCN—is modified from a multiscale semantic segmentation network known as MA-FCN [7]. The structure of AT-MAFCN is illustrated in Fig. 6. We employ the same VGG-16 [63] encoder as MA-FCN. There are three main differences between AT-MAFCN and MA-FCN.

First, two parallel attention modules are placed between the last layer of the encoder and the first layer of the decoder, one of which is the channel attention module and the other is the position attention module [64], to enhance the representation ability of the deep fully convolutional network (FCN) structure.

Second, we believe that learning from the output feature maps at a wider scale can minimize the impact of different
sizes of buildings and various resolutions of images. Therefore, the output feature maps of all five scales in the decoder are upsampled, instead of the last four scales, as in the original MA-FCN. The pixel-level classification loss is then calculated using the output feature maps obtained from scales (1/16, 1/8, 1/4, 1/2, and 1/1) with corresponding weights [0.25, 0.25, 0.25, 0.25, 0.5] in the training process. Equation (4) is the final loss

$$L_{ce}(x_{aug}) = 0.25 \sum_{i=1}^{4} L_i + 0.5L_5$$

where \(L_i \ (i = 1, 2, 3, 4, 5)\) represent the losses of the five prediction maps at scales (1/16, 1/8, 1/4, 1/2, and 1/1), respectively. \(L_i\) is implemented, as shown in (3).

Third, in the test process, the prediction map is only output from the feature of the last layer (scale 1/1), instead of the concatenated feature maps of five layers in the decoder. This is because, after the multiscale training, the prediction map from the last layer is what we really want, and maps with other scales are not necessary in the testing, and the prediction efficiency is increased.

V. EXPERIMENTAL RESULTS AND ANALYSIS

Experiments are conducted to evaluate the effectiveness of the proposed dataset and method. First, the training settings and the evaluation metrics used in the experiments are described. Second, the experiments conducted in this article to evaluate the WHU-Mix building dataset are reported. Finally, the experimental evaluation of the proposed domain generalization method is provided.

A. Experimental Setup

1) Training settings: In all the experiments, we used a single NVIDIA RTX A6000 48 GB GPU and an Intel Xeon Silver 4109T CPU. Among the building extraction models used in the experiments, AT-MAFCN and UNet [65] employed the VGG-16 [63] backbone pretrained on the ImageNet dataset [40] in the training process, and DeepLab V3+ [66] employed the ResNet-101 [4] backbone pretrained on the ImageNet dataset. The structures and weights of the encoder and decoder of the style transfer network (as shown in Fig. 4) were the same as those in Huang and Belongie [44]. The building extraction models were trained by the Adam optimizer [67], with a weight decay of \(5 \times 10^{-5}\) and an initial learning rate of \(1 \times 10^{-4}\). The learning rate was dynamically adjusted by a poly learning strategy with decay parameter of 0.9. To fit the GPU capacity, the batch size was set to 16, and the images used in the experiments were uniformly cropped into the size of 512 \(\times\) 512 pixels. During the test process, the images were fed into the trained model with a batch size of 1.

2) Evaluation metrics: The test set of the WHU-Mix building dataset was used to assess the building extraction accuracy. The IoU for the test data of the five cities was computed. The IoU score is the ratio between the intersection and union of the true positive pixels and the building pixels predicted by the model, which can be defined as follows:

$$\text{IoU} = \frac{TP}{TP + FP + FN}$$

where TP, FP, and FN are the number of true positive, false positive, and false negative pixels, respectively.

As a global metric, the mIoU over all the images in the five cities of the test set is reported. The values of IoU and mIoU are in the range of 0 to 1 (they are reported as percentages (%) in the experiments). The closer the values are to 1, the better the building extraction results. In addition, as an auxiliary indicator, the training times of the building extraction models using the different datasets and domain generalization methods are also reported.

B. Evaluation of the WHU-Mix Building Dataset

The first experiment compared the performance of AT-MAFCN, pretrained on different building extraction datasets, on the test set of the WHU-Mix building dataset, to evaluate the impact of the size and diversity of the dataset. The second experiment was dedicated to the evaluation of the influence of data quality on the building extraction performance.

1) Comparison with other datasets: In this section, we describe how we trained AT-MAFCN by utilizing different building datasets, and evaluated the trained models on the test set of the WHU-Mix building dataset. The details of AT-MAFCN are described in Section IV-B. The compared building datasets were the Massachusetts buildings dataset [17], the Inria dataset [19] (with Kitsap County excluded), the CrowdAI dataset [35], the WHU building dataset [18], and the newly built WHU-Mix building dataset (trainval set). As the geographical regions of the WHU-Mix test set are far from the regions of all of the above-mentioned datasets, the results obtained can be used to fairly measure the ability of the various datasets to handle out-of-distribution data.

The results are presented in Table III. The highest mIoU is achieved with the WHU-Mix building dataset (trainval set), which can be attributed to the diversity and large scale of the WHU-Mix building dataset, enabling the model to learn domain-invariant and robust features that facilitate the extraction of buildings that have not been seen before. The Massachusetts buildings dataset gives the most unsatisfactory result, which means that a dataset of both a very small size and low diversity cannot meet the requirement of a strong generalization

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
The Inria dataset (with Kitsap County excluded) and the WHU building dataset yield mIoU scores that are 6–8% lower than that for the WHU-Mix building dataset (trainval set), due to the relatively small scale and low diversity of these two datasets. Moreover, the label quality of the Inria dataset is unsatisfactory, which affects the performance of the building extraction model. Surprisingly, although the size of the CrowdAI dataset far exceeds that of the WHU-Mix building dataset, the performance with this dataset is very poor, with a 14.68% mIoU gap to the WHU-Mix building dataset. We believe that the main reason for this is that the images in the CrowdAI dataset lack variation, and the buildings lack intraclass diversity, which was verified in our visual check, which results in the features learned by the model being domain-specific. From the abovementioned results, it can be seen that both the diversity and quantity of data have a great impact on the generalization performance of the building extraction model, and the diversity can play a more important role, because a dataset with a very large quantity but low diversity will still lead to a weak generalization capability.

When taking a closer look at the IoU score for each city, it can be found that, even the best-performing dataset, i.e., the WHU-Mix building dataset, achieves very low IoU scores with the challenging DE_Potsdam and SD_Khartoum data. These results show the difficulty of practical building extraction, i.e., the data distributions of some test images are very far from the sample space covered by the training datasets. Therefore, there is a great need to develop domain generalization methods to further improve the generalization ability of building extraction models.

Fig. 7 shows 512 × 512 image patches from the five cities in the test set of the WHU-Mix building dataset. It can be seen that, in the US_Kitsap, CN_Wuxi, and NZ_Dunedin scenes, most of the datasets produce relatively accurate segmentation results. However, for the DE_Potsdam and SD_Khartoum scenes, the building extraction results are not as good. Overall, using the WHU-Mix building dataset (trainval set) results in the best performance, and produces the most complete building masks, especially in the red box areas of Fig. 7.

2) Influence of label quality: In this section, we investigate the influence of label quality on the building extraction performance. The experiment was conducted on three subsets of the Inria dataset—Western Tyrol, Vienna, and Kitsap County [19]—which are also included in the WHU-Mix building dataset. The original labels for Western Tyrol, Vienna, and Kitsap County have some significant errors and are of a poor quality (see Fig. 3). The labeling errors were corrected, one by one, manually and carefully, in the construction of the WHU-Mix building dataset. We selected the images of Western Tyrol and Vienna as the training set to train AT-MAFCN, and the Kitsap County images were used as the test set, resulting in four cases: 1) training and testing with the original training and test sets; 2) training with the original training set and testing with the revised test set; 3) training with the revised training set and testing with the original test set; and 4) training and testing with the revised training and test sets.

The results are listed in Table IV. First, we compare the results for the first case and the fourth case. The IoU score for the first case (using the original low-quality training and test sets) is only 49.89%, and the score for the fourth case (using the revised sets) is as high as 69.33%. The IoU score gap between the two results is as large as 19.44%. By comparing the results for the second case and the fourth case, training with the revised training set improves the IoU score from 59.76% to 69.33%, indicating that training data with a low label quality introduce a lot of noisy information that prevents the model from learning the effective information. By comparing the results for the third and fourth cases, testing with the revised test set results in an improvement of 13.13% in IoU score over testing with the original test set, indicating that a test set with inaccurate labels cannot reflect the true ability of a method. In summary, we conclude that the label quality of both the training set and test set have a significant influence on the building extraction performance. Therefore, developing a dataset with high-quality labels for practical building extraction is very important and necessary, which was one of the aims of this article.

Fig. 8 gives two examples of 512 × 512 image patches from the Kitsap County test set. Fig. 8(b) shows the original labels of the two image patches with obvious errors, and Fig. 8(c) and (d) shows the prediction maps obtained by training with the original and revised training sets, respectively. It can be observed by comparing Fig. 8(b) and (d) that the use of low-quality labels for the evaluation leads to a lower performance and a failure of the fair evaluation. The comparison of Fig. 8(c) and (d) indicates the importance of a high-quality training set for improving the building extraction model performance.

### Table III

| Dataset         | Number of image patches | mIoU (%) | IoU (%) | Training time (h) |
|-----------------|-------------------------|---------|---------|-------------------|
|                 |                         |         | US               | CN  | NZ  | DE  | Potsdam | SD  |                 |
| Massachusetts   | 1359                    | 22.37   | 28.39   | 14.84           | 46.42 | 7.50 | 19.97   | 2.1  |
| Inria           | 14 400                  | 52.46   | 68.09   | 63.95           | 63.75 | 37.70 | 29.79   | 22.4 |
| CrowdAI         | 341 058                 | 44.21   | 34.86   | 52.23           | 58.36 | 41.68 | 29.64   | 86.2 |
| WHU             | 12 430                  | 51.33   | 59.75   | 61.35           | 84.53 | 20.06 | 35.07   | 10.7 |
| WHU-Mix (trainval set) | 43 727               | 58.89   | 69.76   | 79.61           | 83.02 | 24.02 | 39.60   | 52.9 |

### Table IV

| Training/Kitsap County | Testing/Kitsap County | Original | Revised |
|------------------------|-----------------------|----------|---------|
| Tyrol                  | Original              | 49.89    | 59.76   |
| Vienna                 | Revised               | 56.20    | 69.33   |
Fig. 7. Sample prediction results obtained by training AT-MAFCN with different building datasets for five $512 \times 512$ image patches. (a) Images. (b) Labels (cyan mask). (c)–(g) Results for the Massachusetts buildings dataset [17], inria dataset [19] (with kitsap county excluded), CrowdAI dataset [35], WHU building dataset [18], and WHU-Mix building dataset (trainval set), respectively.

Fig. 8. Two examples from the kitsap county test set. (a) Images. (b) Original labels. (c) Predictions obtained by the model trained on the original training set. (d) Predictions obtained by the model trained on the revised training set.
C. Evaluation of the Proposed Domain Generalization Method

In this section, we first conduct experiment to compare the proposed domain generalization method, i.e., BSM, with the existing domain generalization methods. Second, the ablation study was conducted for BSM, in which we evaluated the functions of the three submodules. Third, we focus on the universal adaptability of the proposed BSM method in different building extraction models.

1) Comparison with other domain generalization methods: In this section, we compare the generalization ability of the proposed BSM method and the recent domain generalization methods. AT-MAFCN was used as the classification model, and the methods were trained on the trainval set of the WHU-Mix building dataset, and tested in the test set of the WHU-Mix building dataset. The compared methods were a recently proposed image-level method known as Mixstyle [46], three feature-level methods, i.e., IBN-Net [52], MMLD [51], and LDDG [48], and an image/feature combined method known as RobustNet [53]. The compared methods, as well as the proposed method, are the few domain generalization methods that can be used in the absence of domain labels. A description of the compared methods has been given in Section II. The image-level methods are basically independent of the classification model. However, the feature-level and image/feature combined methods required us to revise AT-MAFCN or add new losses in the training process. Taking IBN-Net as an example, we placed an instance normalization layer and a rectified linear unit (ReLU) layer at the end of the first three scales of the encoder of AT-MAFCN, followed by the IBN-b structure described in Pan et al. [52].

The experimental results of the different methods are listed in Table V. The result of AT-MAFCN trained without any domain generalization method is also listed as the baseline. The proposed BSM method obtains a mIoU score of as high as 72.06%, and shows a significant improvement (13.17%) over the baseline, representing a great success in improving the model's generalization ability. BSM also outperforms all the compared methods with respect to the mIoU score. Among the three feature-level methods, only IBN-Net improves the mIoU score of the baseline slightly, while MMLD and LDDG make the results even worse. The image-level Mixstyle method shows a slight improvement over the baseline. The image/feature combined method—RobustNet—is the second-best method; however, there is a 4.59% mIoU gap between RobustNet and the proposed BSM. We take the abovementioned results as an indication that image-level generalization is more stable than feature-level generalization, and can be used as a prerequisite for the feature-level generalization, to make it more robust. However, feature-level generalization may not be necessary as the proposed BSM method beats the very recent RobustNet by a large margin, demonstrating that a well-designed image-level method can be more powerful than a complex image/feature combined method.

At the individual city scale, it can be seen that the proposed BSM method improves the results of the baseline greatly in the most challenging DE_Potsdam (34% IoU improvement) and SD_Khartoum (9% IoU) scenarios, while the other methods are far less effective. RobustNet, in particular, even results in a 9% IoU decrease with the SD_Khartoum data. For the US_Kitsap, CN_Wuxi, and NZ_Dunedin data, the improvements of the proposed BSM method are relatively modest, and are at the same level as RobustNet. This can be explained by the fact that a higher improvement can be accessed when the room for improvement is larger. Nevertheless, the results for the DE_Potsdam and SD_Khartoum data indicate that the BSM method is more robust and shows a significant advantage over the other methods.

The training of the proposed method consumes 18 h more than the baseline, due to the data augmentation steps. However, it has a comparable efficiency with the compared methods. Considering the significant improvement of the proposed BSM method over the other five methods and the baseline, its efficiency can be considered to be satisfactory.

A visual comparison of the results of the different domain generalization methods is shown in Fig. 9. It can be seen that the proposed BSM method generates more accurate building masks than the other methods, especially with the DE_Potsdam and SD_Khartoum data.

2) Ablation study: To further verify the effectiveness of the different parts of the proposed BSM domain generalization method, we investigated the performance gain with the use of the different submodules, i.e., geometric augmentation (GA), color augmentation (CA), style mixing (SM), and their combinations. The results are listed in Table VI. “AT-MAFCN (Baseline)” means training AT-MAFCN without BSM. By comparing the “GA,” “CA,” and “SM” submodules, it can be seen that GA achieves the best result. This is because buildings are highly variable in both shape and size, and only GA can handle the variations in sizes and shapes. For CA and SM, both of which

| Method                | mIoU (%) | IoU (%) | Training time (h) |
|-----------------------|----------|---------|-------------------|
| AT-MAFCN (Baseline)   | 58.89    | 69.76   | 52.9              |
| Mixstyle              | 60.86    | 76.45   |                   |
| IBN-Net               | 59.91    | 75.77   |                   |
| MMLD                  | 56.77    | 68.06   |                   |
| LDDG                  | 58.02    | 71.21   |                   |
| RobustNet             | 67.47    | 81.16   |                   |
| BSM (proposed)        | 72.06    | 80.93   |                   |

The proposed BSM method achieves the best results in almost all of the compared methods. However, the feature-level and image/feature combined methods are basically independent of the classification model. However, the feature-level and image/feature combined methods required us to revise AT-MAFCN or add new losses in the training process. Taking IBN-Net as an example, we placed an instance normalization layer and a rectified linear unit (ReLU) layer at the end of the first three scales of the encoder of AT-MAFCN, followed by the IBN-b structure described in Pan et al. [52].

The experimental results of the different methods are listed in Table V. The result of AT-MAFCN trained without any domain generalization method is also listed as the baseline. The proposed BSM method obtains a mIoU score of as high as 72.06%, and shows a significant improvement (13.17%) over the baseline, representing a great success in improving the model’s generalization ability. BSM also outperforms all the compared methods with respect to the mIoU score. Among the three feature-level methods, only IBN-Net improves the mIoU score of the baseline slightly, while MMLD and LDDG make the results even worse. The image-level Mixstyle method shows a slight improvement over the baseline. The image/feature combined method—RobustNet—is the second-best method; however, there is a 4.59% mIoU gap between RobustNet and the proposed BSM. We take the abovementioned results as an indication that image-level generalization is more stable than feature-level generalization, and can be used as a prerequisite for the feature-level generalization, to make it more robust. However, feature-level generalization may not be necessary as the proposed BSM method beats the very recent RobustNet by a large margin, demonstrating that a well-designed image-level method can be more powerful than a complex image/feature combined method.

At the individual city scale, it can be seen that the proposed BSM method improves the results of the baseline greatly in the most challenging DE_Potsdam (34% IoU improvement) and SD_Khartoum (9% IoU) scenarios, while the other methods are far less effective. RobustNet, in particular, even results in a 9% IoU decrease with the SD_Khartoum data. For the US_Kitsap, CN_Wuxi, and NZ_Dunedin data, the improvements of the proposed BSM method are relatively modest, and are at the same level as RobustNet. This can be explained by the fact that a higher improvement can be accessed when the room for improvement is larger. Nevertheless, the results for the DE_Potsdam and SD_Khartoum data indicate that the BSM method is more robust and shows a significant advantage over the other methods.

The training of the proposed method consumes 18 h more than the baseline, due to the data augmentation steps. However, it has a comparable efficiency with the compared methods. Considering the significant improvement of the proposed BSM method over the other five methods and the baseline, its efficiency can be considered to be satisfactory.

A visual comparison of the results of the different domain generalization methods is shown in Fig. 9. It can be seen that the proposed BSM method generates more accurate building masks than the other methods, especially with the DE_Potsdam and SD_Khartoum data.

2) Ablation study: To further verify the effectiveness of the different parts of the proposed BSM domain generalization method, we investigated the performance gain with the use of the different submodules, i.e., geometric augmentation (GA), color augmentation (CA), style mixing (SM), and their combinations. The results are listed in Table VI. “AT-MAFCN (Baseline)” means training AT-MAFCN without BSM. By comparing the “GA,” “CA,” and “SM” submodules, it can be seen that GA achieves the best result. This is because buildings are highly variable in both shape and size, and only GA can handle the variations in sizes and shapes. For CA and SM, both of which...
help the model adapt to the variations of image colors and styles, SM shows a better performance than CA, showing its superiority in diversifying the image styles. It is worth noting that SM is the only submodule that improves the performance of AT-MAFCN on the SD_Khartoum data, while the other two submodules have a negative effect. The combinations of two submodules show better performances than the individual submodules, which shows that the three submodules are complementary and mutually reinforcing. Not unexpectedly, the best performance is achieved by the joint combination of the three submodules, verifying the effectiveness of the proposed BSM domain generalization method.

Compared with GA, CA, and their combinations, SM and the combinations including SM take a slightly longer time for the training; however, their efficiency is at the same level.

3) Evaluation of the universal adaptability of the domain generalization method: In the experiments described previously, we implemented AT-MAFCN as the baseline model for the domain generalization baseline to extract buildings. However, it is important to note that the proposed domain generalization method is not especially designed for AT-MAFCN, and is a plug-and-play data augmentation module, which can be coupled with arbitrary modern deep learning based building extraction models. We imbedded BSM into three popular semantic segmentation models, i.e., DeepLab V3+ [66], UNet [65], and TransUnet [68] to further investigate the robustness and performance of BSM. Among them, DeepLab V3+ and UNet are CNN-based models and TransUnet is a combined model of CNN and transformer. Training with the trainval set of the WHU-Mix building dataset and testing with the test set, we obtained the results listed in Table VII. It can be seen that BSM significantly improves the performance of the models, demonstrating its ability to improve the generalization ability of the existing models. There is a performance gap between UNet
and other methods, indicating that we should obtain a better performance when we equip a better building extraction model in the future with BSM. But more importantly, the performance difference between BSM imbedded AT-MAFCN and TransUnet are very small, indicating that developing domain generalization is currently more important than developing different segmentation structures, as the latter have reached a peak after several years of extensive studies.

From the abovementioned comprehensive evaluation of the proposed BSM method, it is concluded that BSM not only exhibits a superior performance and robustness, compared to the commonly used data augmentation strategies and the more recent sophisticated methods, but it also has the advantage of being a portable and plug-and-play module. This indicates that style mixing, which is mainly a binary operation, is a powerful and universal strategy, as effective as and supplementary to the widely used geometric and color augmentation, which is a unary operation, for enlarging the distribution space of training data, which has not been recognized by the remote sensing community. We believe that BSM could become a new domain generalization baseline, not only for building extraction, but also for most applications involved with remote sensing image processing.

VI. DISCUSSION

In our method, we use deep learning style transfer method, AdaIN, for style mixing. However, it is only an option for style mixing in BSM. In fact, more style transfer methods, such as Wallis filter [62] and FDA [69], can be used as the style mixing module. In the future, it is worth exploring more effective and efficient style transfer methods to be embedded in BSM.

Our method utilizes the large-scale dataset and domain generalization techniques to improve building extraction performance for arbitrary remote sensing images, instead of domain adaptation techniques that require the target images. However, domain adaptation has its advantage: when the target data is available, we may alignment better the distributions of source images and target images. Hence, the combination of domain generalization and domain adaptation may be worth to explore.

VII. CONCLUSION

In this article, we have described how we built a new diverse, large-scale, and high-quality building dataset, the WHU-Mix building dataset, to promote building extraction research toward more practical applications. The comparison experiments conducted between different building datasets verified the advantage of the WHU-Mix building dataset in promoting the performance of a building extraction model and properly evaluating the generalization ability of a model, and also the importance of diversity and high quality.

We also proposed a domain generalization method named BSM, which can significantly improve the generalization ability of a building extraction model in a convenient, efficient, and plug-and-play manner. The classic data augmentation methods and the novel style mixing in BSM complement each other, providing the building extraction model with a vast data distribution space to learn the domain-invariant information. The extensive experiments demonstrated the superiority and robustness of BSM over both image-level and feature-level domain generalization methods, and also its ease of use. We believe that BSM could serve as a new domain generalization baseline for remote sensing image processing, and will not be limited to building extraction in the future.

### References

[1] R. N. Clark and T. L. Roush, “Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications,” *J. Geophysical Res.*, vol. 89, pp. 6329–6340, 1984.

[2] S. Freire et al., “Introducing mapping standards in the quality assessment of buildings extracted from very high resolution satellite imagery,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 90, pp. 1–9, 2014.

[3] D. Griffiths and J. Boehm, “Improving public data for building segmentation from convolutional neural networks (CNNs) for fused airborne lidar and image data using active contours,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 154, pp. 70–83, 2019.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[5] C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[6] Y. Liu et al., “ARC-net: An efficient network for building extraction from high-resolution aerial images,” *IEEE Access*, vol. 8, pp. 154997–155010, 2020.

[7] S. Wei, S. Ji, and M. Lu, “Toward automatic building footprint delineation from aerial images using CNN and regularization,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 2176–2189, Mar. 2020.

[8] S. Wei and S. Ji, “Graph convolutional networks for the automated production of building vector maps from aerial images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 60, Mar. 2021, Art. no. 5602411.

[9] W. Zhao, C. Persello, and A. Stein, “Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 119–131, 2021.
[10] Y. Zhou et al., “BOMSC-net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery,” IEEE Trans. Geosci. Remote Sens., vol. 60, no. 60, Feb. 2022, Art. no. 5618617.

[11] Z. Tang et al., “Capable–encoder-decoder: A method for generalizable building extraction from remote sensing images,” Remote Sens., vol. 14, no. 5, 2022, Art. no. 1235.

[12] Z. Zou, T. Shi, W. Li, Z. Zhang, and Z. Shi, “Do game data generalize well for remote sensing image segmentation?,” Remote Sens., vol. 12, no. 2, 2020, Art. no. 275.

[13] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, “Object detection in optical remote sensing images: A survey and a new benchmark,” ISPRS J. Photogrammetry Remote Sens., vol. 159, pp. 296–307, 2020.

[14] X. Qi et al., “MLRNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding,” ISPRS J. Photogrammetry Remote Sens., vol. 169, pp. 337–350, 2020.

[15] M. Luo and S. Ji, “Cross-Domain Next-Generation land-cover classification from VHR remote sensing images with deep learning based domain adaptation,” ISPRS J. Photogrammetry Remote Sens., vol. 191, pp. 105–128, 2022.

[16] D. Wittich and F. Rottensteiner, “Appearance based deep domain adaptation for the classification of aerial images,” ISPRS J. Photogrammetry Remote Sens., vol. 180, pp. 82–102, 2021.

[17] V. Mnih, Machine Learning for Aerial Image Labeling. Toronto, ON, Canada: Univ. of Toronto, 2013.

[18] S. J. Si, S. Wei, and M. Lu, “Fully convolutional networks for multi-source building extraction from an open aerial and satellite imagery data set,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 1, pp. 574–586, Jan. 2019.

[19] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “Can semantic labeling methods generalize to any city? The aerial image labeling benchmark,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2017, pp. 3226–3229.

[20] Y. Hu and F. Guo, “Automatic building extraction based on high resolution aerial images,” in Proc. IEEE Int. Conf. Electron. Inf. Technol. Comput. Eng., 2019, pp. 1017–1020.

[21] M. Chen et al., “Removing an improved network for building extraction from high resolution remote sensing image,” Remote Sens., vol. 13, pp. 1–19, 2021.

[22] D. Wittich and F. Rottensteiner, “Appearance based deep domain adaptation for the classification of aerial images,” ISPRS J. Photogrammetry Remote Sens., vol. 180, pp. 82–102, 2021.

[23] R. Zhang, S. Newsam, Z. Shao, X. Huang, J. Wang, and D. Li, “Multi-scale adversarial network for vehicle detection in UAV imagery,” ISPRS J. Photogrammetry Remote Sens., vol. 180, pp. 283–295, 2021.

[24] M. Long, H. Zhu, J. Wang, and M. I. Jordan, “Deep transfer learning with joint adaptation networks,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 2208–2217.

[25] Y. Tsai, K. Shu, S. Schulter, and M. Chandraker, “Domain adaptation for structured output via discriminative patch representations,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 1456–1465.

[26] L. Yan, B. Fan, H. Liu, C. Hsu, S. Xiang, and C. Pan, “Triplet adversarial domain adaptation for pixel-level classification of VHR remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 5, pp. 3558–3573, May 2020.

[27] Z. Zhang, K. Doi, A. Iwasaki, and G. Xu, “Unsupervised domain adaptation of high-resolution aerial images via correlation alignment and self training,” IEEE Geosci. Remote Sens. Lett., vol. 18, no. 4, pp. 746–750, Apr. 2021.

[28] C. Gan, T. Yang, and B. Gong, “Learning attributes equals multi-source domain generalization,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 87–97.

[29] K. Muandet, D. Balduzzi, and B. Schölkopf, “Domain generalization via causal feature representation,” in Proc. 30th Int. Conf. Mach. Learn., 2013, pp. 1120–1128.

[30] H. Li, S. J. Pan, S. Wang, and A. C. Kot, “Domain generalization with adversarial feature learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 5400–5409.

[31] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, “Playing for data: Ground truth from computer games,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 102–118.

[32] M. Cords et al., “The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3213–3223.
[59] R. Xu, G. Li, J. Yang, and L. Lin, “Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation,” in Proc. Int. Conf. Comput. Vis., 2019, pp. 1426–1435.

[60] J. Choi, T. Kim, and C. Kim, “Self-ensembling with GAN-based data augmentation for domain adaptation in semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 6829–6839.

[61] M. Kim and H. Byun, “Learning texture invariant representation for domain adaptation of semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 12972–12981.

[62] D. Li, M. Wang, and D. Li, “Auto-dodging processing and its application for optical RS images,” Geomatics Inf. Sci. Wuhan Univ., vol. 31, pp. 753–756, 2006.

[63] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” in Adv. Neural Inf. Process. Syst., pp. 2017–2025, 2015.

[64] J. Fu et al., “Dual attention network for scene segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3141–3149.

[65] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention, 2015, vol. 9351, pp. 234–241.

[66] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proc. Eur. Conf. Comput. Vis., 2018, vol. 11211, pp. 801–818.

[67] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representations, 2015.

[68] J. Chen et al., “Transunet: Transformers make strong encoders for medical image segmentation,” 2021, arXiv:2102.04306.

[69] Y. Yang and S. Soatto, “FDA: Fourier domain adaptation for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 4084–4094.

Muying Luo received the B.E. degree in 2019 from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree.

Her research interests include remote sensing image processing and deep learning.

Shunping Ji (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2007.

He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University, China. His research interests include photogrammetry, remote sensing image processing, mobile mapping system, and machine learning.

Shiqing Wei (Member, IEEE) received the M.Sc. degree in 2019 from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree.

His research interests include remote sensing image processing and machine learning.