MASDG: Multiview Augmented Single-Source Domain Generalization Method for Robust Remote Sensing Building Extraction

Yunjiao Liu, Yuanyuan Liu, *Member, IEEE*, Kejun Liu, Yuxuan Huang, Chang Tang, *Senior Member, IEEE*, Wujie Zhou, *Senior Member, IEEE*, Zhe Chen, Wei Xiang, *Senior Member, IEEE*, and Hongyan Zhang, *Senior Member, IEEE*

***Abstract*—Despite advances in deep learning for remote sensing building extraction (RSBE), multitarget domain RSBE (MD-RSBE) remains challenging, as it requires transferring knowledge from a labeled source domain to multiple unlabeled target domains, with domain shifts in texture, style, and semantics. Existing domain adaptation (DA) and generalization (DG) methods face significant limitations: DA requires target-domain training, while DG needs multisource training, leading to high training costs and low generalization in practical MD-RSBE scenarios. To address this, we propose a multiview augmented single-source DG (MASDG) method, which effectively mitigates domain shifts across the RS source and target domains for robust MD-RSBE performance by enriching the diversity of the source domain through multiview augmentation and enforcing semantic consistency. Specifically, MASDG consists of three key components: texture-level domain augmentation (TDA) module, style-level domain augmentation (SDA) module, and semantic-invariant representation learning (SRL). To mitigate texture-level domain shift, TDA first introduces parameter-optimized multilayer random convolution to modify the texture of the source image, generating texture-augmented image pairs for simulating real-world texture diversity across various RS domains. Then, with each image pair from TDA, SDA employs two paralleled encoders, namely, the general feature encoder and the batch-guided style encoder, to formulate multiview building features, further mitigating style-level domain shift. Finally, SRL ensures SRL via a dual mechanism, including multiview segmentation loss and semantic consistency loss. The former generates predictions from diverse feature views (original, texture-augmented, and style-augmented), while the latter performs semantic alignment by minimizing distribution discrepancies among predictions, bridging semantic inconsistency to enable robust segmentation. Extensive experiments across three different MD-RSBE settings with seven different target domains demonstrate that our MASDG outperforms existing state-of-the-art methods by a significant margin.**

***Index Terms*—Multiview semantic consistency, remote sensing building extraction (RSBE), single-source domain generalization (DG), style augmentation, texture augmentation.**

I. Introduction

REMOTE sensing building extraction (RSBE) involves identifying building regions in images by assigning a class label to each pixel, playing a vital role in urban planning, natural resource protection, land resource monitoring, and so on [1], [2], [3]. Although deep learning-based algorithms have significantly advanced RSBE, they face two major limitations. First, these models require a large amount of densely annotated training data, which is both time-consuming and costly to obtain. Second, in complex real-world scenarios, deep learning models suffer from domain shift, namely, significant discrepancies in texture, style and semantics between the training data (source domain) and the testing data (target domain) in RSBE, e.g., imaging mechanisms (optical and SAR), sensors (spectrum and resolution), environments (illumination and climate), and locations (urban and rural areas) [4], [5], [6]. To address these challenges, we focus on a more practical task setting, namely, multitarget domain RSBE (MD-RSBE). In this setting, the model is trained on a single labeled source domain and is expected to generalize to multiple unseen and unlabeled target domains. This setup better reflects real-world RSBE applications, where diverse environmental conditions and imaging variations make generalizing models across domains a pressing challenge.

To mitigate the above-mentioned domain shift problem in MD-RSBE, domain adaptation (DA) [4], [7], [8] and domain generalization (DG) [9], [10], [11] methods have been developed. DA methods align the data distribution between different domains, exploring better generalization of features learned in the source domain to the target domain, as shown in Fig. 1(a). For instance, BDL [7] employs image-level alignment by introducing bidirectional learning to facilitate

Received 19 April 2025; revised 12 July 2025 and 20 October 2025; accepted 24 November 2025. Date of publication 4 December 2025; date of current version 11 December 2025. This work was supported in part by the Natural Science Foundation of Hubei Province under Grant 2023AFB572 and in part by Hubei Key Laboratory of Intelligent Geo-Information Processing under Grant KLIGIP-2022-B10. *(Corresponding author: Yuanyuan Liu.)*

Yunjiao Liu, Yuanyuan Liu, Kejun Liu, Yuxuan Huang, Chang Tang, and Hongyan Zhang are with the School of Computer Science, China University of Geosciences (Wuhan), Wuhan 430074, China (e-mail: liuyunjiao@cug.edu.cn; liuyy@cug.edu.cn; liukejun@cug.edu.cn; cosinehuang@cug.edu.cn; tangchang@cug.edu.cn; zhanghongyan@whu.edu.cn).

Wujie Zhou is with Zhejiang University of Science and Technology, Hangzhou 310000, China (e-mail: wujiezhou@163.com).

Zhe Chen is with the School of Computing, Engineering and Mathematical Sciences, and the Cisco-La Trobe Centre for Artificial Intelligence and Internet of Things, La Trobe University, Melbourne, VIC 3086, Australia (e-mail: zhe.chen@latrobe.edu.au).

Wei Xiang is with the School of Computing, Engineering and Mathematical Sciences, La Trobe University, Melbourne, VIC 3086, Australia (e-mail: w.xiang@latrobe.edu.au).

Digital Object Identifier 10.1109/TGRS.2025.3640112

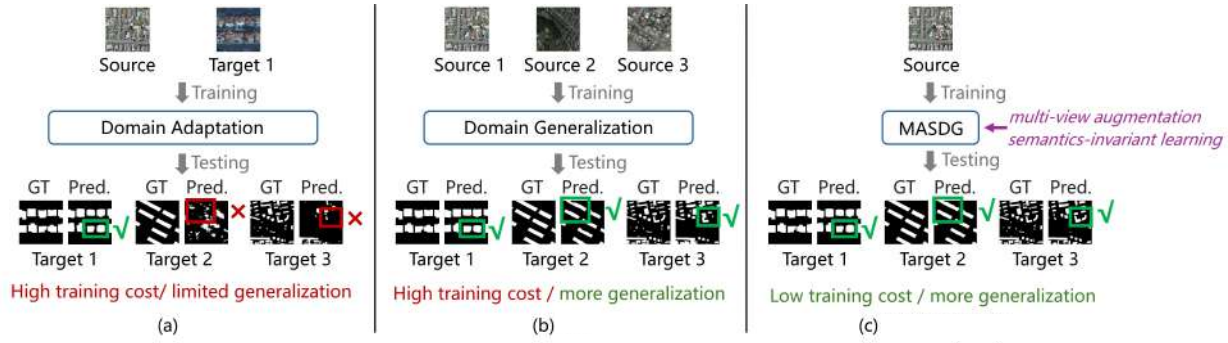

Fig. 1. Comparison of three ways to solve the domain shift problem in MD-RSBE, where GT denotes the GT, Pred. denotes the prediction of each method. (a) DA methods rely on a specific target domain to learn a target-adapted model, struggling to generalize to multiple target domains and requiring a high training cost. (b) DG methods utilize multiple source domains to learn domain-invariant features, requiring an extremely high data collection budget and high training cost. (c) Our MASDG method proposes multiview augmentation and learns semantic-invariant features only on a single-source domain, enabling the model to generalize to multiple target domains with low training cost and extensive generalization.

the reciprocal enhancement of the image translation model and segmentation model. FSDAN [8] aligns source and target images at the image, feature, and output-level through a two-stage adversarial learning based on a generative adversarial network. Despite progress, DA still requires access to specific unlabeled target domain data during training to learn a target-adapted model. As a result, the model must be retrained for each new target domain, exposing high training cost and limiting its generalization capability in multitarget domains.

Compared to DA, DG is independent of target domains, but requires multiple labeled source domains for training to learn domain-invariant representations, as shown in Fig. 1(b). Li et al. [9] learn universal feature representations by aligning the distributions of multiple source domains to generalize on target domains. L2A-OT [11] introduces a data generator to synthesize pseudo domains with distributions different from all source domains to further enhance the diversity of available training data. Despite the progress in multitarget domain shift, DG methods with multiple labeled source domains for training are very costly and labor-intensive, especially for the RSBE.

More recently, due to high efficiency and low labor costs, single-source DG (SDG) methods have been proposed in the general computer vision field for natural images, aiming to learn domain-invariant representations from a single labeled source domain to multitarget domains [12], [13], [14]. For example, RandConv [12] implements random convolution to generate extra images with diverse textures to expand the source domain, and incorporates the Kullback–Leibler divergence to enforce semantics invariance under texture changes. L2D [14] introduces a style-complement module to synthesize samples with unseen styles, and enforces semantic-invariant features by maximizing mutual information. Although these methods make progress in addressing the multitarget domain shift problem for general computer vision tasks, they still face two significant challenges for more complex MD-RSBE as follows.

1) *Diverse Domain Shift in MD-RSBE:* Remote sensing (RS) images possess richer spectral, textural, and stylistic characteristics, and include unique attributes not present in natural images, resulting in the domain shift arises from multiple factors, primarily the texture-level and style-level domain shifts [6], [15]. However, existing SDG methods typically employ single-view augmentation to address either texture-level or style-level domain shift, fail to comprehensively cover these domain shifts, affecting building extraction accuracy and restricting generalization.
2) *Augmentation-Induced Semantic Inconsistency in MD-RSBE:* Although the texture and style augmentations can simulate extensive distributions of multiple unseen RS target domains through domain expansion, they may introduce noise features and distort the original semantics, resulting in inconsistent outputs for the same semantics category, known as semantic inconsistency. In MD-RSBE, this semantic inconsistency can lead to inaccuracies in building extraction, such as blurred building boundaries and misclassifications of buildings and nonbuildings, hindering the generalization to target domains.

To address the challenges, we propose a novel multiview augmented SDG (MASDG) method for robust MD-RSBE performance. MASDG enhances source diversity through multiview augmentation and learns semantic-invariant representations to mitigate domain shifts across RS domains. Specifically, MASDG consists of three main components: the texture-level domain augmentation (TDA) module, the style-level domain augmentation (SDA) module, and the semantic-invariant representation learning (SRL). To mitigate the texture-level domain shift, TDA utilizes the parameter-optimized multilayer random convolution to modify the texture of the source image for generating texture-augmented image pairs. To further mitigate the style-level domain shift, SDA employs a dual encoder structure, the general feature encoder for standard feature extraction and the batch-guided style encoder for feature stylization to generate diverse feature views guided by batch-wise style variance. Finally, SRL enforces semantic consistency across different feature views by minimizing distribution differences among their predictions using multiview segmentation loss and semantic consistency loss.

In summary, our contributions are as follows.

1) We propose MASDG method tailored for the MD-RSBE task, which effectively transfers knowledge from a single RS source domain to multiple unseen RS target domains through multiview augmentation and semantic-invariant building representation learning. To the best of our knowledge, this is the first effective method for MD-RSBE.
2) To address texture- and style-level domain shifts in various RS domains, TDA and SDA are introduced to diversify the source domain through multiview augmentation. TDA generates texture-augmented image pairs while SDA further performs feature stylization that formulates multiview building features, simulating real-world texture and style diversities.
3) To further bridge semantic inconsistency among multiview augmentation for semantic-invariant building representation learning, SRL aligns multiview predictions from different feature views by using multiview segmentation loss and semantic consistency loss, thereby enabling robust building extraction across unseen target domains.
4) We conduct extensive experiments on three MD-RSBE settings, including WHU→Others, SAB→Others, and Crowd→Others, where the model is trained with a single RS source domain and evaluated on seven unseen RS target domains. The results show that the proposed MASDG outperforms existing SDG methods across all seven different target domains, confirming the effectiveness and generality of our proposed approach.

II. Related Work

A. RS Building Extraction

In recent years, deep learning-based models have flourished in RSBE due to their powerful feature extraction and nonlinear modeling ability, broadly categorized into CNN-based and hybrid transformer-based methods. CNNs excel at capturing local contextual information. MAP-Net [16] utilizes channel attention and pyramid pooling modules to fuse multiscale features. MHA-Net [17] further designs a multipath hybrid dilated convolution for enhanced multiscale building extraction. Refined-UNet [1] inherits the classic encoder–decoder structure and incorporates a refined skip connection to extract multiscale building features. CSA-UNet [18] integrates channel-spatial attention into a classic encoder–decoder model to capture discriminative RS building features. While these CNN-based frameworks struggle to capture global context for building extraction, transformers, with strong global modeling capabilities, offer a promising solution. However, local details remain crucial for building extraction. Hybrid transformer-based methods integrate CNNs' local feature extraction with transformers' global contextual modeling, notably advancing building extraction. CMTFNet [19] utilizes multiscale transformer blocks to process multiscale features extracted from CNN, further perform feature fusion for obtaining both local and global contextual information. BCT-Net [20] proposes a dual-branch framework combining both convolution and transformer encoders to capture both local and global contexts. Despite the process, these methods often output blurred building boundaries due to occlusion and noise interference. MSHFormer [21] addresses this by enhancing edge representations and suppressing background noise in low-level features. Similarly, EGAFNet [22] passes the edge features extracted from shallow layers to the decoder for supplementing building boundary information. However, these methods focus on a specific data source and struggle to generalize to out-of-distribution target domains.

B. DA in RSBE

DA utilizes both the labeled source domain data and the unlabeled target domain data during the training phase, with the goal of developing a model that could generalize to the target domain. Several studies have explored the application of DA to address the domain shift in building extraction tasks. Na et al. [23] utilize adversarial attack to generate target-like source domain images for achieving image-level alignment between the source and target domains. Similarly, JRPNet [24] employs the CycleGAN to modify the style of source domain images, obtaining the target-like source images. FDANet [4] makes full use of image-, feature-, and output-level information to adapt the model from the source domain to the target domain. FDANet employs the Wallis filter method to convert the source domain images to target-like ones for image-level alignment, adopts an adversarial learning module for feature-level alignment, and utilizes the mean-teacher model to achieve consistency regularization for output-level alignment. FLDA-NET [15] also employs full-level alignment consisting of image-level style transfer, feature-level entropy distribution minimization, and output-level cotraining algorithm for category-level alignment.

C. Domain Generalization

DG aims to train models on multiple source domains, so that they can be generalized to unseen target domains without having access to target data. Niu et al. [25] enhance the DG by fusing multiple SVM classifiers. MVDG [26] is a meta-learning-based DG method that employs multiple optimization paths to determine the best direction for model update. GTR-LTR [27] and WildNet [28] use external style datasets (e.g., paintings and ImageNet) for image-level style transfer. CCFP [29] introduces a learnable feature perturbation module to diversify the style of features while enforcing semantic consistency. Recently, there are also several DG methods in the field of RS. BSM [30] diversifies the style of the RS source domain images by randomly mixing the styles of the samples within the same batch, thus mitigating the domain shift in RSBE. FosMix [10] performs style randomization in the frequency domain, consisting of the full mix that integrates the style of the reference image into the RS source image as much as possible, while the optimal mix retains the frequencies crucial for segmentation and randomizes the remaining frequencies. DGMaskRCNN [31] integrates the domain adversarial modules at the image-, instance-, and pixel level for domain-invariant feature learning, so that the model trained on the RS source domain could generalize to unseen target domains.

D. Single-Source DG

SDG is a more challenging and practical setting where the model is trained on a single-source domain and tested on multiple unseen target domains. With limited data diversity, most SDG methods rely on data augmentation to enhance variability and improve generalization. RandConv [12] applies random convolutions to source images, altering textures while preserving shapes, to generate diverse samples for domain expansion. PDEN [32] generates additional images with different styles and textures in a progressive manner to expand the source domain. UDP [33] uses information theory metrics to minimize the correlation between original images and augmented ones, improving the diversity of augmented images. Despite prosperity in the field of natural images, SDG remains underexplored in RS. In the hyperspectral cross-scene classification tasks, Wang et al. [34] employ the encoder–decoder framework to divide the features into variant and invariant features, and generate new samples to expand the source domain by perturbing the variant features. TSDANet [35] utilizes the generative adversarial network to generate expand the diversity of the source domain, and introduces a spectral learning branch to eliminate the effect of noise on sample generation. In the RS semantic segmentation task, CCDR [6] randomly transforms the texture and style of the RS source image to generate additional images for domain expansion, thus facilitating the domain-invariant features. However, SDG receives little attention in RSBE.

III. Proposed Approach

A. Problem Definition

The MD-RSBE is formulated as follows: given a single labeled RS source domain $S = \{(x_i, y_i)|i = 1, \ldots, n_S\}$, where $x_i \in \mathbb{R}^{H\times W\times 3}$ is an RS image in the source domain, $y_i \in \mathbb{R}^{H\times W\times N_c}$ is its corresponding semantic label, i indexes the examples in S, N_c represents the number of semantic categories, n_S represents the size of S, and the multiple unlabeled RS target domains $T^m = \{(x_j^m)|j = 1, \ldots, n_T^m\}$, where j indexes over the examples in the mth target domain T^m, and n_T^m represents the size of T^m. It is worth noting that the source domain S and each target domain T^m share the same semantic categories, but exhibit different data distributions. The goal of MD-RSBE is to train a model using the single labeled source domain S, enabling the model to accurately identify buildings in new, unseen RS images from each target domain T^m. However, the domain shift from texture, style, and semantic disparities poses a significant challenge for MD-RSBE to transfer knowledge learned from the single-source domain to multiple unseen target domains.

B. Model Overview

We propose the MASDG framework for MD-RSBE, which transfers knowledge from a single labeled source domain to multiple unlabeled target domains. The MASDG consists of three key components, namely, the TDA module, the SDA module, and the SRL. First, given a source image, TDA applies parameter-optimized multilayer random convolution to generate texture-augmented variants, addressing the texture-level domain shift. Then, SDA takes both original and texture-augmented images as input and uses dual encoders—a general encoder for standard features and a batch-guided style encoder for style-augmented features—to produce multiview representations, mitigating the style-level domain shift. Finally, SRL promotes semantic-invariant learning via a dual mechanism: a multiview segmentation loss for diverse feature predictions and a semantic consistency loss for aligning their distributions. By integrating TDA, SDA, and SRL, MASDG effectively addresses texture-level and style-level domain shifts across RS domains and learns semantic-invariant building representations, enabling more effective knowledge transfer from a single source to multiple unseen target domains. Sections III-C–III-E detail the proposed TDA, SDA, and SRL.

C. TDA Module

The texture-level domain shift caused by the variations of material, density, and geographical location is one of the main factors for domain shift between different RS domains, making it challenging to generalize the RSBE model to multiple unseen target domains [6], [15]. To mitigate the texture-level domain shift, we propose the TDA module to perform stochastic texture transformation at the image level, generating rich texture-augmented images to simulate real-world texture diversity.

As shown in Fig. 2, given an input image from the source domain $x_i \in S$, the TDA module introduces the parameter-optimized multilayer random convolution to constantly generate the texture-augmented image x_i' during training, which presents different textures compared to the original source image x_i, thereby increasing the texture diversity of source domain data. Specifically, the TDA is a two-step process, the first step aims to perform parameter optimization to diversify the texture transformation parameters, obtaining the optimal parameter setting for current texture augmentation, ensuring even parameter space exploration, and enhancing the diversity of texture-augmented images. The second step utilizes the selected texture transformation parameters obtained from the first step, aiming to perform texture modification for the source input image through multilayer random convolution following RandConv [12].

1) Step 1: Parameter Optimization: To ensure the variability of texture-augmented images, we first introduce a parameter optimization strategy that diversifies the combinations of texture transformation parameters for each iteration. The texture transformation parameters consist of the number of random convolution layers M, the kernel size k, and the mixing ratio α. By varying the combinations of M, k, and α, the generated images will exhibit different textures. Following [12], we constrain the value range of each texture transformation parameter, where M is sampled from $M \in \{1, 2, \ldots, M_{\text{upper}}\}$ ($M_{\text{upper}} \geq 4$, denoted the upper bound of M), k is sampled from $\{1, 3, 5, 7\}$ and α is sampled from $[0, 1]$. To ensure comprehensive parameter space exploration and obtain the optimal parameter setting for each iteration, we introduce a parameter memory to store the texture transformation parameters M, k, and α used by previous iterations, which are managed through three parameter queues

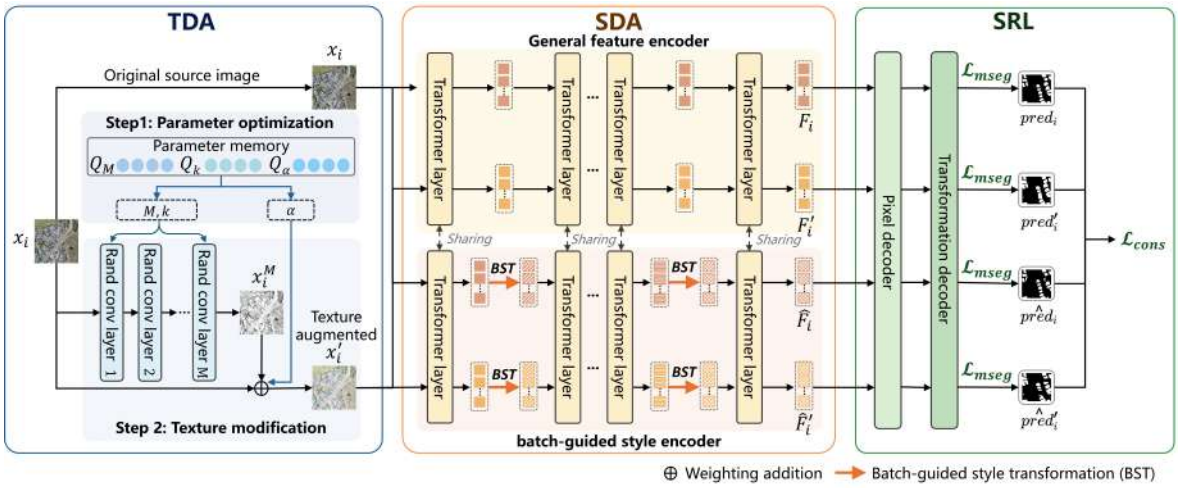

Fig. 2. Overview of the proposed MASDG. Given $x_i \in S$, TDA adopts parameter-optimized multilayer random convolution for texture augmentation, generating the texture-augmented image x_i' with different textures compared to the original source image x^i. Then, with the x_i and x_i', SDA employs dual paralleled encoders to obtain multiview building features, where the general feature encoder performs standard feature extraction to generate F_i and F_i', the batch-guided style encoder employs the BST to perform feature stylization, producing the style-augmented $\hat{F}_i$ and $\hat{F}_i'$. Finally, MSL adopts the multiview segmentation loss $\mathcal{L}_{\text{mseg}}$ to generate a prediction for each feature view, obtaining pred_i, pred_i', $\hat{\text{pred}}_i$, $\hat{\text{pred}}_i'$, and employs the $\mathcal{L}_{\text{cons}}$ for final semantics-consistent prediction.

Q_M, Q_k, and Q_α with length of *len*, respectively. Then, we count the occurrences of values in the parameter queues Q_M, Q_k, and Q_α, and identify parameters with the lowest usage frequency within their respective value ranges, which helps to evenly explore the parameter space and avoid over-reliance on frequently used parameter combinations, thereby enhancing the diversity of generated texture-augmented images.

2) Step 2: Texture Modification: After acquiring the texture transformation parameters M, k, and α for current iteration, we utilize the multilayer random convolution to modify the texture of source image x_i in a progressive manner, where the number of random convolution layers is M, the kernel size of each random convolution layer is k. We initialize the weight of lth random convolution layer by randomly sampling from a Gaussian distribution $\mathcal{N}(0, (1/3k^2))$, where $l \in \{1, 2, 3, \ldots, M\}$. The convolution operation for each layer can be represented as

$$x_i^{(l)} = \Theta^{(l)} * x_i^{(l-1)} \tag{1}$$

where $x_i^{(l-1)}$ is the input to the lth random convolution layer, $\Theta^{(l)}$ is the weight of the lth random convolution layer initialized from $\mathcal{N}(0, (1/3k^2))$, and * is convolution operation. Finally, the output of Mth layer x_i^M is further mixed with the original source image x_i through a linear combination at the ratio of α, obtaining the texture-augmented image x_i', which is formulated as

$$x_i' = \alpha x_i + (1 - \alpha)\, x_i^M. \tag{2}$$

Based on the above parameter-optimized multilayer random convolution operation, we can constantly obtain the texture-augmented image x_i' with different random local textures compared to the original source image x^i during training iterations, thus imitating the real-world texture diversity and addressing the texture-level domain shift within different RS domains.

D. SDA Module

While TDA addresses texture-level domain shift, the style-level shift caused by lighting, sensors, and environmental factors remains a key challenge. To mitigate this, the SDA module enriches style diversity by using the original source image x_i and its texture-augmented image x_i' from TDA. SDA employs two parallel encoders, which includes a general feature encoder (G-encoder) for standard feature extraction and a batch-guided style encoder (S-encoder) for generating style-augmented features, producing multiview building representations.

1) General Feature Encoder (G-Encoder): Taking each image pair x_i and x_i' from TDA as input, the general feature encoder aims to perform standard feature extraction layer by layer, obtaining the F_i and F_i', which is formulated as

$$F_i = \text{G-encoder}\,(x_i), \quad F_i' = \text{G-encoder}\,(x_i'). \tag{3}$$

2) Batch-Guided Style Encoder (S-Encoder): The S-encoder processes the image pair x_i and x_i' in parallel with the general feature encoder, generating style-augmented features $\hat{F}_i$ and $\hat{F}_i'$, to simulate real-world style diversity during training. The formulation of the S-encoder is as follows:

$$\hat{F}_i = \text{S-encoder}\,(x_i), \quad \hat{F}_i' = \text{S-encoder}\,(x_i'). \tag{4}$$

Specifically, in the S-encoder, we introduce a **batch-guided style transformation (BST)** strategy in intermediate layers (7th, 11th, 15th, and 23rd) of the backbone used in [36], to diversify the style of the output features of these layers by utilizing batch-wise style variance, as depicted in Algorithm 1. Given x_i or x_i' as input, we first obtain the output feature from one of these intermediate layers as F_s. Then, we compute the feature statistics of F_s, obtaining the feature mean μ_s and the feature standard deviation σ_s, representing the style of F_s. Since the target domain is unknown in real-world scenarios, there is inherent uncertainty regarding the style-level domain shift between the RS source and target domains.

Algorithm 1 BST

Input: Output features F_s from intermediate layers (7th, 11th, 15th, and 23rd) of the backbone

Output: Style-augmented features $\hat{F}_s$

1: **Initialize:** $\mu_1, \mu_2, \ldots, \mu_b$ and $\sigma_1, \sigma_2, \ldots, \sigma_b$ represent the mean and standard deviation of the features within the same batch as F_s, where b denotes the batch size.

2: Compute feature mean of F_s:
$$\mu_s = Mean(F_s)$$

3: Compute feature standard deviation of F_s:
$$\sigma_s = Std(F_s)$$

4: Multivariate Gaussian distribution assumption:
$$\mu_s \sim \mathcal{N}(\mu_s, \Sigma_\mu^2)$$
$$\sigma_s \sim \mathcal{N}(\sigma_s, \Sigma_\sigma^2)$$

5: Approximate Σ_μ^2 and Σ_σ^2 using variance of each feature statistics within the same batch:
$$\Sigma_\mu^2 = Var(\mu_1, \mu_2, \ldots, \mu_b),\ \mu_s \in [\mu_1, \ldots, \mu_b]$$
$$\Sigma_\sigma^2 = Var(\sigma_1, \sigma_2, \ldots, \sigma_b),\ \sigma_s \in [\sigma_1, \ldots, \sigma_b]$$

6: Sample new feature mean:
$$\hat{\mu}_s \leftarrow \mathcal{N}(\mu_s, \Sigma_\mu^2)$$

7: Sample new feature standard deviation:
$$\hat{\sigma}_s \leftarrow \mathcal{N}(\sigma_s, \Sigma_\sigma^2)$$

8: Apply $\hat{\mu}_s$ and $\hat{\sigma}$ to obtain $\hat{F}_s$:
$$\hat{F}_s = \hat{\sigma}_s \left(\frac{F_s - \mu_s}{\sigma_s}\right) + \hat{\mu}_s$$

9: **return** $\hat{F}_s$

As a result, it is infeasible to determine the direction and magnitude of changes in feature statistics. To model uncertain style-level domain shifts, we assume feature statistics follow multivariate Gaussian distributions: $\mu_s \sim \mathcal{N}(\mu_s, \mathbf{\Sigma_\mu^2})$ and $\sigma_s \sim \mathcal{N}(\sigma_s, \mathbf{\Sigma_\sigma^2})$, where means represent original stats and variances capture potential deviations. Following MixPatch [37], we utilize the variance of feature means and feature standard deviations within the same batch to approximate the $\mathbf{\Sigma_\mu^2}$ and $\mathbf{\Sigma_\sigma^2}$. After acquiring the $\mathbf{\Sigma_\mu^2}$ and $\mathbf{\Sigma_\sigma^2}$, we construct Gaussian distributions and sample new statistics as, $\hat{\mu}_s \sim \mathcal{N}(\mu_s, \Sigma_\mu^2)$ and $\hat{\sigma}_s \sim \mathcal{N}(\sigma_s, \Sigma_\sigma^2)$. Next, we replace the original feature statistics of F_s with the sampled feature statistics following AdaIN [38], obtaining the style-augmented feature $\hat{F}_s$. The BST is formulated as follows:

$$\hat{F}_s = \hat{\sigma}_s \left(\frac{F_s - \mu_s}{\sigma_s}\right) + \hat{\mu}_s. \tag{5}$$

Finally, the $\hat{F}_s$ is propagated through transformer layers, alternating feature extraction and BST-based style transformation, yielding the final style-augmented features $\hat{F}_i$ and $\hat{F}_i'$ for x_i and x_i', respectively.

Overall, the parallel G-encoder and S-encoder in SDA generate multiview features—original (F_i, F_i') and style-augmented ($\hat{F}_i$, $\hat{F}_i'$)—to simulate real-world style diversity and mitigate style-level domain shift between RS source and target domains.

E. Semantics-Invariant Representation Learning

The multiview building features F_i, F_i', $\hat{F}_i$, and $\hat{F}_i'$ from TDA and SDA enhance texture and style diversity. However, augmentations may introduce noise features and distort the original semantics, affecting generalization. To address this, SRL ensures semantic-invariant learning using multiview segmentation loss for predictions and a semantic consistency loss to align predictions, bridging semantic inconsistencies for robust building extraction.

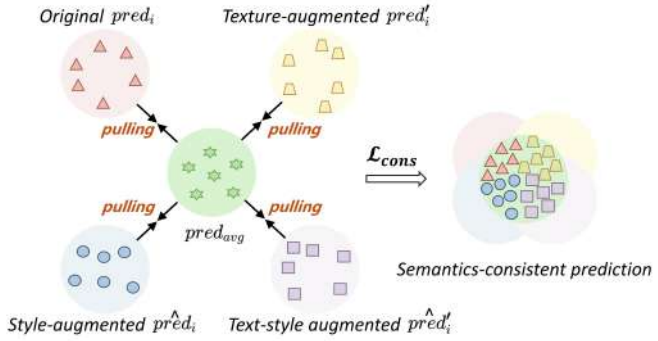

Fig. 3. Mechanism of the $\mathcal{L}_{\text{cons}}$ for semantic alignment in SRL. The $\mathcal{L}_{\text{cons}}$ promotes consistency by averaging multiview predictions (pred_{avg}), and pulling each prediction toward this average for semantics-consistent outputs.

1) Multiview Segmentation Loss: Using the multiview building features F_i, F_i', $\hat{F}_i$, and $\hat{F}_i'$, we devise the multiview segmentation loss with the segmentation head, to generate multiview predictions as pred_i, pred_i', $\hat{\text{pred}}_i$, $\hat{\text{pred}}_i'$. The multiview segmentation loss $\mathcal{L}_{\text{mseg}}$ encompasses four distinct basic segmentation loss functions, each corresponding to a feature view. For each feature view, the basic segmentation loss $\mathcal{L}$ is composed of the standard cross-entropy loss $\mathcal{L}_{\text{ce}}$, the binary cross-entropy loss $\mathcal{L}_{\text{bce}}$, and the dice loss $\mathcal{L}_{\text{dice}}$, following Mask2Former [39]. Formally, the $\mathcal{L}_{\text{mseg}}$ and $\mathcal{L}$ are formulated as

$$\mathcal{L} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} \tag{6}$$

$$\mathcal{L}_{\text{mseg}} = \mathcal{L}\left(\text{pred}_i, y_i\right) + \mathcal{L}\left(\text{pred}_i', y_i\right) + \mathcal{L}\left(\hat{\text{pred}}_i, y_i\right) + \mathcal{L}\left(\hat{\text{pred}}_i', y_i\right) \tag{7}$$

where the λ_{dice}, λ_{dice}, and λ_{dice} denotes the weights of each loss, and y^i denotes the ground truth (GT) of the original source input image x^i. Here, we set $\lambda_{\text{dice}} = 2$, $\lambda_{\text{dice}} = 5$ and $\lambda_{\text{dice}} = 5$ according to [39].

2) Semantic Consistency Loss: Intuitively, the RSBE model, which is robust against domain shift, should produce semantics-consistent predictions for the same semantic category, regardless of variations in texture and style. In this point of view, SRL further incorporates a semantic consistency loss that employed Jensen–Shannon (JS) divergence [40] to minimize the discrepancy among these multiview predictions, achieving semantic alignment of these predictions and finally obtaining the semantics-consistent prediction, as depicted in Fig. 3. More concretely, we first calculate the mean of the multiview predictions, denoted by pred_{avg}. Subsequently, we calculate the JS divergence between each prediction and the mean, and accumulate all of the obtained JS divergences to obtain the final semantic consistency loss $\mathcal{L}_{\text{cons}}$, which is defined as follows:

$$\text{pred}_{\text{avg}} = \frac{1}{4}(\text{pred}_i + \text{pred}_i' + \hat{\text{pred}}_i + \hat{\text{pred}}_i') \tag{8}$$

TABLE I
DOMAIN GAPS BETWEEN DIFFERENT RS DATASETS

Dataset	Spatial Resolution (m)	Image Resolution (pixels)	Shooting Area	Filming Angle
SAB	0.290	500 × 500	Beijing, Shanghai, Shenzhen, Wuhan	Orthographic, non-orthographic
WHU	0.075	512 × 512	Christchurch(New Zealand)	Orthographic
Crowd	0.300	300 × 300	Las Vegas, Paris, Shanghai, Khartum	Orthographic
UBC	0.5∼0.8	600 × 600	Beijing, Munich	Orthographic
Potsdam	0.05	512 × 512	Potsdam (Germany)	Orthographic
Vaihingen	0.09	512 × 512	Vaihingen (Germany)	Orthographic
Massachusetts	1	500 × 500	Boston(USA)	Orthographic
Inria	0.03	512 × 512	Austin, Chicago, Kitsap County, Western Tyrol, Vienna	Orthographic

$$\mathcal{L}_{\text{cons}} = \text{JS}(\text{pred}_{\text{avg}}, \text{pred}_i) + \text{JS}(\text{pred}_{\text{avg}}, \text{pred}'_i) + \text{JS}(\text{pred}_{\text{avg}}, \hat{\text{pred}}_i) + \text{JS}(\text{pred}_{\text{avg}}, \hat{\text{pred}}'_i). \quad (9)$$

3) Overall Objective Function: The overall loss function $\mathcal{L}_{\text{total}}$ is composed by the multiview segmentation loss $\mathcal{L}_{\text{mseg}}$ and the semantic consistency loss $\mathcal{L}_{\text{cons}}$. These losses work synergistically for domain-invariant building feature learning to mitigate the RS domain shift involving the texture, style, and semantic disparities, enabling the RSBE model to generalize to multiple unseen target domains. The $\mathcal{L}_{\text{total}}$ is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mseg}} + \omega \mathcal{L}_{\text{cons}} \quad (10)$$

where ω denotes the weighting factor of the semantic consistency loss. In our experiments, ω is set to 1.0. More detailed discussions about the weight of $\mathcal{L}_{\text{cons}}$ can be seen in Section IV-E.

IV. EXPERIMENTS

A. Datasets

To evaluate our proposed approach, eight RS building datasets were used: Chinese Typical Urban Building Instance Dataset (SAB) [41], WHU Aerial Image Dataset (WHU) [42], CrowdAI Mapping Challenge Dataset (Crowd) [43], UBC Satellite Image Dataset (UBC) [44], ISPRS Potsdam Dataset (Potsdam) [45], ISPRS Vaihingen Dataset (Vaihingen) [46], Massachusetts building dataset (Massachusetts) [47], and Inria Aerial Image Labeling Dataset (Inria) [48]. The more detailed information about the eight datasets was shown in Table I.

1) Chinese Typical Urban Building Instance Dataset [41]: The Chinese Typical Urban Building Instance Dataset (hereinafter referred to as SAB) is captured from Beijing, Shanghai, Shenzhen, and Wuhan. There are 7260 images, with a total of 63 886 buildings in SAB. Among them, 5985 images are used for training, and 1275 images are used for testing. Each RS image in SAB has dimensions of 500×500 pixels and a spatial resolution of 0.29 m.

2) Whu Aerial Image Dataset [42]: The WHU Aerial Image Dataset (hereinafter referred to as WHU) is captured from Christchurch, New Zealand. Each image has dimensions of 512×512 pixels, and the spatial resolution is 0.075 m. The training set contains 4736 images, and the test set contains 2416 images.

3) CrowdAI Mapping Challenge Dataset [43]: The CrowdAI Mapping Challenge Dataset (hereinafter referred to as Crowd) is captured from multiple cities, including Los Angeles, Paris, and Shanghai, with a spatial resolution of 0.3 m. Each image in CrowdAI has dimensions of 300 × 300 pixels. The training set contains 280 741 images, and the test set contains 60 697 images.

4) UBC Satellite Image Dataset [44]: The UBC Satellite Image Dataset (hereinafter referred to as UBC) is from Beijing, China, and Munich, Germany, and the spatial resolution is 0.5–0.8 m. Each image in UBC has dimensions of 600 × 600 pixels. The training set contains 560 images, and the test set contains 160 images.

5) ISPRS Potsdam Dataset [45]: The ISPRS Potsdam Dataset (hereinafter referred to as Potsdam) contains 38 fine-resolution images of size 6000 × 6000 pixels. The images are from Potsdam city in Germany with a spatial resolution is 0.05 m. The training set contains 24 images, and the test set contains 14 images. The dataset contains six categories, namely, surfaces, buildings, low vegetation, trees, cars, and clutter/background. Since only the building category is concerned, we process the annotation file, keep only the building category, and set other categories as background. Because the resolution of the original image is too high, we crop the raw images into 512 × 512 patches.

6) ISPRS Vaihingen Dataset [46]: The ISPRS Vaihingen Dataset (hereinafter referred to as Vaihingen) is composed of 33 images with an average size of 2494 × 2064 pixels. The images come from a small village, including multiple independent buildings and smaller multistory buildings, with a spatial resolution of 0.09 m. The dataset category is the same as Potsdam, so we process the origin image as Potsdam does and crop the original image to 512 × 512 patches using the dataset processing method provided by mmsegmentation [49].

7) Massachusetts Building Dataset [47]: The Massachusetts Building Dataset (hereinafter referred to as Massachusetts) is captured from Boston, the United States, and the spatial resolution is 1 m. The training set contains 137 images, the val set contains four images, and the test set contains ten images. Each image in Massachusetts has dimensions of 1500 × 1500 pixels. Since the original image resolution is too high, we crop the raw image to 500 × 500 patches.

8) Inria Aerial Image Labeling Dataset [48]: Since the label of the test set is not released, we use the provided training set for testing. The training set of Inria Aerial Image Labeling Dataset (hereinafter referred to as Inria) contains 180 images with dimensions of 5000×5000 pixels, which are from Austin, Chicago, Kishap County, West Tyrol, and Vienna, with a spatial resolution of 0.03 m. We crop the raw images into 500 × 500 patches.

B. Evaluation Protocol

In order to evaluate the performance of our proposed method, we employed the widely used intersection over union (IoU) [50] and $F1$-score ($F1$) [51] as the evaluation metrics. IoU and $F1$ are defined as

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{11}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

where TP, FP, and FN indicate the true positive, false positive, and false negatives, respectively. Note that higher $F1$ and IoU denote better overall performance.

C. Implementation Details and Task Settings

The code implementation was based on MMSegmentation [49]. We utilized Dinov2 [52] as the backbone for feature extraction and the segmentation head of Mask2Former [39] to generate pixel-level predictions. During training, we set the learning rate to $1e^{-5}$ and $1e^{-4}$ for the backbone and segmentation head, respectively. We employed AdamW as the optimizer with a batch size of 4, and the model was trained for 80 000 iterations on a Linux Platform with NVIDIA GeForce RTX 2080 Ti GPU and NVIDIA GeForce RTX 3090 GPU. During all training processes, we scaled and resized the input images to 256×256 pixels. We performed the experiments on three cross-domain scenarios, the definitions of the source domain and the target domains, as below in the form of source→targets:

1) Whu→Others: We selected the WHU as the source domain and selected the other seven datasets (SAB, Crowd, UBC, Vaihingen, Massachusetts, Potsdam, and Inria) as the target domains for evaluation.

2) Sab→Others: We selected the SAB as the source domain and selected the other seven datasets (WHU, Crowd, UBC, Vaihingen, Massachusetts, Potsdam, and Inria) as the target domains for evaluation.

3) Crowd→Others: We selected the Crowd as the source domain and selected the other seven datasets (SAB, WHU, UBC, Vaihingen, Massachusetts, Potsdam, and Inria) as the target domains for evaluation.

In the comparative experiment, we employed Rein [36] as our baseline model, which incorporates a parameter-efficient fine-tuning strategy to adapt the visual foundation model (VFM) for domain-generalized semantic segmentation. Rein employs transformer-based VFM as a backbone and incorporates the widely used segmentation head of Mask2former for pixel-level predictions, along with the basic data augmentations. Besides, Rein introduces learnable tokens to each transformer layer and computes cross-attention between features and tokens, thus associating each token to different instances in the image to facilitate instance-level feature refinement. Benefiting from the strong generalization capabilities of VFM and an effective fine-tuning strategy, Rein significantly outperforms existing DG methods.

We also selected several state-of-the-art SDG methods proposed in recent years for comparison under the above three cross-domain settings, including RobustNet [53], MDGVR [25], MVDG [26], SHADE [54], SiamDoGe [55], Dual-Level [56], HGFormer [57], DIIA [58], CCDR [6], BlindNet [59], CMFormer [60], and CPerb [37]. Besides, we followed the original network structures of these methods to avoid structural modification biases. RobustNet [53] removes style information that is sensitive to domain shift from feature covariance, while retaining domain-invariant content information. MDGVR [25] learns domain-invariant patterns through low-rank regularization imposed on the weights of multiple trained models. MVDG [26] leverages multiview optimization trajectories in training and multiview predictions in testing for more stable predictions. SHADE [54] uses the basis styles of the source domain to generate samples with novel styles in the image space, minimizing the JS divergence between the predictions of the original sample and the generated one. SiamDoGe [55] diversifies the feature space by mixing the feature statistics of two color-jittered versions from the same original sample with AdaIN [38]. Dual-level [56] introduces a two-stage domain augmentation, the first stage generates enhanced samples at the image level while the second stage diversifies per-class features at the feature level. HGFormer [57] groups pixels into part-level masks and further aggregates part-level masks into whole-level masks, using two scales to generate the final prediction results. DIIA [58] utilizes domain-invariant edge and semantic layout information to facilitate generalization on unseen target domains. CCDR [6] randomly transforms the texture and style of the RS source image to generate additional images for domain expansion. BlindNet [59] incorporates extra loss constraints to ensure the encoder generates style-invariant features, and employs semantic consistency contrast learning to improve the robustness of the segmentation prediction against domain shifts. CMFormer [60] proposes the content-enhanced mask attention for domain-invariant content representation learning and handling style variation. CPerb [37] employs multiview augmentation at both image and feature levels to augment the original single-source domain.

D. Discussions of Experimental Results

1) WHU→Others: We used the WHU dataset as the source domain and the SAB, Crowd, UBC, Vaihingen, Massachusetts, Potsdam, and Inria datasets as target domains to evaluate the performance of our MASDG and compare it with existing state-of-the-art SDG methods. The performance results are shown in Table II. The overall performance of Rein and our proposed MASDG on seven target domains far exceeded previous SDG methods, which may be attributed to both Rein and our MASDG utilizing VFM as a backbone for feature extraction. Since VFM is pretrained on a large-scale dataset, it possesses significant feature abstraction and generalization capabilities. As a result, the VFM-based SDG methods demonstrate better generalization when applied to unknown RS target domains. In addition, compared with Rein, our MASDG

TABLE II
PERFORMANCE COMPARISON BETWEEN THE PROPOSED MASDG (OURS) AND EXISTING SDG METHODS ON WHU→OTHERS

Method	SAB		Crowd		UBC		Vaihingen		Massachusetts		Potsdam		Inria		AVG	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
RobustNet [53]	33.11	49.75	45.03	62.10	23.09	37.52	47.53	64.44	12.67	22.50	47.53	64.44	46.53	63.51	36.50	52.04
MDGVR [25]	20.63	34.21	42.4	59.27	20.97	34.67	39.08	56.2	13.6	23.94	33.62	50.21	38.44	55.53	29.82	44.86
MVDG [26]	26.32	41.67	49.13	65.89	24.53	39.4	49.36	66.1	15.96	27.53	38.93	56.04	41.88	59.03	35.16	50.81
SHADE [54]	37.37	54.41	39.51	56.64	19.99	33.32	51.84	68.28	18.66	31.44	53.54	69.74	45.54	62.58	38.06	53.77
SiamDoGe [55]	37.36	54.40	50.55	67.16	27.50	43.13	54.86	70.85	17.28	29.47	45.50	62.50	45.77	62.80	39.83	55.76
Dual-Level [56]	34.47	52.86	44.15	62.28	21.69	35.70	50.14	65.62	16.44	26.66	43.70	60.85	43.86	59.99	36.35	51.99
HGFormer [57]	30.18	46.36	46.59	63.56	12.68	22.50	52.50	68.85	32.91	49.53	45.57	62.61	51.58	68.05	38.86	54.50
DIIA [58]	31.84	48.31	40.17	57.32	25.53	40.67	41.85	59.00	14.91	25.95	49.50	66.22	42.87	60.01	35.24	51.07
CCDR [6]	34.62	51.43	51.30	67.81	30.17	46.35	60.01	75.01	14.02	24.60	49.58	66.29	43.23	60.36	40.42	55.98
BlindNet [59]	34.04	50.79	52.22	68.61	30.50	46.75	51.00	67.55	10.46	18.94	50.01	66.68	45.83	62.86	39.15	54.60
CMFormer [60]	38.72	55.82	42.03	59.19	27.4	43.01	60.65	75.5	40.4	57.23	63.18	77.44	53.58	69.78	46.56	62.56
CPerb [37]	32.81	49.41	43.77	60.89	23.14	37.58	54.36	70.43	16.15	27.81	50.32	66.95	44.10	61.21	37.81	53.47
Rein (baseline) [36]	54.26	70.35	73.43	84.64	37.26	54.50	80.86	89.34	33.94	51.92	78.88	88.19	63.49	77.67	60.30	73.80
Ours	**59.47**	**74.58**	**79.46**	**88.55**	**52.05**	**68.47**	**84.13**	**91.38**	**42.50**	**59.65**	**81.99**	**90.10**	**67.49**	**80.59**	**66.73**	**79.04**

TABLE III
PERFORMANCE COMPARISON BETWEEN THE PROPOSED MASDG (OURS) AND EXISTING SDG METHODS ON SAB→OTHERS

Method	WHU		Crowd		UBC		Vaihingen		Massachusetts		Potsdam		Inria		AVG	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
RobustNet [53]	50.94	67.49	52.59	68.93	60.55	75.42	54.41	70.47	25.04	40.05	57.87	73.31	59.56	74.66	51.57	67.19
MDGVR [25]	36.26	53.22	46.06	63.72	40.92	57.93	47.5	64.12	16.76	28.71	41.25	58.62	50.42	67.04	39.88	56.19
MVDG [26]	37.13	54.15	50.94	67.49	42.41	59.56	56.05	71.84	24.8	39.75	42.62	59.77	58.19	73.57	44.59	60.88
SHADE [54]	43.92	61.03	50.60	67.20	56.15	71.92	56.15	71.92	27.02	42.55	54.98	70.95	58.66	73.94	49.64	65.64
SiamDoGe [55]	43.98	61.09	50.88	67.45	49.74	66.43	59.84	74.87	27.26	42.84	49.93	66.60	51.22	67.74	47.55	63.86
Dual-Level [56]	45.70	62.73	52.95	69.24	50.04	66.70	53.25	69.49	29.96	46.22	50.54	67.14	56.08	71.86	48.36	64.77
HGFormer [57]	60.70	75.55	54.86	70.85	57.00	72.62	51.86	68.30	**38.44**	**55.54**	48.21	65.06	64.28	78.25	53.62	69.45
DIIA [58]	49.56	66.27	53.47	69.68	60.02	75.02	59.31	74.46	22.5	36.74	59.13	74.31	59.26	74.42	51.89	67.27
CCDR [6]	44.50	61.59	50.84	67.41	48.09	64.95	64.09	78.12	28.42	44.26	51.83	68.28	50.76	67.34	48.36	64.56
BlindNet [59]	52.01	68.51	54.66	70.69	**65.42**	**79.10**	64.75	78.60	23.26	37.74	46.66	63.63	57.66	73.14	52.06	67.34
CMFormer [60]	55.45	71.34	55.54	71.42	52.5	68.85	49.08	65.84	34.33	50.96	51.68	68.14	63.9	77.97	51.78	67.79
CPerb [37]	50.04	66.70	51.37	67.87	47.71	64.60	55.18	71.11	26.29	41.63	47.61	64.51	53.22	69.47	47.35	63.70
Rein (baseline) [36]	62.66	77.05	70.61	82.77	58.25	73.61	81.18	89.61	35.13	52.00	83.54	91.03	**68.03**	**80.98**	65.63	78.15
Ours	**64.70**	**78.57**	**74.92**	**85.66**	62.02	76.56	**83.33**	**90.91**	35.18	52.05	**83.60**	**91.07**	67.66	80.71	**67.34**	**79.36**

further relatively improved the average IoU and $F1$ on seven target domains by 10.02% and 6.75%, respectively, achieving state-of-the-art performance on all RS target domains. This improvement is attributed to the novel multiview augmentation and semantic-invariant learning strategy we proposed, which effectively increases the texture and style diversity of RS source domain training data and learn semantic-invariant building representations that mitigate semantics inconsistency. As a result, the RSBE model employed VFM as a backbone can effectively obtain domain-invariant representations from diverse source domain training data, which further improves its generalization ability on target domains.

2) SAB→Others: We employed the SAB dataset as the source domain and the WHU, Crowd, UBC, Vaihingen, Massachusetts, Potsdam, and Inria datasets as multiple target domains, to compare our proposed MASDG with existing state-of-the-art methods. As can be observed from Table III, our proposed MASDG achieves the best overall performance on all the target domains. Compared to previous SDG methods employing ResNet as the backbone, our MASDG improves the average IoU and $F1$ by more than 15% and 10%, respectively. Compared with Rein, that also employing VFM as backbone, our method further relatively improves the average IoU and $F1$ by 2.61% and 1.55%, respectively, on all target domains. It is proven that the multiview augmentation and SRL strategy can effectively mitigate the texture and style divergencies between different RS source and target domains and enforce semantic consistency, thus leading to improved generalization performance across various target domains. However, our proposed MASDG is slightly inferior to BlindNet and HGFormer on SAB→UBC and SAB→Massachusetts, respectively. This may arise from the small and dense buildings contained in these two target domains, making it difficult for our model to accurately identify object boundaries during inference.

3) Crowd→Others: We evaluated MASDG on different RS domains using the Crowd dataset as source domain and SAB, WHU, UBC, Vaihingen, Massachusetts, Potsdam, and Inria datasets as target domains. As can be seen from the performance in Table IV, our proposed MASDG outperforms other compared methods, obtaining the best average IoU of 64.38% and the best average $F1$ of 77.19%. Compared with Rein, which performs best among previous SDG methods, our method further relatively improves the average IoU and $F1$ by 3.37% and 2.40%, respectively, on all target domains. This indicates that our MASDG can effectively address the domain shift problem between various RS domains, thus generalizing better to arbitrary unseen target domains.

4) Analysis of Computational Efficiency: To thoroughly and fairly assess the computational efficiency of our proposed method, we performed comparative experiments on the WHU→Others setting in terms of trainable parameters and inference speed under different backbone architectures. As

TABLE IV
PERFORMANCE COMPARISON BETWEEN THE PROPOSED MASDG (OURS) AND EXISTING SDG METHODS ON CROWD→OTHERS

Method	SAB		WHU		UBC		Vaihingen		Massachusetts		Potsdam		Inria		AVG	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
RobustNet [53]	43.37	60.50	62.74	77.10	48.01	64.87	54.67	70.69	27.22	42.79	57.79	73.25	55.11	71.06	49.84	65.75
MDGVR [25]	34.03	50.78	47.58	64.48	37.81	54.96	51.83	67.66	21.98	34.44	46.11	63.12	49.83	66.51	41.31	57.42
MVDG [26]	42.06	59.22	50.47	67.09	43.37	60.5	59.03	74.24	24.34	39.16	52.13	68.53	51.53	68.01	46.13	62.39
SHADE [54]	39.29	56.42	46.01	63.02	34.97	51.81	48.31	64.99	14.94	26.00	35.13	51.99	41.23	58.39	37.13	53.23
SiamDoGe [55]	44.47	61.56	48.95	65.73	46.25	63.25	52.39	68.76	29.01	44.98	58.42	73.75	50.11	66.76	47.09	63.54
Dual-Level [56]	41.17	58.32	46.63	63.60	40.44	57.60	58.84	74.09	27.40	43.02	56.17	71.93	46.17	63.18	45.26	61.68
HGFormer [57]	38.78	55.89	65.45	79.12	43.40	77.84	55.88	71.10	**36.81**	**53.81**	56.51	72.21	54.06	70.18	50.13	68.59
DIIA [58]	44.5	61.59	58.63	73.92	49.78	66.47	54.88	70.87	27.89	43.62	57.83	73.28	55.16	71.1	49.81	65.84
CCDR [6]	39.69	56.82	50.40	67.02	38.52	55.62	58.79	74.05	26.52	41.92	49.41	66.14	47.71	64.60	44.43	60.88
BlindNet [59]	41.49	58.65	55.37	71.28	45.79	62.81	59.06	74.26	18.94	31.85	61.23	75.96	52.23	68.62	47.73	63.35
CMFormer [60]	40.45	57.6	65.81	79.38	40.17	57.31	53.23	69.47	34.67	51.49	62.01	76.55	47.21	64.14	49.08	65.13
CPerb [37]	42.04	59.20	57.63	73.12	41.40	58.56	54.48	70.53	28.81	44.73	56.65	72.32	49.65	66.36	47.24	63.55
Rein (baseline) [36]	54.61	70.64	**69.57**	**82.05**	52.84	69.14	79.36	88.49	31.31	47.69	83.22	90.84	65.04	78.82	62.28	75.38
Ours	**57.10**	**72.69**	68.90	81.59	**56.24**	**71.99**	**82.69**	**90.52**	35.79	52.72	**83.81**	**91.19**	**66.15**	**79.63**	**64.38**	**77.19**

TABLE V
COMPUTATIONAL EFFICIENCY COMPARISON OF DIFFERENT METHODS ON WHU→OTHERS

Methods	Backbone	Params (M) ↓	Inference speed(s) ↓	Average IoU ↑
RobustNet [53]	ResNet50	45.068	0.0125	36.50
MDGVR [25]	ResNet50	39.045	0.0249	29.82
MVDG [26]	ResNet50	40.224	0.5461	35.16
SHADE [54]	ResNet50	45.068	0.0110	38.06
SiamDoGe [55]	ResNet50	40.224	0.0094	39.83
Dual-level [56]	ResNet50	47.117	0.0122	36.35
DIIA [58]	ResNet50	51.698	0.0381	35.24
CCDR [6]	ResNet50	39.238	0.0097	40.42
BlindNet [59]	ResNet50	50.092	0.0095	39.15
CPerb [37]	ResNet50	40.224	0.0361	37.81
Ours	ResNet50	**22.09**	**0.0087**	**46.85**
HGFormer [57]	Swin-T	51.572	0.0996	38.86
CMFormer [60]	Swin-T	48.291	0.0849	46.56
Ours	Swin-T	**23.59**	**0.0186**	**50.41**
Rein (baseline) [36]	DINOv2	23.590	0.0334	60.30
Ours	DINOv2	**23.590**	**0.0333**	**66.73**

presented in Table V, our proposed method inherently achieves the fewest trainable parameters by utilizing a frozen backbone, which can be well-adapted to low-computing-power scenarios (e.g., edge devices or lightweight GPUs) that are common in real-world applications. For inference speed, we quantified the time taken by each method to process a single image. Table V shows our method consistently achieves comparable or better inference speeds than existing approaches across various backbone configurations. More importantly, our method delivers superior generalization performance in all backbone setups. Overall, our method not only excels in generalization capability but also maintains competitive efficiency across different backbone architectures.

E. Ablation Studies and Discussions

1) Effect of Different Components: In order to evaluate the effect of each component in our MASDG, we progressively added TDA, SDA, and SRL to the baseline Rein framework on the WHU→Others setting, the results are shown in Table VI. Based on the baseline, adding the TDA module alone relatively increases the average IoU on all target domains by 8.36% and the average $F1$ by 5.50%. Adding the SDA module alone relatively increases the average IoU on all target domains by 1.51% and the average $F1$ by 1.65%. The integration of TDA and SDA attains relative performance gains of 9.87% in average IoU and 6.56% in average $F1$ on all target domains, demonstrating the complementary of multiview texture-style augmentation by tackling both texture-level and style-level domain shift between different RS domains, overcoming the limitations of the ones that solely focus on one aspect, thereby effectively mitigating the limited diversity of a single-source domain. Finally, the further addition of SRL lead to a relative performance improvement of 10.67% in average IoU and 7.10% in average $F1$, which proves the advantage of the SRL strategy in preserving semantic consistency, thus more effectively leveraging the diverse source domain training data obtained by TDA and SDA for robust building extraction.

2) Effect of Different Texture Augmentation Methods: To verify the effect of different texture augmentation methods used in TDA, Table VII shows the performance results of using three different texture augmentation methods under the WHU→Others setting. Our proposed TDA, RandConv [12], and Pro-RandConv [61] all use random convolution to transform the texture of the source input image. The difference lies in the parameter setting of random convolution. RandConv uses a single layer of random convolution with a kernel size is randomly selected. Pro-RandConv uses multiple layers of random convolution with the same weights, and the size of the convolution kernel and the number of convolution layers are randomly selected. Our proposed TDA uses parameter-optimized multilayer random convolution for texture augmentation. The key texture transformation parameters are not random, but from a well-designed parameter optimization strategy, which diversifies the combination of texture transformation parameters, thereby increasing the texture diversity of the RS source domain. As can be seen from Table VII, our method outperforms the other two texture augmentation methods, showing the highest texture diversity of images generated by TDA and the most superior generalization performance across multiple RS target domains.

3) Effect of Different Style Transformation Strategies in SDA: To investigate the effect of different style transformation

TABLE VI
EFFECT OF DIFFERENT COMPONENTS ON WHU→OTHERS

Component				SAB		Crowd		UBC		Vaihingen		Massachusetts		Potsdam		Inria		AVG	
Baseline	TDA	SDA	SRL	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
✓				54.26	70.35	73.43	84.64	37.26	54.50	80.86	89.34	33.94	51.92	78.88	88.19	63.49	77.67	60.30	73.80
✓	✓			**59.82**	**74.86**	78.87	88.18	50.68	67.27	84.25	91.45	37.70	54.76	79.62	88.66	66.45	79.84	65.34	77.86
✓		✓		54.56	70.6	76.48	86.67	36.15	53.1	81.4	89.75	36.92	53.93	80.47	89.18	64.81	78.65	61.54	74.55
✓	✓	✓		59.22	74.39	**79.59**	**88.64**	50.80	67.38	**84.39**	**91.53**	41.46	58.61	81.07	89.54	67.24	80.41	66.25	78.64
✓	✓	✓	✓	59.47	74.58	79.46	88.55	**52.05**	**68.47**	84.13	91.38	**42.50**	**59.65**	**81.99**	**90.10**	**67.49**	**80.59**	**66.73**	**79.04**

TABLE VII
EFFECT OF DIFFERENT TEXTURE AUGMENTATION METHODS IN TDA ON WHU→OTHERS

Methods	SAB		Crowd		UBC		Vaihingen		Massachusetts		Potsdam		Inria		AVG	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
Baseline	54.26	70.35	73.43	84.64	37.26	54.50	80.86	89.34	33.94	51.92	78.88	88.19	63.49	77.67	60.30	73.80
RandConv [12]	**59.71**	**74.77**	77.97	87.62	49.88	66.56	83.65	91.10	38.63	55.73	81.93	90.07	66.53	79.90	65.47	77.96
Pro-RandConv [61]	58.12	73.52	**79.75**	**88.73**	50.86	67.42	**84.16**	**91.40**	36.10	53.04	80.60	89.26	66.80	80.09	65.20	77.64
Ours	59.47	74.58	79.46	88.55	**52.05**	**68.47**	84.13	91.38	**42.50**	**59.65**	**81.99**	**90.10**	**67.49**	**80.59**	**66.73**	**79.04**

TABLE VIII
EFFECT OF DIFFERENT STYLE TRANSFORMATION STRATEGIES IN SDA ON WHU→OTHERS

Methods	SAB		Crowd		UBC		Vaihingen		Massachusetts		Potsdam		Inria		AVG	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
Baseline	54.26	70.35	73.43	84.64	37.26	54.50	80.86	89.34	33.94	51.92	78.88	88.19	63.49	77.67	60.30	73.80
MixStyle [62]	**59.71**	**74.77**	77.97	87.62	49.88	66.56	83.65	91.10	38.63	55.73	81.93	90.07	66.53	79.90	65.47	77.96
TFS-Token [63]	59.32	74.47	78.96	88.24	**52.24**	**68.62**	83.63	91.09	38.28	55.36	81.50	89.81	66.48	79.86	65.77	78.21
Ours	59.47	74.58	**79.46**	**88.55**	52.05	68.47	**84.13**	**91.38**	**42.50**	**59.65**	**81.99**	**90.10**	**67.49**	**80.59**	**66.73**	**79.04**

TABLE IX
EFFECT OF DIFFERENT SEMANTIC CONSISTENCY LOSSES IN SRL ON WHU→OTHERS

Losses	SAB		Crowd		UBC		Vaihingen		Massachusetts		Potsdam		Inria		AVG	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
Baseline	54.26	70.35	73.43	84.64	37.26	54.50	80.86	89.34	33.94	51.92	78.88	88.19	63.49	77.67	60.30	73.80
L1 [64]	59.03	74.24	79.17	88.37	50.50	67.11	84.50	91.60	39.22	56.34	**82.01**	**90.12**	66.98	80.23	65.92	78.29
L2 [65]	59.23	74.47	79.22	88.42	52.02	68.11	**84.62**	**91.59**	40.95	56.80	80.58	89.84	67.11	80.14	66.25	78.48
Ours	**59.47**	**74.58**	**79.46**	**88.55**	**52.05**	**68.47**	84.13	91.38	**42.50**	**59.65**	81.99	90.10	**67.49**	**80.59**	**66.73**	**79.04**

strategies used in SDA for feature stylization, Table VIII shows the comparative results of three different style transformation strategies, MixStyle, TFS-Token, and our proposed BST under the WHU→Others setting. MixStyle [62] swaps the feature statistics of samples in the same batch to generate intermediate features with new styles. The style transfer mechanism of TFS-Token [63] is similar to MixStyle, but TFS-Token retains some of the original features. In the experimental setting, we constrained the feature layers for the style transfer of MixStyle and BST kept consistent. Since TFS-Token needs to randomly specify the feature layers used for style transfer, we guarantee that the number of layers used for style transfer in TFS-Token is consistent with MixStyle and BST. As can be observed from Table VIII, the BST used in our MASDG achieves the best performance, showing that our BST can effectively diversify the style of features by modeling the feature statistics as uncertain variables sampled from multivariate Gaussian distributions, thus effectively mitigating the uncertain style-level domain shift between different RS domains.

4) Effect of Different Semantic Consistency Loss Functions in SRL: To evaluate the effect of different semantic consistency loss functions used in SRL for semantic alignment of multiview predictions, Table IX shows the impact of using three different semantic consistency loss functions in our SRL on the model performance under the WHU→Others setting. We conducted experiments using three commonly used loss functions for similarity measurement, Manhattan distance (L1) [64], Euclidean distance (L2) [65], and JS divergence (JS) [40] used in our MASDG. Among the three different semantic consistency loss functions, the model obtains the best performance using JS divergence, which indicates that the JS divergence used in our SRL can effectively regularize the discrepancy among multiview predictions and mitigate semantic inconsistency.

5) Effect of the Key Parameters in TDA: To evaluate the effect of key parameters in TDA, we adjusted different values for the queue length *len* of Q_M, Q_k, and Q_α, along with the upper bound M_{upper} of the number of random convolution layers M on the model performance and provided the results in Figs. 4 and 5, respectively. Fig. 4(a) and (b) shows the average IoU and average $F1$ of the model on all target domains by varying *len* under the WHU→Others setting. As the *len*

TABLE X
GENERALIZATION COMPARISONS WITH STATE-OF-THE-ART BUILDING EXTRACTION METHODS ON WHU→OTHERS

Methods	SAB		Crowd		UBC		Vaihingen		Massachusetts		Potsdam		Inria		AVG	
	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
UANet [66]	17.75	30.15	41.54	58.70	13.09	23.14	30.54	46.79	21.94	35.98	38.72	55.83	49.02	65.79	30.37	45.20
LWGANet [67]	35.56	52.46	32.94	49.55	2.46	4.80	26.23	41.56	28.85	44.78	35.19	52.06	52.99	69.27	30.60	44.93
Ours	**59.47**	**74.58**	**79.46**	**88.55**	**52.05**	**68.47**	**84.13**	**91.38**	**42.50**	**59.65**	**81.99**	**90.10**	**67.49**	**80.59**	**66.73**	**79.04**

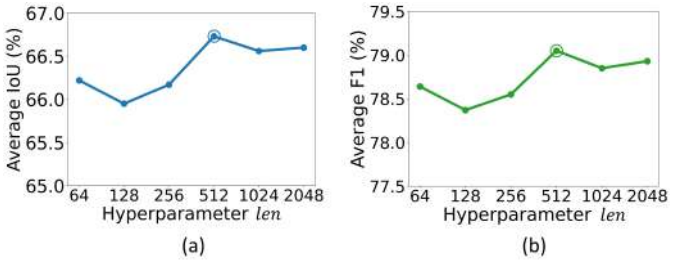

Fig. 4. Model performance with the change of queue length *len* in TDA. (a) Average IoU across all target domains. (b) Average $F1$ across all target domains.

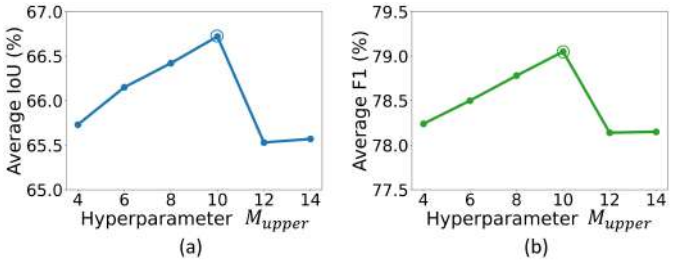

Fig. 5. Model performance with the change of M_{upper} in TDA, which is the upper bound of the number of random convolution layers. (a) Average IoU across all target domains. (b) Average $F1$ across all target domains.

increases, both IoU and $F1$ exhibit an increasing trend, and then achieve a fairly stable once the *len* exceeds 512. This is because when *len* is too small, it will lead to incomplete parameter space exploration, resulting in limited texture diversity of generated images. Similarly, Fig. 5(a) and (b) presents the average IoU and average $F1$ of the model across all target domains, respectively, under the WHU→Others setting as M_{upper} is changed. The result indicates that, as the M_{upper} increases, the overall generalization performance will initially increase and then subsequently decrease. This is because too small M_{upper} leads to limited texture diversity of texture-augmented images and too large M_{upper} may yield excessive M that impairs image semantics when performing random convolution. Therefore, we set *len* as 512 and M_{upper} as 10 for obtaining the optimal performance.

6) Effect of the Weighting Factor ω of $\mathcal{L}_{Cons}$ in SRL: In order to assess the effect of the weighting factor ω of $\mathcal{L}_{\text{cons}}$ in SRL on the performance [see (10)], we experimented with different values and analyzed the results. Fig. 6(a) and (b) shows the average IoU and average $F1$ of the model on all target domains by varying the value of ω under the WHU→Others setting. We can observe when ω is too small, the model cannot efficiently deal with semantic disparity between different RS domains, leading to poor performance. In contrast, when ω is too large, it will lead to insufficient

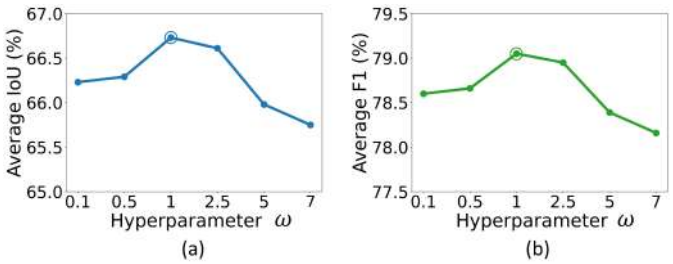

Fig. 6. Model performance with the change of the weight factor ω of $\mathcal{L}_{\text{cons}}$ in SRL. (a) Average IoU across all target domains. (b) Average $F1$ across all target domains.

domain-invariant representation learning, also resulting in poor performance. Therefore, we set the ω as 1.0 since it reaches the performance peak as shown in Fig. 6.

F. Generalization Comparison With State-of-the-Art Building Extraction Methods

To compare the generalization performance of our proposed MASDG and current state-of-the-art building extraction methods on unseen target domains, we selected the latest open-source building extraction methods, including UANet [66] and LWGANet [67] for generalization comparison under the WHU→Others cross-domain setting. The results are presented in Table X. Despite high testing performance on the source domain, UANet and LWGANet show poor generalization performance when directly testing on unseen target domains. This is due to the domain shift caused by the difference of imaging mechanisms, sensors, environments, and location between the source and target domains. In contrast, our proposed method demonstrates superior performance in handling domain shift compared to these current state-of-the-art building extraction methods, which is attribute to the design of multiview augmentation and semantic-invariant learning that effectively mitigates the domain shift, showing strong adaptability and robustness in more realistic cross-domain scenarios that domain shift is inevitable.

G. Visual Experimental Results and Discussions

1) Visualization Results on the WHU→Others: We compared the building segmentation results of our MASDG and other SDG methods on the WHU→Others setting, as shown in Fig. 7, where white represented the building and black represented the background. Since the WHU dataset is collected from a single region with low spatial resolution, the buildings are characterized by high density, small scale, and high similarity, resulting in an obvious domain shift from

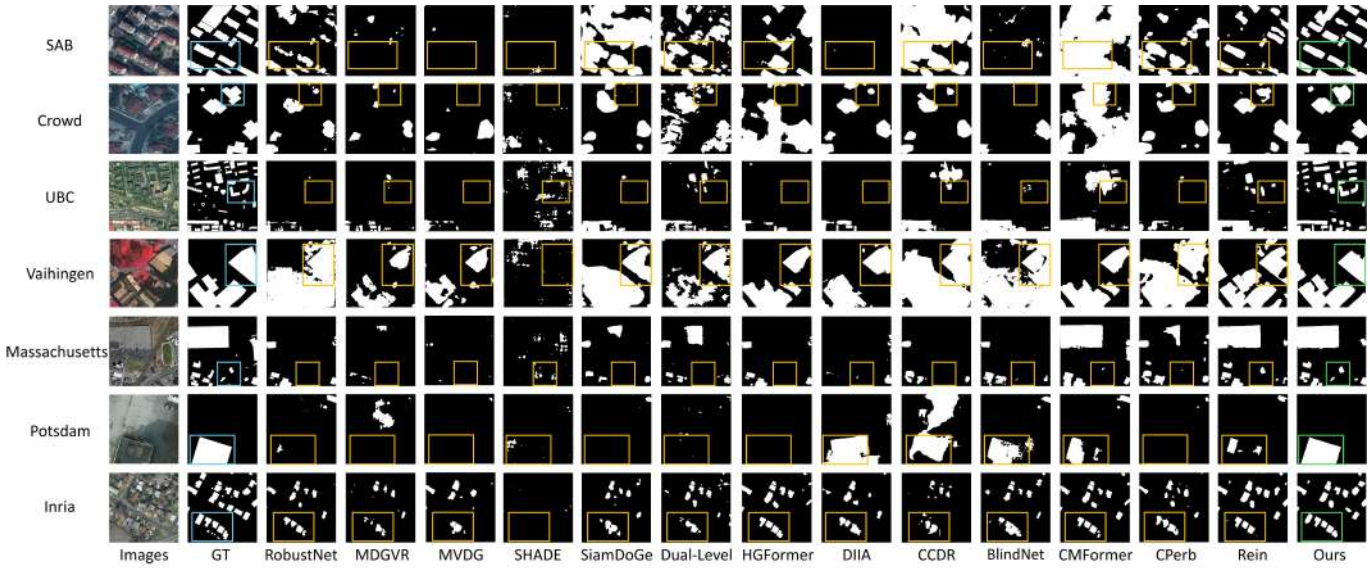

Fig. 7. Visualization results of different SDG methods under WHU→Others. The first column represents the testing images from various target domains, the second column indicates the GT, and the rest columns show the predictions obtained by different SDG methods and our proposed MASDG, respectively. We use boxes to mark areas with significant differences between different methods, where the blue boxes indicate the GT, the yellow boxes point out regions where previous SDG methods have missed detections or false detections, and the green boxes highlight the superiority of our proposed MASDG in more accurately identifying buildings.

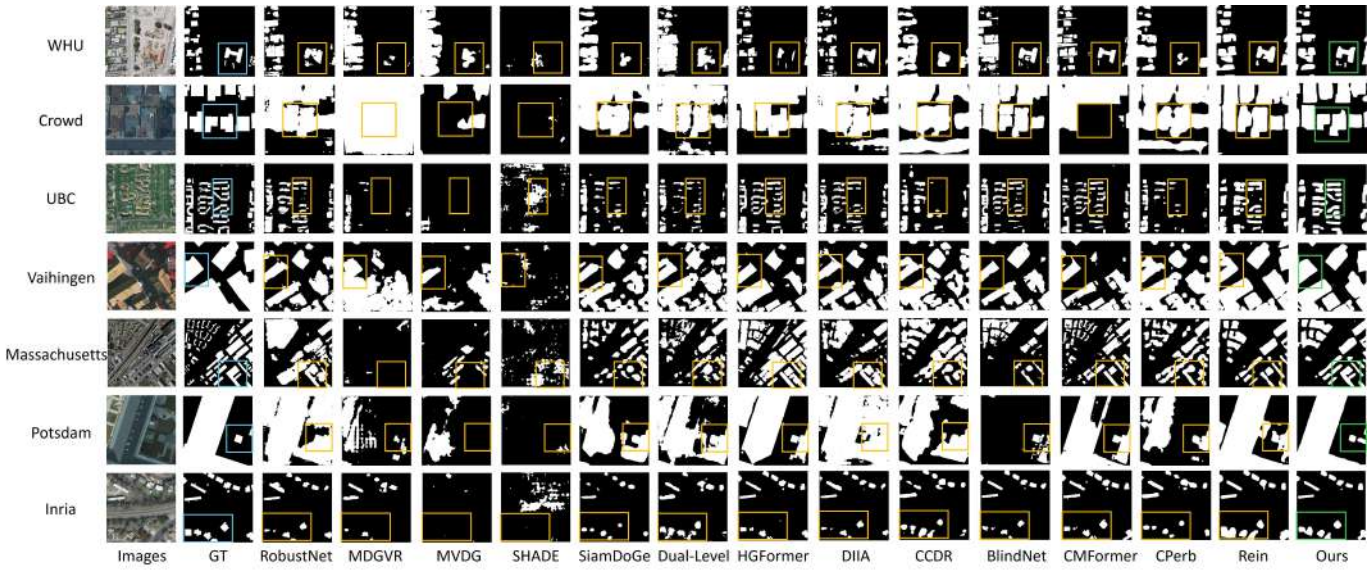

Fig. 8. Visualization results of different SDG methods under SAB→Others. The first column represents the testing images from various target domains, the second column indicates the GT, and the rest columns show the predictions obtained by different SDG methods and our proposed MASDG, respectively. We use boxes to mark areas with significant differences between different methods, where the blue boxes indicate the GT, the yellow boxes point out regions where previous SDG methods have missed detections or false detections, and the green boxes highlight the superiority of our proposed MASDG in more accurately identifying buildings.

other RS domains. The comparison methods are far from satisfactory and cannot effectively distinguish between building and background (see the 3rd~14th columns of Fig. 7). This is primarily because they only address either texture- or style-level domain shift, failing to effectively and comprehensively tackle the domain shift between different RS domains. Especially for the UBC domain (see the 3rd row of Fig. 7), these methods exhibit significant missed detections due to the structural, stylistic, and scaling differences between WHU and UBC.

In contrast, the segmentation results of Rein have been greatly improved. This is because Rein benefits from the generalization ability of pretrained VFMs for unknown scenes, and further refines the feature maps at the object-level for each instance, achieving better segmentation results. However, compared with our method, Rein still suffers from issues such as missed building detections, incomplete building segmentation, and blurred boundaries, as seen at the 15th column of Fig. 7. Based on Rein, our MASDG introduces the multiview augmentation and SRL to address both texture- and style-level RS domain shifts. Therefore, MASDG can generate more building instances, more complete building masks, and clearer building boundaries, which closely resembles the GT, see the 16th row of Fig. 7.

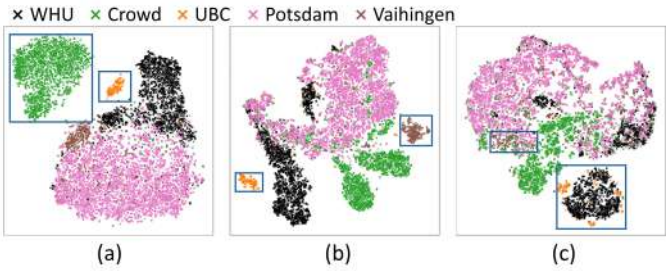

Fig. 9. Comparison of feature distributions using different SDG methods under SAB→Others. (a) BlindNet. (b) Rein. (c) Our MASDG. The more uniform distribution of features across target domains indicates better results.

2) Visualization Results on the SAB→Others: Fig. 8 illustrates the comparison between our MASDG and other SDG methods on the SAB→Others setting. Although previous comparison methods can also detect some building instances, the generated masks often lacked accuracy. They may either fail to completely encompass the building instances or include excessive background, see the annotated areas at 3rd∼14th columns of Fig. 8, particularly for the Crowd domain (see the 2nd row). In contrast, Rein reduces the false detection rate, however, it still faces issues such as incomplete masks for building instances and missed detections, see the 15th column of Fig. 8. Our MASDG can effectively distinguish buildings from their surrounding backgrounds, reduce the missed detections, and generate complete segmentation masks with clear boundaries.

3) Feature Distribution Visualization: To analyze the generalization performance of different SDG methods on target domains, Fig. 9(a)–(c) compared the feature distributions extracted by BlindNet, Rein, and our proposed MASDG method under the SAB→Others setting. As shown in Fig. 9, the feature distribution generated by our proposed MASDG for various unseen RS target domains is more uniform, especially within the annotated regions. Notably, the features of the Crowd and UBC datasets extracted by BlindNet were dispersed from others, and the features of the UBC and Vaihingen datasets extracted by Rein were also dispersed from the rest, whereas our MASDG yielded a more uniform distribution. This observation highlights the superiority of MASDG in extracting domain-invariant features, contributing to its superior generalization performance on unseen RS target domains.

V. Conclusion

In this article, we proposed **MASDG**, a novel framework for the challenging MD-RSBE task, which transfers knowledge from a single RS source domain to multiple unlabeled target domains. MASDG introduces a multiview augmentation strategy that captures both texture- and style-level domain shifts, and employs SRL to address inconsistencies across views. Extensive experiments on three cross-domain RS building datasets demonstrate that MASDG outperforms existing SDG methods, achieving state-of-the-art performance. We also validated the contributions of each component and visualized segmentation and feature distributions. While MASDG effectively mitigates domain shifts, its performance drops with extremely small-scale objects. Future work will explore multiscale feature enhancement and prompt-based strategies with pretrained large models to further improve model-level generalization.

References

[1] W. Qiu, L. Gu, F. Gao, and T. Jiang, "Building extraction from very high-resolution remote sensing images using refine-UNet," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[2] J. Bai et al., "Building extraction from high-resolution remote sensing images using improved HRNet method," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2024, pp. 7982–7985.

[3] B. Song, W. Shao, P. Shao, J. Wang, J. Xiong, and C. Qi, "DHI-Net: A novel detail-preserving and hierarchical interaction network for building extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[4] D. Peng, H. Guan, Y. Zang, and L. Bruzzone, "Full-level domain adaptation for building extraction in very-high-resolution optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607317.

[5] F. Zhang et al., "Multitarget domain adaptation building instance extraction of remote sensing imagery with domain-common approximation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4702916.

[6] C. Liang, W. Li, Y. Dong, and W. Fu, "Single domain generalization method for remote sensing image segmentation via category consistency on domain randomization," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5616116.

[7] Y. Li, L. Yuan, and N. Vasconcelos, "Bidirectional learning for domain adaptation of semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6936–6945.

[8] S. Ji, D. Wang, and M. Luo, "Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3816–3828, May 2021.

[9] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5400–5409.

[10] R. Iizuka, J. Xia, and N. Yokoya, "Frequency-based optimal style mix for domain generalization in semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4501114.

[11] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Learning to generate novel domains for domain generalization," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 561–578.

[12] Z. Xu, D. Liu, J. Yang, C. Raffel, and M. Niethammer, "Robust and generalizable visual representation learning via random convolutions," in *Proc. Int. Conf. Learn. Represent.*, 2021. [Online]. Available: https://openreview.net/forum?id=BVSM0x3EDK6

[13] D. Rui, K. Guo, X. Zhu, Z. Wu, and H. Fang, "Progressive diversity generation for single domain generalization," *IEEE Trans. Multimedia*, vol. 26, pp. 10200–10210, 2024.

[14] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 814–823.

[15] Y. Meng et al., "Cross-domain land cover classification of remote sensing images based on full-level domain adaptation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 11434–11450, 2024.

[16] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.

[17] J. Cai and Y. Chen, "MHA-Net: Multipath hybrid attention network for building footprint extraction from high-resolution remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5807–5817, 2021.

[18] X. Shi, H. Huang, C. Pu, Y. Yang, and J. Xue, "CSA-UNet: Channel-spatial attention-based encoder–decoder network for rural blue-roofed building extraction from UAV imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[19] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2004612.

[20] L. Xu, Y. Li, J. Xu, Y. Zhang, and L. Guo, "BCTNet: Bi-branch cross-fusion transformer for building footprint extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4402014.

[21] P. Zhu, Z. Song, J. Liu, J. Yan, X. Luo, and Y. Tao, "MSHFormer: A multiscale hybrid transformer network with boundary enhancement for VHR remote sensing image building extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5616316.

[22] M. Yang, L. Zhao, L. Ye, W. Jia, H. Jiang, and Z. Yang, "EGAFNet: An edge guidance and scale-aware adaptive fusion network for building extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 4700513.

[23] Y. Na, J. H. Kim, K. Lee, J. Park, J. Y. Hwang, and J. P. Choi, "Domain adaptive transfer attack-based segmentation networks for building extraction from aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5171–5182, Jun. 2021.

[24] L. Shi, Z. Wang, B. Pan, and Z. Shi, "An end-to-end network for remote sensing imagery semantic segmentation via joint pixel- and representation-level domain adaptation," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 11, pp. 1896–1900, Nov. 2021.

[25] L. Niu, W. Li, and D. Xu, "Multi-view domain generalization for visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4193–4201.

[26] J. Zhang, L. Qi, Y. Shi, and Y. Gao, "MVDG: A unified multi-view framework for domain generalization," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 161–177.

[27] D. Peng, Y. Lei, L. Liu, P. Zhang, and J. Liu, "Global and local texture randomization for synthetic-to-real semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 6594–6608, 2021.

[28] S. Lee, H. Seong, S. Lee, and E. Kim, "WildNet: Learning domain generalized semantic segmentation from the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9936–9946.

[29] C. Li, D. Zhang, W. Huang, and J. Zhang, "Cross contrasting feature perturbation for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1327–1337.

[30] M. Luo, S. Ji, and S. Wei, "A diverse large-scale building dataset and a novel plug-and-play domain generalization method for building extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4122–4138, 2023.

[31] E. Durakli, P. Bosilj, C. Fox, and E. Aptoula, "A domain generalized mask R-CNN for building instance segmentation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2024, pp. 9684–9687.

[32] L. Li et al., "Progressive domain expansion network for single domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 224–233.

[33] K. Guo et al., "Single domain generalization via unsupervised diversity probe," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 2101–2111.

[34] B. Wang, Y. Xu, Z. Wu, S. Zheng, Z. Wei, and J. Chanussot, "Hyperspectral images single-source domain generalization based on nonlinear sample generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5516613.

[35] X. Wang, J. Liu, Y. Ni, W. Chi, and Y. Fu, "Two-stage domain alignment single-source domain generalization network for cross-scene hyperspectral images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5527314.

[36] Z. Wei et al., "Stronger, fewer, & superior: Harnessing vision foundation models for domain generalized semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 28619–28630.

[37] D. Zhao, L. Qi, X. Shi, Y. Shi, and X. Geng, "A novel cross-perturbation for single domain generalization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 11, pp. 10903–10916, Nov. 2024.

[38] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[39] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1290–1299.

[40] P. W. Lamberti, A. P. Majtey, M. Madrid, and M. E. Pereyra, "Jensen–Shannon divergence: A multipurpose distance for statistical and quantum mechanics," *AIP Conf. Proc.*, vol. 913, pp. 32–37, Jan. 2007.

[41] W. Kaishun et al., "A dataset of building instances of typical cities in China," *China Sci. Data*, vol. 6, no. 1, pp. 182–190, 2021.

[42] S. P. Mohanty et al., "Deep learning for understanding satellite imagery: An experimental survey," *Frontiers Artif. Intell.*, vol. 3, Nov. 2020, Art. no. 534696.

[43] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource 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.

[44] X. Huang et al., "Urban building classification (UBC)—A dataset for individual building detection and classification from satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1412–1420.

[45] (2025). *Potsdam*. Accessed: Jan. 19, 2025. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/Default.aspx

[46] *Vaihingen*. Accessed: Jan. 19, 2025. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/Default.aspx

[47] V. Mnih, *Machine Learning for Aerial Image Labeling*. Toronto, ON, Canada: University of Toronto, 2013.

[48] *Inria*. Accessed: Jan. 19, 2025. [Online]. Available: https://project.inria.fr/aerialimagelabeling/

[49] MMSegmentation Contributors.(2020). *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. [Online]. Available: https://github.com/open-mmlab/mmsegmentation

[50] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408715.

[51] *F1-Score*. Accessed: Jan. 19, 2025. [Online]. Available: https://en.wikipedia.org/wiki/F-score

[52] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*.

[53] S. Choi, S. Jung, H. Yun, J. T. Kim, S. Kim, and J. Choo, "RobustNet: Improving domain generalization in urban-scene segmentation via instance selective whitening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11580–11590.

[54] Y. Zhao, Z. Zhong, N. Zhao, N. Sebe, and G. H. Lee, "Style-hallucinated dual consistency learning for domain generalized semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 535–552.

[55] Z. Wu, X. Wu, X. Zhang, L. Ju, and S. Wang, "SiamDoge: Domain generalizable semantic segmentation using Siamese network," in *Proc. 17th Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 603–620.

[56] S.-J. Chang, C.-Y. Lu, P.-K. Huang, and C.-T. Hsu, "Single-domain generalization for semantic segmentation via dual-level domain augmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 2335–2339.

[57] J. Ding, N. Xue, G.-S. Xia, B. Schiele, and D. Dai, "HGFormer: Hierarchical grouping transformer for domain generalized semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15413–15423.

[58] M. Liao, S. Tian, Y. Zhang, G. Hua, W. Zou, and X. Li, "Domain-invariant information aggregation for domain generalization semantic segmentation," *Neurocomputing*, vol. 546, Aug. 2023, Art. no. 126273.

[59] W.-J. Ahn, G.-Y. Yang, H.-D. Choi, and M.-T. Lim, "Style blind domain generalized semantic segmentation via covariance alignment and semantic consistence contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 3616–3626.

[60] Q. Bi, S. You, and T. Gevers, "Learning content-enhanced mask transformer for domain generalized urban-scene segmentation," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2023, pp. 819–827.

[61] S. Choi, D. Das, S. Choi, S. Yang, H. Park, and S. Yun, "Progressive random convolutions for single domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10312–10322.

[62] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "MixStyle neural networks for domain generalization and adaptation," *Int. J. Comput. Vis.*, vol. 132, no. 3, pp. 822–836, Mar. 2024.

[63] M. Noori et al., "TFS-ViT: Token-level feature stylization for domain generalization," *Pattern Recognit.*, vol. 149, May 2024, Art. no. 110213.

[64] *L1-Loss*. Accessed: Jan. 19, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Taxicab_geometry

[65] *L2-Loss*. Accessed: Jan. 19, 2025. [Online]. Available: https://en.wikipedia.org/wiki/Euclidean_distance

[66] J. Li, W. He, W. Cao, L. Zhang, and H. Zhang, "UANet: An uncertainty-aware network for building extraction from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5608513.

[67] W. Lu, X. Yang, and S.-B. Chen, "LWGANet: Addressing spatial and channel redundancy in remote sensing visual tasks with light-weight grouped attention," 2025, *arXiv:2501.10040*.