Frequency-based Optimal Style Mix for Domain Generalization in Semantic Segmentation of Remote Sensing Images

Reo Iizuka, Junshi Xia, *Senior Member, IEEE,* and Naoto Yokoya, *Member, IEEE,*

***Abstract*—Supervised learning methods assume that training and test data are sampled from the same distribution. However, this assumption is not always satisfied in practical situations of land cover semantic segmentation when models trained in a particular source domain are applied to other regions. This is because domain shifts caused by variations in location, time, and sensor alter the distribution of images in the target domain from that of the source domain, resulting in significant degradation of model performance. To mitigate this limitation, domain generalization has gained attention as a way of generalizing from source domain features to unseen target domains. One approach is style randomization, which enables models to learn domain-invariant features through randomizing styles of images in the source domain. Despite its potential, existing methods face several challenges, such as inflexible frequency decomposition, high computational and data preparation demands, slow speed of randomization, and lack of consistency in learning. To address these limitations, we propose a Frequency-based Optimal Style Mix (FOSMix), which consists of three components: 1) Full Mix enhances the data space by maximally mixing the style of reference images into the source domain, 2) Optimal Mix keeps the essential frequencies for segmentation and randomizes others to promote generalization, and 3) regularization of consistency ensures that the model can stably learn different images with the same semantics. Extensive experiments that require the model's generalization ability, with domain shift caused by variations in regions and resolutions, demonstrate that the proposed method achieves superior segmentation in remote sensing.**[1]

***Index Terms*—Domain Generalization, Style Randomization, Semantic Segmentation.**

I. Introduction

SEMANTIC segmentation of remote sensing images has undergone significant expansion in recent years due to its diverse range of applications, including monitoring environment, automatic mapping, and detecting land abuse. In machine learning, the effectiveness of semantic segmentation models is typically evaluated by training them on large annotated datasets of images and testing their performance on separate test datasets. It is commonly assumed that the training and test datasets are sampled from the same distribution. However, in many real-world situations, distributions of the training and test datasets may diverge, a phenomenon known as domain shift [1], [2]. For instance, in the field of remote sensing, the appearance of images like hue, texture, and resolution in the training and test datasets may vary due to differences in factors such as location, time, and sensor. Domain shift can significantly degrade model performance when a model trained on the source domain (i.e., the training data) is evaluated on the target domain (i.e., the test data), as shown in Fig. 1. The simplest solution is to collect and annotate target domains' images, but such a strategy requires tremendous labor and cost.

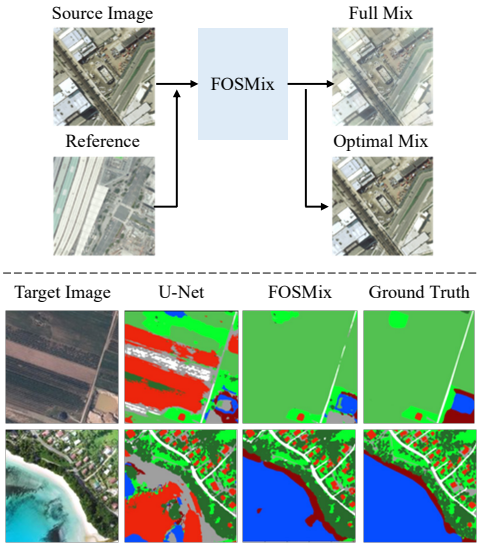


Fig. 1. Our proposed FOSMix enables the generation of diverse mixtures in frequency space by incorporating other image styles. The figure below shows the images of the target domain, the predictions of the baseline (U-Net), the predictions of FOSMix, and the ground truth. Our proposal demonstrates its robustness to the domain shift in the target domain compared to the baseline.

Unsupervised domain adaptation (UDA) is a technique for handling the domain shift by utilizing sufficient labeled data from the source domain and unlabeled data from the target domains to enhance the performance on target domains [3]–[10]. Domain adaptation seeks to train a model specialized for a specific domain, which is valuable especially when acquiring labeled data from the target domain is challenging. UDA methods are instrumental in this endeavor, transferring

This work was partly supported by JST, FOREST under Grant Number JPMJFR206S, and JSPS, KAKENHI under Grant Number 22H03609. (Corresponding author: Naoto Yokoya)

R. Iizuka is with the Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, the University of Tokyo, Chiba 277–8561, Japan (e-mail: reoiizuka97@gmail.com)

J. Xia is with the RIKEN Center for Advanced Intelligence Project, Tokyo 103–0027, Japan (e-mail: junshi.xia@riken.jp)

N. Yokoya is with the Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, the University of Tokyo, Chiba 277–8561, Japan, and also with the RIKEN Center for Advanced Intelligence Project, Tokyo 103–0027, Japan (e-mail: yokoya@k.u-tokyo.ac.jp)

[1]The source code is available at https://github.com/Reo-I/FOSMix

knowledge from a source domain to unlabeled target domains to align their distributions. However, it's important to note that even when UDA is targeted at multiple domains [11], it still requires retraining or fine-tuning for new target domains. Consequently, this means that a single model cannot be universally applied to diverse, unobserved domains. The recent emphasis on domain generalization (DG) (also known as out-out-distribution generalization) has resulted from its ability to generalize effectively to unobserved domains without access to the target domain [12]–[15]. The generalization to the target domain is attained by explicitly regularizing the model to enhance its resistance to domain shifts or supplementing the training data. This approach facilitates the development of a singular model that can be applied across various domains without retraining or fine-tuning. A popular technique in DG is style randomization [16]–[20], which increases the input space by introducing variability in image style. In particular, one approach to style randomization is Frequency Space Domain Randomization (FSDR) [21], which augments images in frequency space. FSDR randomizes domain-invariant frequency components (DIFs) with reference images that are neither in the source domain nor in the target domain while preserving domain-variant frequency components (DVFs), assuming that DIFs are essential features for the segmentation. However, applying FSDR to images with different sizes is challenging due to the way the frequency components are divided. Furthermore, Histogram Matching (HM) [22] for randomization is sluggish and time-consuming. In addition, this approach requires a large number of labeled reference images, which is different from the source domain, and a significant number of preliminary experiments, which is both time-consuming and resource intensive. Finally, the randomized images in different ways may induce the model to learn in a different direction, negatively impacting overall performance.

To address these limitations, we propose a novel DG mechanism called Frequency-based Optimal Style Mix (FOSMix) as illustrated in Fig. 1. FOSMix determines DVFs without preliminary experiments and mixes them more quickly, both in terms of time and with fewer iterations per image with elements of the reference image in frequency space. DVFs, which are to be mixed with reference images, are based on the masks created by the generator for each input image. Since we do not specifically know what the DVFs/DIFs are, the generator can acquire knowledge about the frequencies that do/don't impact segmentation through the incorporation of segmentation loss and mask loss. Subsequently, the generator produces a mask that represents the proportion of DVFs for each frequency component on a scale of 0 to 1, effectively serving as an indicator of the essential frequency components for segmentation. In addition, exact feature distribution mixing (EFDMix) [23] is used for mixing, enabling faster and more fine-grained mixture. Finally, the model simultaneously learns three images with different styles: a full mixed image that maximally incorporates the reference image's style, an optimal mixed image randomized based on the generated mask, and the original image. However, these images have identical semantics, and therefore, the model may encounter unstable learning due to divergent learning directions. Therefore, we introduce consistency regularization among predictions of their images to enable harmonious learning.

The contributions of this paper are summarized as follows:

- We propose a new module for the DG task, namely frequency-based optimal style mix (FOSMix), which addresses the deficiencies of current techniques in DG, such as inflexible division, high computational and data preparation demands, slow speed of randomization, and lack of consistency in learning.
- FOSMix consists of a full mix that maximally includes the style of the reference image, an optimal mix that leaves the frequency components necessary for segmentation and randomizes the rest, and a consistency loss that reconciles the training of the two different images randomized by the full and optimal mixes.
- Extensive two DG experiments in the field of remote sensing show that our proposal outperforms other methods.

II. Related Works

A. Domain Generalization

Domain generalization (DG) aims to learn a generalizable model from source domains for unseen target domains to address domain shifts between source and target domains. The DG methods can be divided into three categories [14], [15]: representation learning, learning strategy, and data manipulation.

1) Representation Learning: This category of methods is the most popular in DG. A domain-invariant representation is a feature representation consistent across different domains and not affected by domain-specific variations. By learning domain-invariant representations, the model can mitigate the domain gap by focusing on the task-specific features that are important for the classification or prediction task rather than being distracted by domain-specific variations. Typical methods include kernel-based methods [13], [24], adversarial learning [25]–[27], domain agnostic representation learning [28]–[30], and Maximum Mean Discrepancies (MMD) [31].

2) Learning Strategy: The second category focuses on the learning strategy to improve generalization performance. Since the parameters are updated by gradient descent to learn the model, it is straightforward to update the parameters in such a way as to acquire generality. In the gradient operation methods [32], [33], the gradients are adjusted by assuming that the gradient directions should be the same across the domains so that they do not conflict with each other. Ensemble learning, which enhances generalization performance by using multiple models, has the assumption that any samples are formed by the integration of multiple domains and use weighted sums of classifiers [34] and domain-specific batch normalization parameters [35]. Meta-learning methods [36], [37] are other learning strategies. In order to learn general representation, meta-learning methods for DG divide the source domain into meta-train and meta-test sets to simulate domain shifts.

3) Data Manipulation: Our work is mostly related to data manipulation. Data manipulation methods for the DG task aim to improve the performance of the model on unseen data by

increasing the data space through augmentation or generation [38], [39] to reduce the gap across domains [40]–[42]. As the volume of data is always in shortage for machine learning, data manipulation is the cheapest and easiest way to deal with such a situation. It has another advantage of being model-agnostic since it only increases the data pool.

For image augmentation, Adaptive Instance Normalization (AdaIN) [43], Histogram Matching (HM) [22], and Exact Histogram Matching (EHM) [23], [44] are used to randomize images by mixing or changing the style of the reference image with the style of the original image. The AdaIN transforms input source features $\boldsymbol{x} \in \mathbb{R}^m$ into output $\boldsymbol{o} \in \mathbb{R}^m$ whose mean and standard deviation match those of a target features $\boldsymbol{y} \in \mathbb{R}^m$ as follows:

$$\boldsymbol{o} = \frac{\boldsymbol{x} - \mu(\boldsymbol{x})}{\sigma(\boldsymbol{x})}\sigma(\boldsymbol{y}) + \mu(\boldsymbol{y}), \tag{1}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation of reference images, respectively. The HM transforms the input into the output so that the target empirical Cumulative Distribution Function (eCDF) matches the input eCDF. The EHM, on the other hand, adjusts the image pixels so that its histogram matches the one of a reference image more accurately. The sort-matching algorithm [45] can be used to match faster. Sort matching is accomplished by matching two sorted vectors $\{x_{\nu_n}\}_{n=1}^m$ and $\{y_{\kappa_n}\}_{n=1}^m$, where ν_n, κ_n are the indexes of the n-th smallest, $\boldsymbol{x}$ and $\boldsymbol{y}$ elements, respectively, i.e., only swapping them in order. To perform Exact Feature Distribution Mixing (EFDMix) [23] by applying EHM in the features space as follows:

$$\begin{aligned} &\text{EFDMix}(\boldsymbol{x}, \boldsymbol{y}) : \\ &o_{\nu_n} = x_{\nu_n} + (1-\lambda)y_{\kappa_n} - (1-\lambda)\langle x_{\nu_n} \rangle, \end{aligned} \tag{2}$$

where λ is the instance-wise mixing weight, and $\langle \cdot \rangle$ represents the stop-gradient operation [46], allowing for gradient back-propagation in deep learning.

After changing the style of the image in the feature space, it is then returned to the image space. The mapping function from images to features and vice versa can be implemented using encoder-decoder models. It is well-known in the field of research that many studies have been conducted on classification tasks [20], [47], [48], however, the use of generative models in semantic segmentation is not as common as it is in classification models. The reason behind this discrepancy is that while classification models can adapt and perform various modifications without affecting the meaning of the objects, semantic segmentation requires the preservation of the label meanings in the manipulation process. Therefore, in problem settings where the number of samples is not abundant, it is best to refrain from using generative models, which are uncertain whether the semantics will always hold, for stable data manipulation.

B. Style Randomization

Style randomization (SR) is a type of data manipulation performed for augmentation [16]. This approach expanded the input space by randomly changing images' color, texture, and light to address the domain shift from synthetic images to real-world data. DRPC [18] generate source images with the style of the reference image using CycleGAN [49] and learn the model to have features of multiple randomized images in the hidden layer possess similar. GLTR [19] also generates style-mixed images using AdaIN. However, most existing SR methods randomize the spectrum of images, which alters the domain-invariant features in undesirable ways. Although the spectrum is changed in feature space, operations in frequency space can separate DIFs from DVFs. Frequency-based operation methods [21], [50]–[53] have recently grained attention in the context of DA and DG tasks.

FSDR [21] utilizes the division in frequency space to perform domain generalization, in which the DVFs of the image are randomized with corresponding components of the reference images. However, FSDR divides the frequency coefficients fixedly and assigns them to either DVFs or DIFs, resulting in the inhibition of a diverse mixture. Additionally, HM slow down the randomization speed. Furthermore, preliminary experiments are required to determine the DVFs, which adds significant cost.

C. Datasets to Evaluate Domain Shift for Land Cover Mapping

In recent years, there has been a notable increase in the utilization of deep learning methods for analyzing remote sensing imagery. Semantic segmentation in remote sensing involves the process of assigning a class label to each pixel in an image, effectively creating a map of the scene. Many deep learning-based methods [54]–[56] have been proposed for semantic segmentation in remote sensing, including fully convolutional networks (FCNs) [57] and encoder-decoder architectures. These techniques have produced successful results on standard datasets like ISPRS 2D Semantic Labeling [58], SpaceNet [59], DeepGlobe [60], and OpenSentinelMap [61]. Unfortunately, these datasets do not share common labels, so it has not been easy to conduct DG experiments using multiple regional datasets.

Under conditions of the limited availability of comprehensive labeled datasets, recent research has focused on utilizing UDA techniques to adapt models to a specific region of interest once labeled data has been obtained for some regions. In the context of remote sensing, UDA techniques such as self-learning [62] and adversarial learning [63], [64] have been widely used to transfer knowledge from a labeled source domain to an unlabeled target domain. Specifically, utilizing AdaIN [43], MixStyle [20], and SPADE [41], some UDA methods [65]–[67] and the data synthesizing method [68] are proposed. However, from a practical standpoint in the real world, DG methods, which can be applied to any region, are urged rather than UDA methods.

Fortunately, OpenEarthMap (OEM) dataset [69], including high-resolution image datasets from 97 different regions, and French Land cover from Aerospace ImageRy (FLAIR) [70] with 40 regions in France, have recently become available for DG experiments. While BSM [71] is a DG method for building extraction, it depends on the large datasets of more than forty

thousand training images and changes all spectrum of images using AdaIN, which results in degradation of performance for natural class prediction. RobustBDD [72] proposed a method using stochastic weight averaging (SWA) [73] and adaptive batch normalization (AdaBN) [74] for building damage detection in domain shift, which is different problem setting from semantic segmentation. RobustBDD, as well as BCM, only considers buildings and assumes that features of each domain have a similar variance, which may not be satisfied in datasets with a variety of labels other than buildings. Besides, there is a metric for measuring the distance between domains [75], DG classification method for oil tree detection [76], and a representation learning method using knowledge graph for generalized zero-shot scene classification [77]. However, there has been a lack of research on DG of semantic segmentation in the field of remote sensing as multiple domains' datasets have not been developed until OEM was in place. That is where our research plays a vital role, as the proposed method has great potential to contribute to this field.

III. Method

In this section, we introduce our Frequency-based Optimal Style Mix (FOSMix) method as a novel style randomization technique. The FOSMix method consists of three main components: Full Mix, which maximally mixes the style of the reference image with the source domain image; Optimal Mix, which optimally mixes them for generalization; and Regularization of Consistency which helps the model update parameters stably. We first describe the problem definition and explain each component. The overall FOSMix framework is illustrated in Fig. 2.

A. Problem Definition

Given a training set of multiple source domains $\mathcal{D}_s = \{\mathcal{D}_1, \cdots, \mathcal{D}_S\}$ with N_k labeled samples $\{(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k)\}_{i=1}^{N_k}$ in the k-th domain $\mathcal{D}_k$ where $\boldsymbol{x}_i^k \in \mathbb{R}^{3\times H\times W}$ and $y_i^k \in \{0,1\}^{C\times H\times W}$ denote the input images and the corresponding C-class label maps, respectively. The goal is to learn a semantic segmentation model f_θ from source domains $\mathcal{D}_s$ that generalizes well on unseen target domains $\mathcal{D}_t$. In the context of domain generalization (DG), during the model training process, we have access to the training dataset from the source domains $\mathcal{D}_s$, whereas the training data from the target domains $\mathcal{D}_t$ remain inaccessible. Here, θ donates the model parameters to be learned. Subscript i and k are removed for simplification reasons in $\boldsymbol{x}_i^k$ and $\boldsymbol{y}_i^k$.

B. Full Mix in Frequency Space

Full Mix aims to maximally randomize the source image's frequency components (FCs) with those of the reference image $\boldsymbol{r} \in \mathbb{R}^{3\times H\times W}$. This means that all FCs are randomized. As part of the randomization process, the source image is mapped into frequency space using Discrete Cosine Transform (DCT) [78], [79] as in FSDR. Let $\tilde{\boldsymbol{x}}(= DCT(\boldsymbol{x})) \in \mathbb{R}^{3\times H\times W}$ and $\tilde{\boldsymbol{r}}(= DCT(\boldsymbol{r})) \in \mathbb{R}^{3\times H\times W}$ be features representing coefficients of the source domain image and the reference image in frequency space, respectively. The degree of mixing is determined for each element of $\tilde{\boldsymbol{x}}$. As an indicator to determine the degree of mixing, we introduce a mask $\boldsymbol{M} \in [0,1]^{3\times W\times H}$ corresponding to the elements of $\tilde{\boldsymbol{x}}$. When the element of the mask is 0 (i.e., DVFs), it means that the element of the source domain mixes with that of the corresponding reference, while if it is 1 (i.e., DIFs), it is holding. The flexibility to determine the mixing ratio for each frequency enables complex randomization. In Full Mix, to maximally randomize all FCs, the full mask $\boldsymbol{M}^f$ can be expressed as follows:

$$\begin{aligned} & \boldsymbol{M}^f = \boldsymbol{O} \quad \text{(zero matrix)} \\ \iff \quad & M^f_{i,j} = 0, \quad \forall i = 1, \cdots, H, \forall j = 1, \cdots, W \end{aligned} \tag{3}$$

Next, we consider the randomization process. In FSDR, HM is used for randomization. According to the randomization experiment [23], EFDMix randomization speed has been demonstrated to be 85 times faster compared to that of HM. Therefore, we employ EFDMix to make efficient use of GPUs and improve the speed of randomization. The randomized output $\tilde{\boldsymbol{o}}$ with its ν_n-th element $\tilde{o}_{\nu_n}$ in frequency space by EFDMix with mask $\boldsymbol{M}$ can be written as follows:

$$\begin{aligned} \tilde{o}_{\nu_n} &= \text{EFDMix}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{r}}, \boldsymbol{M})_{\nu_n} \\ &= \tilde{x}_{\nu_n} + (1 - M_{\nu_n}) \cdot \tilde{r}_{\kappa_n} - (1 - M_{\nu_n}) \cdot \langle \tilde{x}_{\nu_n} \rangle, \end{aligned} \tag{4}$$

where $\langle \cdot \rangle$ represents the stop-gradient operation to enable the back-propagation in the deep learning model [46]. In Full Mix, since all elements of the mask $\boldsymbol{M}^f$ are zero, we can practically modify Eq. 4 for all elements as:

$$\begin{aligned} \tilde{o}_{\nu_n} &= \text{EFDMix}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{r}}, \boldsymbol{M}^f)_{\nu_n} \\ &= \tilde{x}_{\nu_n} + \tilde{r}_{\kappa_n} - \langle \tilde{x}_{\nu_n} \rangle. \end{aligned} \tag{5}$$

After Full Mix using EFDMix, the mixed one is transformed from frequency space to image space by the inverse Discrete Cosine Transform (iDCT: $DCT^{-1}(\cdot)$). Formally, the full mixed image $\tau(\boldsymbol{x}, \boldsymbol{M}^f)$ can be obtained by:

$$\tau(\boldsymbol{x}, \boldsymbol{M}^f) = DCT^{-1}(\text{EFDMix}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{r}}, \boldsymbol{M}^f)). \tag{6}$$

Fig. 3 illustrates the process of Full Mix by simply replacing mask $\boldsymbol{M}$ with $\boldsymbol{M}^f$ without the generator. Full Mix allows the source domain image space to be expanded, enabling fast randomization without a band-pass filter.

C. Optimal Mix in Frequency Space

We propose Optimal Mix to randomize DVFs to provide diversity while keeping essential features for segmentation. More in detail, the element of the mask $\boldsymbol{M}$ corresponding to the DIFs, which are necessary for segmentation, is close to 1, and the one corresponding to the DVFs is close to 0 so that the mixing ratio of the reference image increases. Note that $\boldsymbol{M}$ takes continuous values rather than deterministic values such as 0 or 1, which allows for a wide variety of mixing.

Fig. 3 illustrates the detailed process of Optimal Mix. Assuming that the DVFs and DIFs are fluid for each source domain's image, the generator $G_{\theta'}$ creates the mask $\boldsymbol{M}$ where θ' represents the generator's parameters. Using the sigmoid function $\phi(\cdot)$ as the final layer's activation function, the

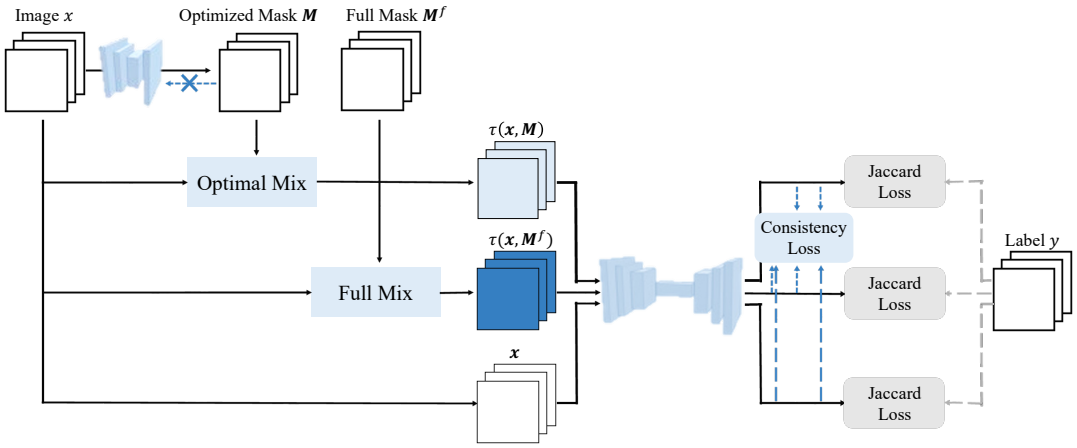


Fig. 2. The overall architecture of FOSMix. $\tau(\boldsymbol{x}, \boldsymbol{M})$ and $\tau(\boldsymbol{x}, \boldsymbol{M}^f)$ are generated by randomizing the source image $\boldsymbol{x}$ using two masks $\boldsymbol{M}$ and $\boldsymbol{M}^f$. Besides the Jaccard loss on each prediction, we also impose a consistency loss so that they harmonize their learning. Note that the mask generator does not update based on the consistency loss.

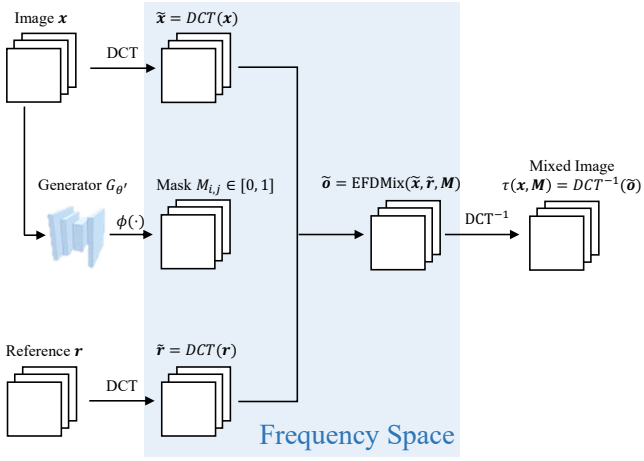


Fig. 3. The detailed architecture of the optimal mix. The frequency coefficients $\tilde{\boldsymbol{x}}$ transformed by DCT are mixed with the ones of reference image $\tilde{\boldsymbol{r}}$ with the mask $\boldsymbol{M}$ generated by the generator $G_{\theta'}$. The light blue background represents the operation in frequency space. As an activation function to make the mask $\boldsymbol{M}$ set from 0 to 1, $\phi(\cdot)$ represents the sigmoid function.

mask's elements are forced between 0 and 1. After the mask generation, the randomization uses EFDMix in Eq. 4. At the end of Optimal Mix, as in Full Mix, iDCT returns mixed features to the image space as follows:

$$\tau(\boldsymbol{x}, \boldsymbol{M}) = DCT^{-1}(\text{EFDMix}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{r}}, \boldsymbol{M})), \tag{7}$$

where

$$\boldsymbol{M} = \phi(G_{\theta'}(\boldsymbol{x})). \tag{8}$$

The generator needs a learning rule to determine which FCs are essential and which can be randomized. First, we define the segmentation loss for the three images: raw images, full mixed images, and optimal mixed images, as follows:

$$\begin{aligned}\mathcal{L}_{seg} = \; & \ell(f_\theta(\boldsymbol{x}), \boldsymbol{y}) + \ell(f_\theta(\tau(\boldsymbol{x}, \boldsymbol{M}^f)), \boldsymbol{y}) \\ & + \ell(f_\theta(\tau(\boldsymbol{x}, \phi(G_{\theta'}(\boldsymbol{x})))), \boldsymbol{y}).\end{aligned} \tag{9}$$

where, $\ell(\cdot)$ is the loss function. In Eq. 9, the third term uses the mixed image combined with the source and reference images. By setting all elements of the mask used in the mixed image to 1, the mixed image reverts to the raw image that has been presented in the first term of Eq. 9. Here, it is more convenient for the segmentation model f_θ to have less complex mixed images, and it can easily predict labels. As a result, the generator attempts to provide the model with the simplest raw image with the style of the source domain. In other words, θ' must be updated so that all mask elements are close to 1 via back-probagation. Then, in order to keep only the essential FCs for segmentation and randomize the others, we impose the mask loss as follows:

$$\mathcal{L}_{msk} = \lambda_1 \mu(\boldsymbol{M}) + \lambda_2 \frac{1}{\sigma^2(\boldsymbol{M})}, \tag{10}$$

where $\mu(\cdot)$ and $\sigma^2(\cdot)$ indicate the mean and variance of the mask, respectively. Particularly, λ_1 and λ_2 are the hyper-parameters balancing the two loss terms. The first term in Eq. 10 updates θ' so that the average value of the mask $\boldsymbol{M}$ is closer to 0. In other words, the mask loss promotes the generator in the direction of increasing the mixing ratio of the reference image. The mask loss function incentivizes the generator to move in a direction opposite to the learning rule outlined in Eq. 9. As a result, the element of the mask is expected to approach 0, while the mask element that is crucial for segmentation approaches 1. Additionally, in such tug-of-war situations, it is easier to differentiate between FCs that are neither fully DIFs nor DVFs by imposing losses that decrease the inverse variance of the mask. The name "Optimal Mix" is derived from the fact that the components of the mask are optimized based on the third term of Eq. 9 and the loss function of Eq. 10. Optimal Mix enables the determination and randomization of DVFs without requiring extensive preliminary experiments.

D. Regularization of Consistency

Our proposal includes a regularization of consistency [19] to ensure the coherence of the mixed images used for training,

which is illustrated in Fig. 2. Specifically, we utilize both full and optimal mixed images in the training process. However, the learning direction may vary due to the simultaneous training of images with different FCs, leading to potentially unstable learning as the input for these images carries the same semantics. By imposing the consistency loss, we aim to stabilize the randomization mechanisms and promote consistent learning. The formulation of consistency loss to all three pairs of outputs can be written as follows:

$$\begin{aligned}\mathcal{L}_{con} =&\lambda_3 \{\mathcal{L}_1\big(f_\theta(\tau(\boldsymbol{x},\boldsymbol{M})), f_\theta(\tau(\boldsymbol{x},\boldsymbol{M}^f))\big)\\ &+\mathcal{L}_1\big(f_\theta(\boldsymbol{x}), f_\theta(\tau(\boldsymbol{x},\boldsymbol{M}^f))\big)\\ &+\mathcal{L}_1\big(f_\theta(\tau(\boldsymbol{x},\boldsymbol{M})), f_\theta(\boldsymbol{x})\big)\},\end{aligned} \tag{11}$$

where $\mathcal{L}_1(\cdot,\cdot)$ represents the $L1$ distance between two different predictions, and λ_3 is the hyper-parameters balancing other loss terms. Note that Eq. 11 is calculated using $f_\theta(\tau(\boldsymbol{x},\boldsymbol{M})) = f_\theta(\tau(\boldsymbol{x},\phi(G_{\theta'}(\boldsymbol{x}))))$. This means that parameters θ' of the generator are updated so that the elements of the mask $\boldsymbol{M}$ approaches 0 or 1. Therefore, we stop updating θ' only for the consistency loss to prevent those parameters from being updated with back-propagation.

Our final objective can be defined as:

$$\min_{\theta,\theta'} \frac{1}{|\mathcal{X}_s|}\sum_{\boldsymbol{x}\in\mathcal{X}_s}(\mathcal{L}_{seg}+\mathcal{L}_{msk}+\mathcal{L}_{con}). \tag{12}$$

where $\mathcal{X}_s$ represents source domain's image space. During inference, we do not generate masks or perform randomization, but only use segmentation networks to make predictions.

IV. Experiments

In our experiments, our FOSMix is evaluated with other methods under the domain shift caused by regional differences. In addition, an ablation study will assess the contribution of each component.

A. Datasets

We evaluate our method in the domain shift caused by different geographic regions with respect to two datasets. In this situation, we use the OpenEarthMap (OEM) [69] dataset and the FLAIR [70]. OEM dataset contains 5,000 high-resolution remote sensing images from 44 countries and 97 geographic regions at a 0.25-0.5m ground sampling distance. It consists of 8 classes of $1,024\times1,024$ resolution in most regions. FLAIR dataset contains $61,712$ images from 40 regions in France, consisting of 18 classes of 512×512 resolution. Note that the background is represented by 13 classes through 18 classes, which collectively account for 0.58% of the total pixel.

For the first DG experiments for the regional shift, we use OEM dataset of 5,000 images. The source domain is represented by 1,531 images from 29 regions, which are divided into a training set of 1,222 images and a validation set of 309 images. Each region is represented by approximately four-fifths in the training set and one-fifth in the validation set. The test set, which serves as the target domain, contained 2,492 images from 47 regions. In addition, we have 977 reference images from 21 regions. Table V in the Appendix represents the classified domain for each region. This split way follows the practice of other DG experiments, such as synthetic-to-real DG experiments [18], [21], [80] and art/cartoon/sketch style images to photo style images DG experiments [20], [81], in which training and validation data are extracted from the same domain. Fig. 4 shows a t-SNE [82] visualization of the average features from the baseline hidden layer for each domain, color-coded by source and target domains. The target domain, enclosed by the green dotted line, is located relatively far from the source domain of the training data. It becomes possible to verify generality by looking at the results of predictions for regions that do not resemble these source domains in features. The results of the visualization will be discussed in the IV-D.

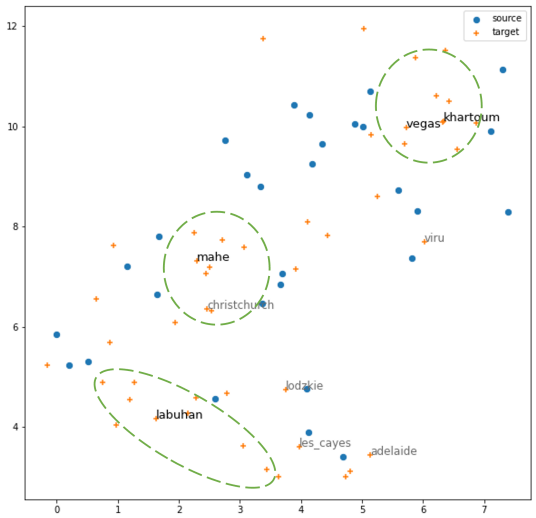


Fig. 4. The visualization of mean features per domain from the hidden layer of the baseline using t-SNE [82]. Many regions in the target domain, enclosed by the green dotted line, are located relatively far from the source domain of the training data. The regions of the target domain with names in black are particularly far from the source domain represented by the blue dots. This implies that they have new styles that the source domains don't possess.

For the FLAIR dataset, to avoid arbitrary assignment of regions, the dataset is divided according to the number assigned to each domain's file name. The file names for each classified domain are provided in Appendix Table VI. The training domain, validation domain, target domain, and reference region are represented by 14,522 images from 9 regions, 14,000 images from 9 regions, 17,375 images from 11 regions, and 15,815 images from 11 regions, respectively.

B. Implementation Details

For fair comparison, we use ResNet-50 [83] and VGG-16 [84] with U-Net [85] as the semantic segmentation network f_θ. The mask generators $G_{\theta'}$ do not need to be large models and utilize light backbones, i.e., ResNet-18 and VGG-11, which are both of the same types as the segmentation network. These models are initialized with pre-trained parameters on ImageNet [86]. The models are trained and evaluated on the

Python 3.6 platform. Specifically, all proposed models are implemented on two NVIDIA A100 GPUs for OEM and four NVIDIA A100 GPUs for FLAIR.

In the training period, we choose Adam [87] optimizer with beta coefficients of 0.9 and 0.999. The initial learning rate is set to 1e-4 and follows the poly learning policy [88] with a poly power of 0.9. The network is trained for a total of 150 epochs for OEM and 50 epochs for FLAIR. As an evaluation metric, we use Intersection over Union (IoU) for each class. The mean IoU (mIoU) is the average of the IoUs for all classes.

For the training set, after simple augmentations like flip, rotate, and Gaussian blur with 10%, we crop both source images and reference images with a size of 512×512 pixels. Besides, it is crucial to expand the data space by selecting images $\boldsymbol{r}$ for blending from both the source and reference domains with a probability of 50%.

During the test phase, the segmentation network f_θ is utilized for predicting labels without the use of the generator $G_{\theta'}$. To enhance the robustness of the prediction, we employed a test-time augmentation (TTA) [89], [90] technique where the predicted label was determined as the average maximum class of predictions with flipping operations.

To avoid focusing too much on source domain-specific tints, color jittering is applied to source domain images with a probability of 50%. The hyper-parameters of the first experiment in Eq. 10 and Eq. 11 that balance the loss term are set to 10 for λ_1, 0.2 for λ_2 and 5e-5 for λ_3, respectively. For the FLAIR dataset, we replace λ_2 with 0.1.

C. Comparison with Other Methods

In the OEM experiment, we evaluate the performance of FOSMix against the baseline, FSDR, RobustNet [91] and SiamDoGe [92], as shown in Table I. Two network backbones, ResNet-50 and VGG-16, are used in the experimentation. It should be noticed that RobustNet [91] and SiamDoGe [92] are based on the Resnet-50 backbones. The results in Table I demonstrate that FOSMix outperformed the compared methods on both network backbones. In particular, FOSMix with ResNet-50 exceeds the baseline in all classes. The highest accuracy is achieved when the backbones are ResNet-50 and VGG-16 in 7 and 6 out of 8 classes, respectively. Notably, the most substantial improvement in performance is observed in the road and agricultural land classes. The complexity of the problem is enhanced by the diversity of road conditions and agricultural land characteristics in both urban and rural areas. The road conditions range from well-paved concrete roads to dirt roads, while agricultural land exhibits variations in color and appearance depending on the region and the crops themselves. The improvement in performance can be attributed to the inclusion of a diverse set of data comprised of different mixed styles of roads and agricultural land. This diversity reinforced data and, thus, improved the model's ability to handle variations in road conditions and agricultural land characteristics that could not be achieved through simple techniques such as color jittering.

Our experimentation revealed a significant issue with the FSDR, as evidenced by the duration of the experiment. The requirement to transfer data from GPUs to CPUs before randomization for the HM algorithm resulted in a huge time overhead. Additionally, we found that utilizing the EFDMix [23] in place of the HM in FSDR took approximately 37 hours. Therefore, the inefficiency of HM has a major impact on how long it takes. In contrast, the time of our proposed method has also increased by a factor of 3 to 4 due to the tripling of the image input, but it can be seen that the experiment is completed in a realistic amount of time. Aslo, the training time of the proposed method is lower than the ones of RobustNet [91] and SiamDoGe [92].

In addition, we also conducted experiments with a large non-remote-sensing image dataset, ImageNet, instead of utilizing remote sensing images from other domains for reference. Note that we eliminated gray-scale images from ImageNet. The mIoU was found to be 43.08, which is approximately 0.9 points lower than when remote sensing images were used as the reference. The mIoU was also found to be 42.81, 1.1 percentage points lower when reference images were randomly selected from ImageNet with a 50% probability. As the performance with non-remote-sensing datasets is greatly improved compared to the source-only case, it is worth considering datasets from other fields, such as ImageNet, as an alternative when remote sensing imagery is unavailable.

Next, Table II displays the results of the FLAIR dataset experimentation, revealing noteworthy progress in mIoU, surpassing the baseline, RobustNet [91] and SiamDoGe [92] in 7 out of the 12 categories. It should be noticed that due to the limited computational resources, we do not show the results of FSDR. Notably, the categories of Bare Soil, Deciduous, Vineyard, Brushwood, and Agricultural Land demonstrated significant improvements. The advancements were attributed to the proposal's capacity to tackle the domain shifts that could vary drastically in look across regions and seasons. When considered globally such as OEM dataset, the FLAIR dataset limited to France alone exhibits no significant domain shift. As a consequence, improving performance across all categories can be challenging. Nonetheless, making progress in numerous categories still represents a noteworthy contribution.

D. Ablation Studies

We conduct an ablation study to examine the role of the FOSMix components in enhancing the generality of the network. Tab. III shows the performance of the baseline model and the impact of incorporating FOSMix components such as Full Mix, Optimal Mix, single consistency loss (Single CL), and all pairs consistency loss (All CL) in the regional difference problem setting.

It is observed that the baseline model trained on the original images alone perform poorly due to domain bias. The incorporation of the full or optimal mix yields the most favorable results in the developed space class, which can be attributed to the lack of diversity in the source domain and can be mitigated by incorporating new styles through mixing. Conversely, in the water category, adding Full Mix leads to a reduction in accuracy, while Optimal Mix improves the performance. This can be attributed to the fact that a

TABLE I
DG PERFORMANCE FOR THE OEM'S REGIONAL SHIFT IN MEAN IOU (MIOU) AND EACH CLASS IOU. TIME REPRESENTS HOURS TAKEN FOR THE TRAINING. THE BEST RESULTS FOR EACH CATEGORY ARE BOLD.

Network	Methods	mIoU	Bare Land	Grass	Developed Space	Road	Tree	Water	Agriculture Land	Building	Time (H)
Resnet-50	Source Only	40.65	14.1	38.9	36.9	48.9	52.9	31.6	35.8	66.1	2.6
	RobustNet [91]	41.86	16.1	37.5	35.4	49.6	54.1	33.4	41.6	67.2	12.3
	SiamDoGe [92]	42.17	17.2	38.1	36.2	49.1	55.2	32.7	42.1	66.8	13.4
	FSDR [21]	39.47	**20.0**	28.6	33.5	47.9	45.8	28.3	40.9	62.8	203.5
	Ours	**43.96**	16.4	**41.2**	**37.9**	**52.2**	**55.6**	**35.3**	**45.8**	**67.3**	7.4
VGG-16	Source Only	40.70	11.2	40.1	37.2	46.1	**55.6**	31.7	36.7	**67.0**	2.3
	FSDR [21]	39.38	11.3	36.7	35.9	48.5	53.2	27.7	37.0	64.8	236.1
	Ours	**43.36**	**13.7**	**40.6**	**38.2**	**49.5**	55.4	**32.0**	**48.7**	66.8	9.9

TABLE II
DG PERFORMANCE FOR THE FLAIR'S REGIONAL SHIFT IN MEAN IOU (MIOU) AND EACH CLASS IOU. TIME REPRESENTS HOURS TAKEN FOR THE TRAINING. THE BEST RESULTS FOR EACH CATEGORY ARE BOLD.

Network	Methods	mIoU	Building	Pervious Surface	Impervious Surface	Bare Soil	Water	Coniferous	Deciduous	Brushwood	Vineyard	Herbaceous Vegetation	Agricultural Land	Plowed Land
Resnet-50	Source Only	44.92	72.7	32.7	64.6	17.6	58.6	**15.1**	52.2	17.9	71.6	43.4	38.5	**54.3**
	RobustNet [91]	46.54	73.2	33.4	65.1	26.8	61.3	15.6	54.7	18.7	74.2	44.8	**46.6**	44.1
	SiamDoGe [92]	46.58	73.5	34.2	64.6	28.1	**62.1**	13.4	54.2	19.2	74.8	**45.6**	44.1	45.2
	Ours	**48.51**	**74.0**	**35.4**	**66.2**	**35.2**	60.9	12.7	**57.4**	**19.6**	**75.7**	45.2	46	53.6

TABLE III
ABLATION STUDIES FOR OEM'S DG TASK CAUSED BY THE REGIONAL SHIFT IN MIOU AND EACH CLASS IOU. FULL MIX AND OPTIMAL MIX ARE FULL MIXED IMAGES $\tau(\boldsymbol{x}, \boldsymbol{M}^f)$ AND OPTIMAL MIXED IMAGES $\tau(\boldsymbol{x}, \boldsymbol{M})$, RESPECTIVELY. SINGLE CL AND ALL CL REPRESENT THE CONSISTENCY LOSS BETWEEN MIXED IMAGES AND THE CONSISTENCY LOSS FOR ALL PAIRS, RESPECTIVELY. THE NETWORK BACKBONE IS THE RESNET-50.

Methods	Full Mix	Optimal Mix	Single CL	All CL	IoU								
					mIoU	Bare Land	Grass	Developed Space	Road	Tree	Water	Agriculture Land	Building
Baseline					40.65	14.1	38.9	36.9	48.9	52.9	31.6	35.8	66.1
FOSMix	✓				42.85	14.7	40.6	**38.4**	49.8	54.6	31.1	46.7	66.8
		✓			42.75	13.9	**41.6**	**38.4**	51.5	53.9	33.7	42.3	66.7
	✓	✓			42.13	14.3	36.3	37.5	50.3	51.3	33.9	**46.9**	66.5
	✓	✓	✓		43.50	15.5	41.5	38.3	**52.4**	54.3	32.2	46.4	**67.4**
	✓	✓		✓	**43.96**	**16.4**	41.2	37.9	52.2	**55.6**	**35.3**	45.8	67.3

consistent frequency is prevalent across domains in the water category and the use of bold style transformations disrupts the original frequency, whereas Optimal Mix increases diversity by selectively altering appropriate frequencies. Additionally, it is evident that the agricultural land category necessitates a diverse range of styles, with vastly distinct distributions in urban and undeveloped regions, and Full Mix plays a significant role in this category. Second, when both Full Mix and Optimal Mix are added simultaneously, the mIoU is lower than when they are used individually. This is due to the fact that different images with the same semantics inhibit learning and introduce conflicts in the learning direction. To address this issue, consistency loss is introduced to alleviate the conflict. The implementation of a single consistency loss between Full mix and Optimal mix is more accurate and stable than any of the previous ones and outperformed the baseline in all categories. Furthermore, FOSMix with all elements results in the highest performance, especially in categories that are sensitive to excessive style mixing such as bare land, trees, and water. Additionally, the FOSMix with all elements added also performed well in both natural and artificial classes such as grass, roads, and buildings, indicating a moderate and diverse style mixing.

From a qualitative perspective, Fig. 5 shows the models' predictions for the baseline, Full Mix, Optimal Mix, and consistency loss in combination. The target domain includes a diverse range of style images, which are divided into two groups for clarity: rural areas and urban areas. In Fig. 4, the names of the target domain regions, represented in black, are relatively far from the source domain (plotted in blue), indicating their features do not present in the source domain. To achieve a more comprehensive understanding, apart from the urban and rural categories, we also compare the predicted labels for the regions of Vegas, Labuhan, Khartoum, and Mahe, which seem to possess a unique style.

In rural areas, due to the limited diversity in the source domain, the baseline model misclassifies crops as buildings, mistakes them for similar-looking grasses, and fails to rec-

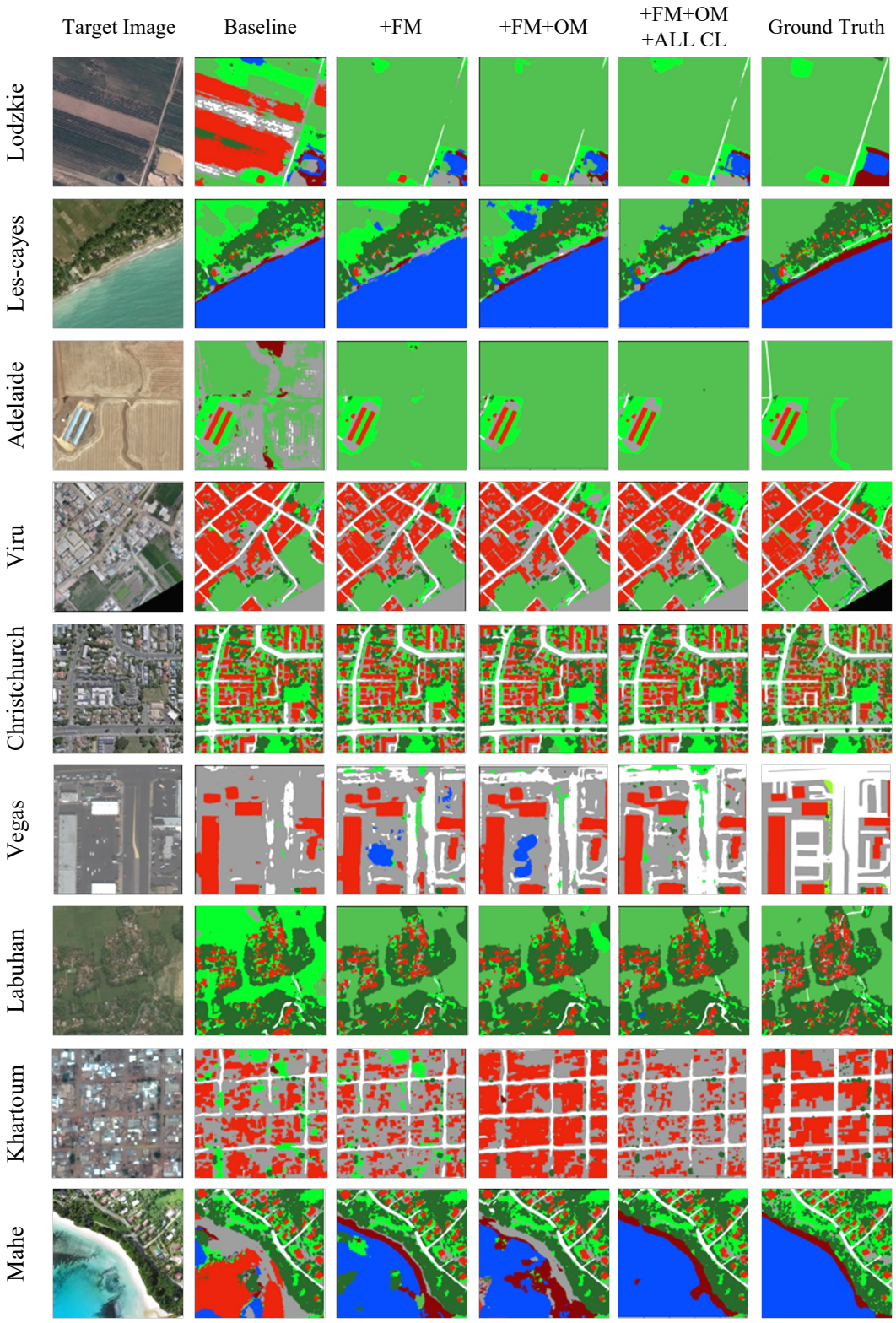


Fig. 5. Qualitative results of DG task caused by the regional shift. FM stands for Full Mix and OM for Optimal Mix.

ognize crops in different seasons accurately. Our proposal shows that even a simple addition of the full mix improves the performance significantly, and incorporating all elements improves the accuracy. Specifically, in Les-cayes, the introduction of consistency loss for all pairs yields more plausible predictions. Conversely, in urban areas, despite the complexity of the ground truth structure in urban areas, the source domain contains enough urban information to provide accurate predictions at baseline. Therefore, the proposed method has limited effect in these areas. Finally, in the regions with the new styles, such as dimly tinted imagery in Vegas and Labuhan, and an abundance of tints not present in the source domain in Mahe, the baseline model based on the source domain's class-specific tints performs poorly. However, incorporating each element of FOSMix shows improvement. Additionally, in Khartoum, while the proposed method still falls short of ground truth, it no longer incorrectly classifies the images as grass.

For FLAIR dataset, the result of ablation studies are shown in Tab. IV. Tab. IV indicates that the inclusion of all the proposed elements results in the best performance as well as OEM's results. Among 12 categories, four of the 12 classes show the highest values. Incorporating the Full Mix in addition to the baseline yields high values in three categories. Notably, the Full Mix model alone demonstrates superior accuracy in Coniferous, presumably due to the presence of features with widely varying distributions in the target domain. In addition, it is clear that the simultaneous inclusion of Full Mix and Optimal Mix reduces accuracy, while the inclusion of consistency loss improves it.

E. Visualization of Mixed Image

In Fig. 6, examples of full mixed and optimal mixed images are presented, which have been found to contribute to improved performance. The variations in brightness, tint, and other stylistic characteristics of the images are dependent on the specific domain. Nevertheless, the full mixed image demonstrates that the style of the image aligns with that of the reference image. It is not an exaggeration to say that the image looks as if it has been sampled from the domain of the reference image, considering only the style. The incorporation of such full mixed images allows for the model to learn a diverse range of features. In the case of optimal mixed images, high-frequency components are fused, while the general tint and other elements remain similar to those of the source domain. Specifically, in rows 5 and 6, the source images whose coefficients of high-frequency components are low are combined with the high-frequency components of the reference image. This feature, which represents texture, is randomized to increase diversity in the high-frequency component and improve generalization performance.

As illustrated in Fig. 7, the generator produces a three-dimensional mask during Optimal Mix process. This mask preserves middle-frequency information while randomizing high frequencies in the red and green channels for each image. Note that the low-frequency information is located in the upper left corner of the image, with middle-frequency information extending from the upper left to approximately a quarter of the width. In contrast, the blue channel of the mask varies for each image, indicating that the features to be retained are also located in the high frequencies. By selectively mixing different frequencies for each image, a variety of randomization can be achieved.

Fig. 6. Illustration of results on Full Mix and Optimal Mix. Full Mix has the same styles of reference images, while the semantics (i.e., objects) of original images are not changed. On the other hand, Optimal Mix has the middle and high frequencies of reference images.

V. Conclusion

This paper presents a frequency-based optimal style mix (FOSMix) for the domain generalization task. The proposed FOSMix method consists of three key components: Full Mix that maximizes the mixing of styles in the frequency space, Optimal Mix that selectively randomizes frequencies not required for segmentation, and the regularization of consistency that ensures harmonious training of the generated images. Through extensive experiments, we demonstrate that FOSMix effectively addresses the limitations of the existing method and achieves superior segmentation performance in land cover

TABLE IV
ABLATION STUDIES FOR FLAIR'S DG TASK CAUSED BY THE REGIONAL SHIFT IN MIOU AND EACH CLASS IOU. THE NETWORK BACKBONE IS THE RESNET-50.

Methods	Full Mix	Optimal Mix	Single CL	All CL	mIoU	IoU											
						Building	Pervious Surface	Impervious Surface	Bare Soil	Water	Coniferous	Deciduous	Brushwood	Vineyard	Herbaceous Vegetation	Agricultural Land	Plowed Land
Baseline					44.92	72.7	32.7	64.6	17.6	58.6	15.1	52.2	17.9	71.6	43.4	38.5	54.3
FOSMix	✓				47.08	**74.0**	35.4	65.6	23.9	58.3	**17.2**	55.8	19.5	73.2	44.0	**46.6**	51.5
		✓			47.85	73.6	**36.2**	65.7	32.1	58.9	12.1	**57.7**	18.1	75.5	**45.2**	44.3	54.7
	✓	✓			46.04	73.1	31.2	63.5	19.8	55.3	15.9	54.4	**20.2**	**79.5**	40.1	42.3	**57.4**
	✓	✓	✓		48.40	73.9	35.3	66.1	**36.4**	60.4	14.9	57.0	19.4	74.1	44.3	41.3	57.2
	✓	✓		✓	**48.51**	**74.0**	35.4	**66.2**	35.2	**60.9**	12.7	57.4	19.6	75.7	**45.2**	46.2	53.6

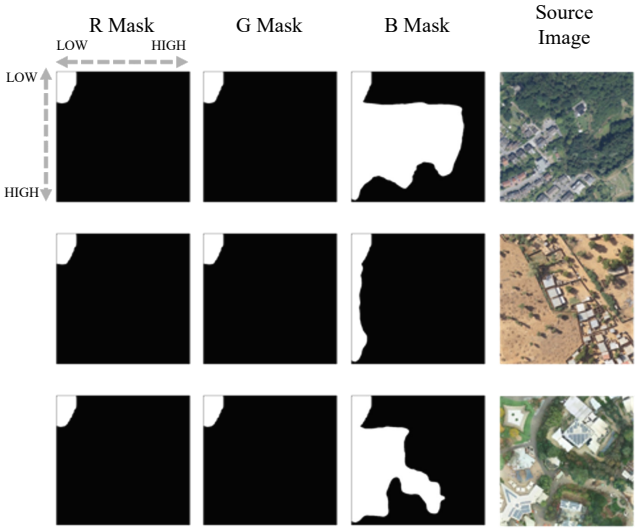


Fig. 7. Illustration of generated masks for each channel. R, G, and B masks are red, green, and blue channel masks, respectively. White represents one, whose frequency is kept, and black represents zero, whose frequency is mixed with the reference image. The upper left part of the mask represents the coefficients for low-frequencies, and the lower right part represents the coefficients for high-frequencies.

semantic segmentation. Additionally, the proposed approach is model-agnostic, allowing for combining other architectures and methods.

APPENDIX A
REGIONAL SELECTION FOR DG

The split of regions used for DG experiments for OEM is represented as Tab. V. The focus of the problem setting is on DG for cases in which there are domain shifts across different regions. In the field of remote sensing, obtaining high-resolution images and corresponding labels can be quite expensive. As a result, only countries and cities with sufficient funding are able to acquire such data, which may then be used to apply models to other regions. For this problem setting, we chose urban-centric regions as the source domain and various other regions as the target domain.

The split of regions used for DG experiments for FLAIR is represented as Tab. VI. Since the labels for the target domains described in the original paper were not available at the time of the experiment, we used only the 40 regions with available labels from the training and validation domains. For the validation domains, we used the same set of domains as in the original paper. We then assigned the target domain to the region with a file name divisible by 3, while the source domain consisted of the remaining files with a file name divisible by either 4 or 5. The remaining region was assigned to the reference region.

REFERENCES

[1] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
[2] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," IEEE signal processing magazine, vol. 32, no. 3, pp. 53–69, 2015.
[3] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in European Conference on Computer Vision, 2010.
[4] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in International conference on machine learning. PMLR, 2015, pp. 1180–1189.
[5] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in International conference on machine learning. Pmlr, 2018, pp. 1989–1998.
[6] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7167–7176.
[7] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), vol. 2, no. 5, Oct 2017, p. 6.
[8] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4893–4902.
[9] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[10] B. Li, Y. Wang, S. Zhang, D. Li, K. Keutzer, T. Darrell, and H. Zhao, "Learning invariant representations and risks for semi-supervised domain adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1104–1113.

TABLE V
REGIONS CHOSEN AS SOURCE DOMAINS, TARGET DOMAINS, AND REFERENCE IMAGES FOR THE FIRST EXPERIMENT. THIS IS AN EXPERIMENT IN DG THAT REPLICATES DOMAIN SHIFTING, WITH URBAN AREAS AS THE SOURCE DOMAIN AND VARIOUS REGIONS AS THE TARGET DOMAIN.

Source Regions (Train)	Target Regions (Test)		Reference Regions
Austin (USA)	Accra (Ghana)	Pomorskie (Poland)	Aachen (Germany)
Bogota (Colombia)	Adelaide (Australia)	Port-a-piment (Haiti)	Abancay (Peru)
Buenos aires (Argentina)	Al qurnah (Iraq)	Rio (Brazil)	Bielefeld (Germany)
Chicago (USA)	Baybay (Philippines)	Rosario (Argentina)	Chiangmai (Tailand)
Chiclayo (Peru)	Chincha (Peru)	San tome (Sao Tome and Principe)	Cox's bazar (Bangladesh)
Chisinau (Moldova)	Christchurch (New Zealand)	Santiago (Chile)	Dolnoslaskie (Poland)
Dhaka (Bangladesh)	Dar es salaam (Tanzania)	Sechura (Peru)	Gorakhpur (Nepal)
Dortmund (Germany)	Dowa (Malawi)	Shanghai (China)	Kampala (Uganda)
Dusseldorf (Germany)	El rodeo (Guatemala)	Slaskie (Poland)	Kujawsko-pomorskie (Poland)
Joplin (USA)	Houston (USA)	Svaneti (Georgia)	Little rock (USA)
Koeln (Germany)	Ica (Peru)	Swietokrzyskie (Poland)	Malopolskie (Poland)
Kyoto (Japan)	Jeremie (Haiti)	Thousand oaks (USA)	Mazowieckie (Poland)
Lambayeque (Peru)	Kagera (Tanzania)	Tonga (Tonga)	Palu (Indonesia)
Lima (Peru)	Khartoum (Sudan)	Tyrol (Austria)	Panama city (USA)
Lohur (Tajikistan)	kinshasa (Congo)	Vegas (USA)	Pisco (Peru)
Lubuskie (Poland)	Kitsap (USA)	Viru (Peru)	Saint-louis-du-sud (Haiti)
Melbourne (Australia)	Labuhan (Malaysia)	Warminsko-mazurskie (Poland)	Santa rosa (USA)
Mexico city (Mexico)	Leilane estates (USA)	Zanzibar (Tanzania)	Tokyo (Japan)
Muenster (Germany)	Les-cayes (Haiti)		Ulaanbaatar (Mongolia)
Niamey (Niger)	Lodzkie (Poland)		Wallace (USA)
Paris (France)	Mahe (Seychelles)		Zachodniopomorskie (Poland)
Pedrogao grande (Portugal)	Maputo (Mozambique)		
Rotterdam (Netherlands)	Monrovia (Liberia)		
Soriano (Uruguay)	Ngaoundere (Cameroon)		
Tulsa (USA)	Oklahoma (USA)		
Tuscaloosa (USA)	Piura (Peru)		
Vienna (Austria)	Podkarpackie (Poland)		
Western (Ghana)	Podlaskie(Poland)		
Wielkopolskie (Poland)	Pointenoire (Congo)		

TABLE VI
REGIONS CHOSEN AS SOURCE DOMAINS, TARGET DOMAINS, AND REFERENCE IMAGES FOR THE SECOND EXPERIMENT. THIS IS AN EXPERIMENT IN DG THAT REPLICATES DOMAIN SHIFTING BASED ON THE RANDOMLY SAMPLED REGIONS.

Source Regions (Train)	Source Regions (Validation)	Target Regions (Test)	Reference Regions
D008	D012	D006	D007
D016	D022	D009	D013
D032	D026	D021	D023
D035	D064	D030	D034
D044	D071	D033	D038
D052	D075	D051	D041
D055	D076	D060	D046
D070	D083	D063	D049
D080	D085	D072	D074
		D078	D086
		D081	D091

[11] S. Saha, S. Zhao, and X. X. Zhu, "Multitarget domain adaptation for remote sensing classification using graph neural network," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2022.

[12] G. Blanchard, G. Lee, and C. Scott, "Generalizing from several related classification tasks to a new unlabeled sample," Advances in neural information processing systems, vol. 24, 2011.

[13] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in International Conference on Machine Learning. PMLR, 2013, pp. 10–18.

[14] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," IEEE Transactions on Knowledge and Data Engineering, 2022.

[15] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[16] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2017, pp. 23–30.

[17] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," in 2018 IEEE international conference on robotics and automation (ICRA). IEEE, 2018, pp. 3803–3810.

[18] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2100–2110.

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

[20] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in International Conference on Learning Representations, 2021.

[21] J. Huang, D. Guan, A. Xiao, and S. Lu, "Fsdr: Frequency space domain randomization for domain generalization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pp. 6891–6902.

[22] G. Rafael C, W. Richard E et al., "Digital image processing," 2002.

[23] Y. Zhang, M. Li, R. Li, K. Jia, and L. Zhang, "Exact feature distribution matching for arbitrary style transfer and domain generalization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 8035–8045.

[24] M. Ghifary, D. Balduzzi, W. B. Kleijn, and M. Zhang, "Scatter component analysis: A unified framework for domain adaptation and domain generalization," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 7, pp. 1414–1430, 2016.

[25] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 5400–5409.

[26] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 624–639.

[27] R. Gong, W. Li, Y. Chen, and L. V. Gool, "Dlow: Domain flow for adaptation and generalization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2477–2486.

[28] H. Wang, Z. He, Z. L. Lipton, and E. P. Xing, "Learning robust representations by projecting superficial statistics out," in International Conference on Learning Representations, 2019.

[29] S. Lee, H. Seong, S. Lee, and E. Kim, "Wildnet: Learning domain generalized semantic segmentation from the wild," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 9936–9946.

[30] D. Peng, Y. Lei, M. Hayat, Y. Guo, and W. Li, "Semantic-aware domain generalized segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 2594–2605.

[31] J. Kang, S. Lee, N. Kim, and S. Kwak, "Style neophile: Constantly seeking novel styles for domain generalization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 7130–7140.

[32] L. Mansilla, R. Echeveste, D. H. Milone, and E. Ferrante, "Domain generalization via gradient surgery," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2021, pp. 6630–6638.

[33] Y. Shi, J. Seely, P. H. S. Torr, N. Siddharth, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization."

[34] M. Mancini, S. R. Bulo, B. Caputo, and E. Ricci, "Best sources forward: domain generalization through source-specific nets," in 2018 25th IEEE international conference on image processing (ICIP). IEEE, 2018, pp. 1353–1357.

[35] M. Segu, A. Tonioni, and F. Tombari, "Batch normalization embeddings for deep domain generalization," Pattern Recognition, vol. 135, p. 109115, 2023.

[36] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1, 2018.

[37] H. Bai, F. Zhou, L. Hong, N. Ye, S.-H. G. Chan, and Z. Li, "Nas-ood: Neural architecture search for out-of-distribution generalization," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 8320–8329.

[38] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[40] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2019, pp. 4401–4410.

[41] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2337–2346.

[42] J. Shi, N. Xu, H. Zheng, A. Smith, J. Luo, and C. Xu, "Spaceedit: Learning a unified editing space for open-domain image color editing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 19 730–19 739.

[43] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2017, pp. 1501–1510.

[44] D. Coltuc, P. Bolon, and J.-M. Chassery, "Exact histogram specification," IEEE Transactions on Image processing, vol. 15, no. 5, pp. 1143–1152, 2006.

[45] J. P. Rolland, V. Vo, B. Bloss, and C. K. Abbey, "Fast algorithms for histogram matching: Application to texture synthesis," Journal of Electronic Imaging, vol. 9, no. 1, pp. 39–45, 2000.

[46] X. Chen and K. He, "Exploring simple siamese representation learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15 750–15 758.

[47] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Learning to generate novel domains for domain generalization," in European conference on computer vision. Springer, 2020, pp. 561–578.

[48] Y. Wang, L. Qi, Y. Shi, and Y. Gao, "Feature-based style randomization for domain generalization," IEEE Transactions on Circuits and Systems for Video Technology, 2022.

[49] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2017, pp. 2223–2232.

[50] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4085–4095.

[51] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A fourier-based framework for domain generalization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pp. 14 383–14 392.

[52] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, "Detecting camouflaged object in frequency domain," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 4504–4513.

[53] X. Li, Y. Zhang, J. Yuan, H. Lu, and Y. Zhu, "Discrete cosin transformer: Image modeling from frequency domain," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 5468–5478.

[54] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016, pp. 1–9.

[55] Y. Su, J. Cheng, H. Bai, H. Liu, and C. He, "Semantic segmentation of very-high-resolution remote sensing images via deep multi-feature learning," Remote Sensing, vol. 14, no. 3, p. 533, 2022.

[56] P. Yin, D. Zhang, W. Han, J. Li, and J. Cheng, "High-resolution remote sensing image semantic segmentation via multiscale context and linear self-attention," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 15, pp. 9174–9185, 2022.

[57] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.

[58] F. Rottensteiner, G. Sohn, M. Gerke, and J. D. Wegner, "Isprs semantic labeling contest," ISPRS: Leopoldshöhe, Germany, vol. 1, p. 4, 2014.

[59] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," arXiv preprint arXiv:2110.08733, 2021.

[60] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 172–181.

[61] N. Johnson, W. Treible, and D. Crispell, "Opensentinelmap: A large-scale land use dataset using openstreetmap and sentinel-2 imagery," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1333–1341.

[62] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–13, 2021.

[63] Z. Li, X. Tang, W. Li, C. Wang, C. Liu, and J. He, "A two-stage deep domain adaptation method for hyperspectral image classification," Remote Sensing, vol. 12, no. 7, p. 1054, 2020.

[64] J. Zhu, Y. Guo, G. Sun, L. Yang, M. Deng, and J. Chen, "Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level context memory," arXiv preprint arXiv:2208.07722, 2022.

[65] O. Tasar, A. Giros, Y. Tarabalka, P. Alliez, and S. Clerc, "Daugnet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 2, pp. 1067–1081, 2020.
[66] O. Tasar, Y. Tarabalka, A. Giros, P. Alliez, and S. Clerc, "Standardgan: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 192–193.
[67] Z. Zhengxia, S. Tianyang, L. Wenyuan, Z. Zhou, and S. Zhenwei, "Do game data generalize well for remote sensing image segmentation?" Remote Sensing, vol. 12, no. 2, p. 275, 2020.
[68] G. Baier, A. Deschemps, M. Schmitt, and N. Yokoya, "Synthesizing optical and sar imagery from land cover maps and auxiliary raster data," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–12, 2021.
[69] J. Xia, N. Yokoya, B. Adriano, and C. Broni-Bediako, "Openearthmap: A benchmark dataset for global high-resolution land cover mapping," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), January 2023.
[70] A. Garioud, S. Peillet, E. Bookjans, S. Giordano, and B. Wattrelos, "Flair: French land cover from aerospace imagery." [Online]. Available: https://codalab.lisn.upsaclay.fr/competitions/8769#learn_the_details
[71] 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," arXiv preprint arXiv:2208.10004, 2022.
[72] V. Benson and A. S. Ecker, "Assessing out-of-domain generalization for robust building damage detection," NeurIPS 2020 Workshop on Artificial Intelligence for Humanitarian Assistance and Disaster Response (AI+HADR 2020), 2020.
[73] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," arXiv preprint arXiv:1803.05407, 2018.
[74] Y. Li, N. Wang, J. Shi, X. Hou, and J. Liu, "Adaptive batch normalization for practical domain adaptation," Pattern Recognition, vol. 80, pp. 109–117, 2018.
[75] C. M. Gevaert and M. Belgiu, "Assessing the generalization capability of deep learning networks for aerial image classification using landscape metrics," International Journal of Applied Earth Observation and Geoinformation, vol. 114, p. 103054, 2022.
[76] J. Zheng, W. Wu, S. Yuan, H. Fu, W. Li, and L. Yu, "Multisource-domain generalization-based oil palm tree detection using very-high-resolution (vhr) satellite images," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1–5, 2021.
[77] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, "Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 179, pp. 145–158, 2021.
[78] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," IEEE transactions on Computers, vol. 100, no. 1, pp. 90–93, 1974.
[79] G. K. Wallace, "The jpeg still picture compression standard," Communications of the ACM, vol. 34, no. 4, pp. 30–44, 1991.
[80] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 464–479.
[81] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 834–843.
[82] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." Journal of machine learning research, vol. 9, no. 11, 2008.
[83] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[84] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[85] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
[86] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, 2009, pp. 248–255.
[87] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[88] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.
[89] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.
[90] G. Wang, W. Li, S. Ourselin, and T. Vercauteren, "Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation," in Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. Springer, 2019, pp. 61–72.
[91] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2021, pp. 11 580–11 590.
[92] Z. Wu, X. Wu, X. Zhang, L. Ju, and S. Wang, "Siamdoge: Domain generalizable semantic segmentation using siamese network," in Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII, 2022, p. 603–620.

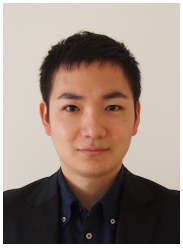
Reo Iizuka received the B.Eng. degree from the Department of Industrial and Management Systems Engineering, Waseda University, Tokyo, Japan in 2021, and the M.S. degree from the Department of Complexity Science and Engineering, The University of Tokyo, Chiba, Japan, in 2023.

Since 2023 he has been with Boston Consulting Group (BCG), Tokyo, where he is currently a Data Scientist. His research interests include domain generalization in remote sensing, machine learning algorithms for understanding remote sensing images, and applications to real-world problem settings.

Junshi Xia (Senior Member, IEEE) received the B.S. degree in geographic information systems and the Ph.D. degree in photogrammetry and remote sensing from the China University of Mining and Technology, Xuzhou, China, in 2008 and 2013, respectively, and the Ph.D. degree in image processing from the Grenoble Images Speech Signals and Automatics Laboratory, Grenoble Institute of Technology, Grenoble, France, in 2014.

From 2014 to 2015, he was a Visiting Scientist with the Department of Geographic Information Sciences, Nanjing University, Nanjing, China. From 2015 to 2016, he was a Postdoctoral Research Fellow with the University of Bordeaux, Bordeaux, France. From 2016 to 2018, he was the Japan Society for the Promotion of Science (JSPS) Postdoctoral Overseas Research Fellow with the University of Tokyo, Tokyo, Japan. Since 2018, he has been with the RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, where he is currently a Senior Research Scientist. His research interests include multiple classifier systems in remote sensing, hyperspectral remote sensing image processing, and deep learning in remote sensing applications.

Dr. Xia was the recipient of the first place prize in the IEEE Geoscience and Remote Sensing Society Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee in 2017. Since 2019, he has been an Associate Editor for the IEEE Geoscience and Remote Sensing Letters (GRSL), Remote Sensing and Frontiers in Remote Sensing, and Guest Editor for Remote Sensing and the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS).



Naoto Yokoya (S'10-M'13) received the M.Eng. and Ph.D. degrees from the Department of Aeronautics and Astronautics, The University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

He was an Assistant Professor at The University of Tokyo from 2013 to 2017. From 2015 to 2017, he was an Alexander von Humboldt Fellow with the German Aerospace Center (DLR), Oberpfaffenhofen, Germany, and the Technical University of Munich (TUM), Munich, Germany. Since 2018, he has been with the RIKEN Center for Advanced Intelligence Project, Tokyo, where he leads the Geoinformatics Team. Since 2020, he has been with The University of Tokyo, where he is currently an Associate Professor. His research is focused on the development of image processing, data fusion, and machine learning algorithms for understanding remote sensing images, with applications to disaster management and environmental assessment.

Dr. Yokoya won first place in the 2017 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee (IADF TC). He was the Chair from 2019 to 2021, a Co-Chair of the IEEE GRSS IADF TC from 2017 to 2019, and also the Secretary of the IEEE GRSS All Japan Joint Chapter from 2018 to 2021. He was an Associate Editor for the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS) from 2018 to 2021. He has been an Associate Editor for the IEEE Transactions on Geoscience and Remote Sensing since 2021.