Remote Sensing Image Synthesis via Semantic Embedding Generative Adversarial Networks

Chendan Wang, Bowen Chen, Zhengxia Zou and Zhenwei Shi*, *Member, IEEE*

***Abstract*—Generating photo-realistic remote sensing images conditioned on semantic masks has many practical applications like image editing, detecting deep fake geography, and data augmentation. Although previous methods achieved high-quality synthesis results for natural images like faces and everyday objects, they still underperform in remote sensing scenarios in terms of both visual fidelity and diversity. The high data imbalance and high semantic similarity of remote sensing object categories make the semantic synthesis of remote sensing images more challenging than natural images. To tackle these challenges, we propose a novel method named Conducted Semantic EmBedding GAN (CSEBGAN) for semantic-controllable remote sensing image synthesis. The proposed method decouples different semantic classes into independent Semantic Embeddings, which explores the regularities between classes to improve visual fidelity and naturally supports semantic-level. We further introduce a novel tripartite cooperation adversarial training scheme that involves a conductor network to provide fine-grained semantic feedback for the generator. We also show that the proposed semantic image synthesis method can be utilized as an effective data augmentation approach on improving the performance of the downstream remote sensing image segmentation tasks. Extensive experiments show the superiority of our method compared with the state-of-the-art image synthesis methods.**

***Index Terms*—Semantic image synthesis, generative adversarial networks, remote sensing images, image segmentation.**

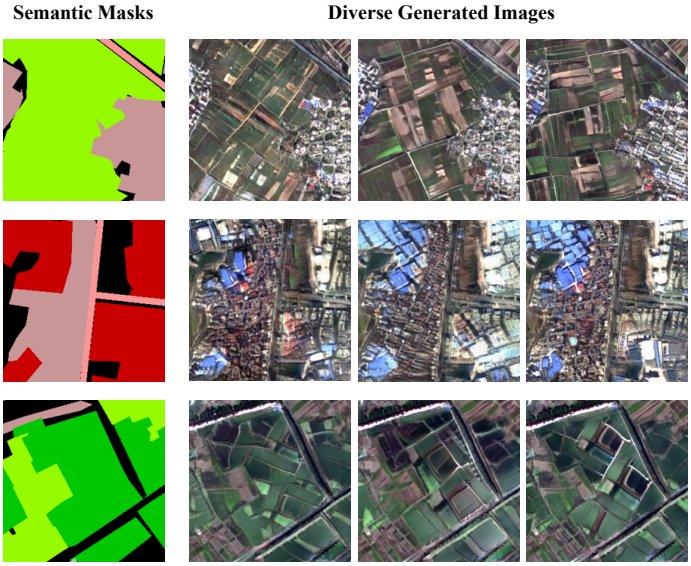

Fig. 1. Demonstration of the diverse results generated by the proposed method. The left column shows the input semantic mask, the right three columns are the generated results conditioned on the input.

I. Introduction

SEMANTIC image synthesis of remote sensing images aims at generating photo-realistic remote sensing images according to their input semantic mask. As a specific form of conditional image synthesis, this task contributes to many practical applications such as semantic image editing [1], [2], detecting deep fake geography [3], and data augmentation [4], and therefore is of great research significance.

To facilitate the aforementioned applications, both visual fidelity and diversity are critical in the semantic image synthesis of remote sensing images. In image synthesis, visual fidelity refers to the realism of synthetic images in spatial and spectral dimensions, including the realism factors such as color, texture, and semantics of the generated ground features. Semantic diversity refers to generating multiple outputs from the same semantic mask [5]–[8]. In the scenario of utilizing image synthesis as superior data augmentation, the reduction in semantic diversity may cause the generator to only produce synthesis outputs that are similar to the original training data and therefore limit the improvement brought by synthetic data. Semantic-level diversity refers to changing specific semantic contents while keeping others untouched [1], [9]. Fine-grained controllability is of great significance in image editing where only editing certain parts in generative images is a common need.

In remote sensing image synthesis, visual fidelity, semantic diversity, and fine-grained controllability are essential. On one hand, many current papers [10]–[14] mainly concentrate on fidelity. They translate semantic masks to images in general image-to-image frameworks by directly feeding semantic masks to an encoder-decoder network [10], [11] or adopting spatially-adaptive normalization [15]. On the other hand, many current methods that concentrate on diversity enhancement also concentrate only on global diversity rather than semantic diversity. They utilize variational autoencoder (VAE) [16] architectures to constrain one-one generation. However, the

The work was supported by the National Key Research and Development Program of China (Grant No. 2022ZD0160401), the National Natural Science Foundation of China under Grant 62125102, the Beijing Natural Science Foundation under Grant JL23005, and the Fundamental Research Funds for the Central Universities. *(Corresponding author: Zhenwei Shi (e-mail: shizhenwei@buaa.edu.cn))*

Chendan Wang, Bowen Chen and Zhenwei Shi are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with the Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Zhengxia Zou is with the Department of Guidance, Navigation and Control, School of Astronautics, Beihang University, Beijing 100191, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

above methods only introduce variety by global noise or variational autoencoder, which leads to a decline in semantic diversity and the absence of fine-grained controllability. Although recent methods attempt to mitigate the two issues through various methods, including concatenating 3D noise [17], replacing convolution layers with group convolutions in generator [1] or sampling from semantic class distributions [9], the improvement brought by synthetic images is still far from satisfactory.

However, the methods mentioned above only focus on everyday objects while failing in remote sensing images due to specific challenges such as the high semantic similarity and high data imbalance of remote sensing object classes. Semantic similarity means that there may be common features at the semantic level between different classes of remote sensing objects, such as the river and the lake. Data imbalance means that there are large differences in the number of pixels between different classes of remote sensing objects, which may lead to difficulties in training. Both are inherent properties of remote sensing images and therefore need to be considered in remote sensing image synthesis. In addition, data imbalance is a key constraint for downstream tasks such as semantic segmentation. If the quality of the synthesized image is good enough, using the synthesized image for data augmentation can improve the performance of downstream tasks. Therefore, it is essential to include the metric of downstream improvement when evaluating synthesized images. However, this metric is often neglected in previous papers [1], [6], [9], [15], [18], [19].

To tickle the above problems, we propose a novel semantic image synthesis method named CSEBGAN (Conducted Semantic EmBedding GAN) based on the special challenges of remote sensing images. Our method not only enables competitive fidelity and fine-grained level diversity but also significantly boosts downstream performance improvement. There are two key components in our method, 1) a novel Semantic Embedding Network and 2) a novel tripartite cooperation training scheme. To enhance diversity and fine-grained controllability, we propose a Semantic Embedding Network (SENet), which encodes different semantic classes to independent Semantic Embeddings. The motivation is that generating specific semantic content is just like painting with certain pigments. During generation, specific semantic contents are generated from semantic vectors sampled from corresponding Semantic Embeddings. Compared to directly feeding semantic mask as input and applying uniform convolutional kernels on the whole mask, SENet dramatically increases diversity and naturally supports controllability over specific semantics. Different from utilizing uniform parameters for all the categories, SENet learns different semantic embeddings for different categories, which increases the attention of generative models on semantic categories with small occupancy and improves the fidelity of the small-occupancy classes. SENet also shares the convolutional parameters for all semantic classes to facilitate the information sharing of different classes. Besides, SENet encodes semantic classes into the embedding space, so that embeddings of similar classes are closer and embeddings of dissimilar classes are farther. For example, embeddings of rivers and lakes are closer, while embeddings of rivers and forests are farther. In this way, it is feasible to utilize the semantic similarity of remote sensing images to reduce the learning difficulty of the generative model.

To further improve the fidelity and consistency of semantic masks, we introduce a novel tripartite cooperation training scheme. In the traditional GAN [20] framework, there are two networks namely generator and discriminator that compete with each other. How semantic information is encoded into the discriminator is a significant design in semantic image synthesis. Most previous methods simply concatenate the semantic mask with the image as the input of discriminator [10], [11], [15], which only provides global feedback and faces the risk of neglecting the semantic mask. Some recent methods [5], [17] redesign the structure of the discriminator to fully utilize the semantic information. In contrast, we introduce a new conductor network to the adversarial training. The conductor itself is restrained by real images to capture meaningful evidence and subsequently provides fine-grained semantic feedback for the generator and thus avoids the possibility of neglecting the semantic mask. The three networks are jointly trained from scratch. Through tripartite cooperation, our method further enhances fidelity and the performance improvement brought by synthetic data. Fig. 1 shows the results of the generated images using the proposed method. Our method generates images with high fidelity and high diversity.

The contributions of this paper are summarized as follows:

1) We propose a new method named SENet for remote sensing image synthesis, which significantly enhances fine-grained generative diversity and utilizes the regularities between semantic classes to improve fidelity.
2) We introduce a conductor in the adversarial training framework, which guides the generator to generate images that are more in accord with our preference.
3) To our best knowledge, we are the first ones to explore the open question of whether semantic image synthesis can be utilized as advanced data augmentation to improve the same domain segmentation performance. The proposed method achieves satisfying improvement.

II. Related Work

A. Semantic Image Synthesis

Semantic image synthesis aims at translating semantic masks to photo-realistic images. Methods based on conditional GANs [21] have seen great progress. Traditional methods fulfill this in a general image-to-image framework [10], [11], [15], [18]. Pix2Pix [10] directly feeds semantic masks to an encoder-decoder architecture. Pix2PixHD [11] improves Pix2Pix by a coarse-to-fine generator and multi-scale discriminators. To prevent the "wash-away" problem caused by normalization layers, SPADE [15] encodes semantic masks into the generator by spatially-adaptive denormalization instead of taking them as inputs. Since then, many methods based on SPADE are proposed to optimize the denormalization parameters. CLADE [6] shows that parameters are almost spatial-invariant within regions with the same semantic class and directly learn semantic-adaptive parameters. SEAN [13], SAFM [19], RESAIL [18] optimize denormalization parameters by

combining parameters from other sources, like style images, position encoding, and image patches in the training set. Apart from adjusting denormalization parameters, CC-FPSE [5] predicts convolutional kernels based on semantic masks while LCGAN [14] combines local and global generators to achieve better fidelity.

The above methods mainly concentrate on improving fidelity. Apart from fidelity, diversity is also of great importance in practice application. To obtain diverse generative results, most methods [5], [6], [11] achieve global diversity through variational autoencoder architectures to constrain one-one generation. BicycleGAN [7] combines constraints in both directions. DSCGAN [8] utilizes regulation to penalize single outputs. To further expand the diversity to the semantic level, GroupDNet [1] leverages group convolutions in the generator and decreases the number of groups in the decoder. OASIS [17] concentrates 3D noise on activation maps. INADE [9] adopts instance-adaptive modulation sampling. However, few methods focus on the downstream performance improvement brought by synthetic data and thus limit the application of adopting them as data augmentation.

Although semantic image synthesis has drawn much attention in the community [13], [14], [18], [19], it is rarely studied in the remote sensing area. Even if the existing methods that achieve high fidelity and diversity for natural images are migrated to remotely sensed images, their performance is still far from satisfactory due to the particular challenges of remote sensing images. Most existing methods adopt the same parameters for all the semantic categories and therefore the models tend to over-concentrate on the quality of categories that occupy a large percentage while neglecting the small-occupied semantic. Furthermore, they simply treat different semantics as independent contents, ignoring the relationship between different semantic categories. Different from most methods, our method learns unique parameters, i.e. semantic embeddings for each semantic category, alleviating the imbalanced attention on semantics and utilizing the regulations of different remote sensing semantics. Also, there are few methods concentrating on semantic-level diversity and performance improvement when serving as augmentation. To the best of our knowledge, we are the first to focus on semantic image synthesis of remote sensing images, propose a method with high output diversity at the semantic level, and explore the possibility to adopt it as superior augmentation in segmentation tasks.

B. GAN-based Image Augmentation

Image augmentation has been proven to be an effective strategy to alleviate the cost of data collecting through sufficiently utilizing existing data [4], [22]. Model-free image augmentation follows artificially defined rules [23]–[25], like flipping, cropping, and rotation to create new images while mode-based image augmentation utilizes pre-trained models to generate new images. The generator for image augmentation aims at capturing the distribution of real data and creating images not existing in the original dataset.

GAN-based image augmentation has been widely studied particularly for the image classification task. Many methods [26]–[29] generate minority-class images based on GANs and therefore obtained higher accuracy classifiers by mitigating the class-imbalanced problem. The methods [30] and [31] concentrated on the medical domain where the original data often contains sensitive information and should be avoided. DAGAN [32] further relieved the need for the class labels, where the generator is trained to produce images that have the same class as input images in the source domain and can be applied in the low-data domain of interest without knowing specific labels.

Given the common practices that augmentation of classification often depends on label-conditional GANs, semantic image synthesis GANs seem a reasonable choice for the augmentation of semantic segmentation. However, due to the low diversity of common semantic image synthesis methods, GAN-based augmentation of segmentation often utilize domain translation methods instead. These methods require well-annotated images in the source domain where the annotations are often generated by computer game engines. SimGAN [33] utilized unlabeled real images to improve the realism of simulated images while preserving annotation information. SUIT [34] proposed a semantic-content loss to preserve the semantic and content information during the translation. [35] jointly trained generator and segmentation network to generalize game data for building segmentation in remote sensing images.

These methods require semantic classes in the target domain to be the subset of those in the source domain, which may limit the applications. In contrast, we investigate the possibility of directly adopting semantic image synthesis as a superior augmentation.

III. Methods

Semantic image synthesis aims at translating a semantic mask $\boldsymbol{m} \in \mathbb{N}^{H \times W}$ to a photo-realistic image $\boldsymbol{o} \in \mathbb{R}^{H \times W \times 3}$ which shares the same height H and width W. Each pixel in $\boldsymbol{m}$ is an index k which denotes a specific semantic class of predefined semantic classes $\{1, 2..., K\}$ and represents the expected semantic of the corresponding location in $\boldsymbol{o}$.

A. Semantic Embedding Network

Given K predefined classes, previous methods usually translate $\boldsymbol{m}$ to a matrix $\in \{0, 1\}^{H \times W \times K}$ and apply uniform convolutional kernel over the whole matrix. This simple operation has several limits. First, only global variety is adopted on the whole map, which limits generation diversity and prevents fine-grained controllability. Second, this operation is essentially equivalent to representing each semantic as a one-hot vector, which ignores the regularities between semantic classes. Inspired by the success of word embedding in Natural Language Processing (NLP) [36], [37], we propose to encode each semantic to an embedding in order to decouple different semantics and leverage the semantic similarity. However, using common embedding as a single vector will face the defect of low diversity. Therefore, we further expand the specific vector of a semantic class to a distribution of embeddings, termed Semantic Embedding. Owing to the sampling processing from

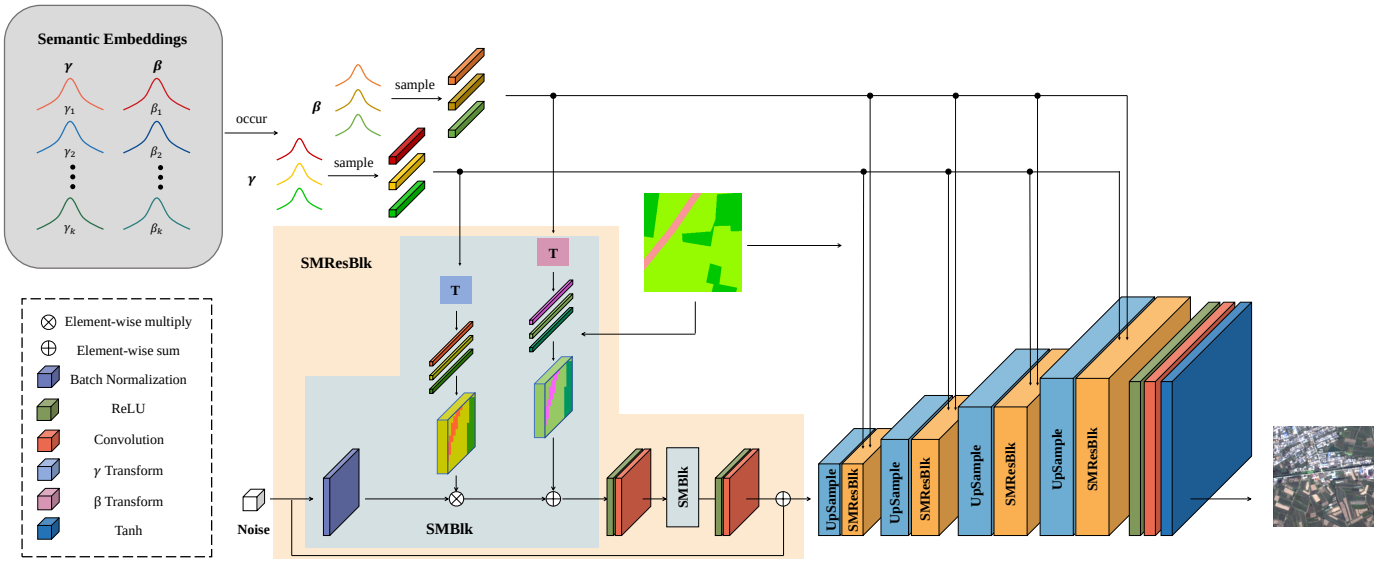

Fig. 2. The overall architecture of the proposed SENet. SENet is composed of several SMResBlks followed by upsampling operations. During the generation, only the semantic classes that occur in the input mask are sampled to get specific semantic vectors. Semantic Embeddings are shared by all the layers and each SMBlk transfers the embeddings to multi-scale parameters.

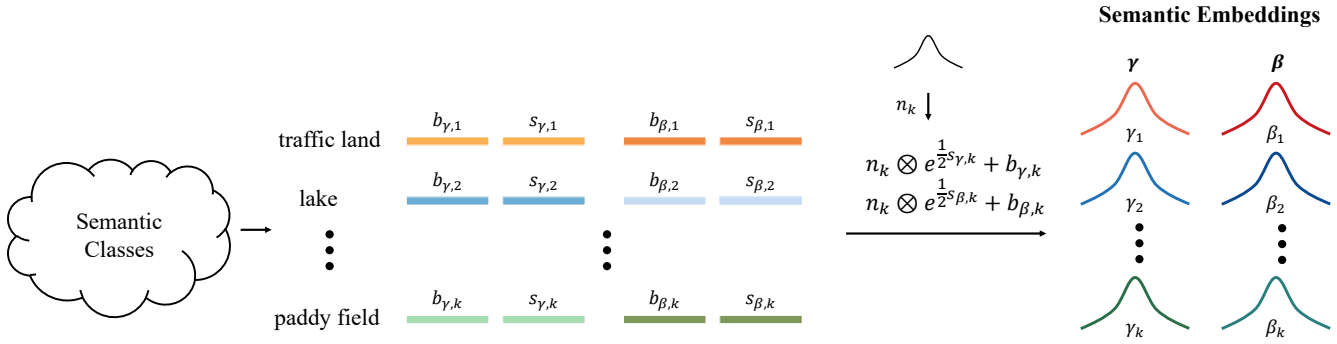

Fig. 3. We decouple different semantic classes into independent Semantic Embeddings. Semantic Embeddings are clusters of embeddings that carry the information of each semantic class.

Semantic Embeddings, generation diversity is fundamentally boosted. We term the generator based on Semantic Embedding as SENet. Fig. 2 illustrates the overall architecture of the SENet.

SENet consists of several SMResBlks with upsampling operations. During the generation, semantic vectors used for the specific process are sampled from corresponding Semantic Embeddings. To guarantee consistency of the same region across generator layers, semantic vectors are shared by all layers while multi-scale modulation parameters are obtained through SMBlks of each layer. Then SMBlks utilize the modulation parameters to control generation contents through semantic-adaptive denormalization.

Fig. 3 shows the process of obtaining Semantic Embeddings. Given the predefined K semantic classes, SENet learns two embedding distributions $\boldsymbol{\gamma}_k$, $\boldsymbol{\beta}_k$ for each semantic class via reparameterization trick [16]. Taking $\boldsymbol{\gamma}_k$ as an example, two vectors $\boldsymbol{b}_{\gamma,k}$, $\boldsymbol{s}_{\gamma,k}$ are assigned to learn the shift and scale of $\boldsymbol{\gamma}_k$ respectively. During the forward propagation, we sample a random noise vector $\boldsymbol{n}_k$ for each semantic from the standard multivariate normal distribution. Then the noise is parameterized by $\boldsymbol{b}_{\gamma,k}$, $\boldsymbol{s}_{\gamma,k}$ of corresponding semantics to get $\boldsymbol{\gamma}_k$:

$$\boldsymbol{\gamma}_k = \boldsymbol{n}_k \otimes e^{\frac{1}{2}\boldsymbol{s}_{\gamma,k}} + \boldsymbol{b}_{\gamma,k} \tag{1}$$

where $\otimes, +$ represents element-wise multiplication and addition respectively. During the backpropagation $\boldsymbol{b}_{\gamma,k}$, $\boldsymbol{s}_{\gamma,k}$ are updated by respective gradients. We utilize the same method to get all the $2 \times K$ semantic embeddings $\{\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_K, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K\}$ which are shared by all layers of SENet.

For a specific SMBlk, let $\boldsymbol{X} \in \mathbb{R}^{H^l \times W^l \times C^l}$ denotes the activation map, and H^l, W^l, C^l denote the height, width, and the number of channels of the activation map respectively. For the R semantic classes that occur in the semantic mask, $\boldsymbol{\gamma}_k, \boldsymbol{\beta}_k$ are then transferred through learned transforms of each SMBlk $f^l_{\gamma,k}: \boldsymbol{\gamma}_k \rightarrow \boldsymbol{\gamma}^l_k, f^l_{\beta,k}: \boldsymbol{\beta}_k \rightarrow \boldsymbol{\beta}^l_k$ respectively to get $\boldsymbol{\gamma}^l_k \in \mathbb{R}^{C^l}$ and $\boldsymbol{\beta}^l_k \in \mathbb{R}^{C^l}$, the semantic modulation parameters for each layer. The learned transforms are achieved by two-layer MLPs in SENet. Then $\{\boldsymbol{\gamma}^l_1, ..., \boldsymbol{\gamma}^l_R\}$, $\{\boldsymbol{\beta}^l_1, ..., \boldsymbol{\beta}^l_R\}$ are tiled to two semantic modulation maps $\boldsymbol{\Gamma} \in \mathbb{R}^{H^l \times W^l \times C^l}$ and $\boldsymbol{B} \in$

$\mathbb{R}^{H^l \times W^l \times C^l}$ respectively according to the specified semantic layout.

In each activation layer, Batch normalization [38] is applied first to get $\hat{\boldsymbol{X}}$. Each activation value $\hat{h}^l_{x,y,c}$ ($x \in W^l, y \in H^l, c \in C^l$) in $\hat{\boldsymbol{X}}$ is defined as

$$\hat{h}^l_{x,y,c} = \frac{h^l_{x,y,c} - \mu^l_c}{\sigma^l_c}, \tag{2}$$

where $h^l_{x,y,c}$ is the corresponding activation value in $\boldsymbol{X}$ and μ^l_c, σ^l_c are the mean and standard deviation of channel c respectively:

$$\mu^l_c = \frac{1}{NH^lW^l} \sum_{x,y} h^l_{x,y,c} \tag{3}$$

$$\sigma^l_c = \sqrt{\frac{1}{NH^lW^l} \sum_{x,y} (h^l_{x,y,c} - \mu^l_c)^2}. \tag{4}$$

Then the semantic-adaptive denormalization is applied by $\boldsymbol{\Gamma}$ and $\boldsymbol{B}$ to control the contents of generated images. The output tensor of SMBlk is

$$\hat{\boldsymbol{X}} \otimes \boldsymbol{\Gamma} + \boldsymbol{B}, \tag{5}$$

where $\otimes, +$ represents element-wise multiplication and addition respectively.

Due to the abundant information in Semantic Embeddings, SENet significantly improves generation diversity. Due to the independence of different semantic classes, SENet inherently supports fine-grained controllability. To take advantage of the high similarity between remote sensing classes, SENet leverages the regularities between classes through embedding space. To tackle the large differences in the percentage of classes, SENet decouples different semantics to avoid over-focusing on large occupancy semantics.

B. Tripartite Cooperation Training Scheme

In the traditional GAN training framework, there are two networks competing with each other. The generator creates fake images while the discriminator aims to distinguish fake images from real ones. Most previous methods [5], [17] design new discriminator network structures to better utilize the semantic masks. This means that the discriminator needs to accomplish judging the fidelity and the coherence with the semantic mask at the same time, which dramatically increases the learning difficulty of the discriminator and can result in the inability to bring effective feedback to the generator.

In contrast, we introduce a new network termed conductor to the traditional training scheme. In the tripartite training scheme, the discriminator provides global fidelity feedback while the conductor provides local semantic feedback. The whole structure is shown in Fig. 4. We design the conductor as a semantic segmentation network that takes images as input and outputs classification for every pixel. We adopt traditional encoder-decoder architecture as the structure of the conductor. We use pixel-wise cross entropy between the output of the conductor and ground truth segmentation mask $\boldsymbol{s}$ as its loss function.

The conductor is jointly trained with the generator and the discriminator. The training objective of the conductor includes two parts, i.e. $\mathcal{L}_c^{real}$, which is related to real images, and $\mathcal{L}_c^{fake}$, which is related to fake ones. When the conductor receives real images, it is trained to learn meaningful information while when it receives fake images, it provides segmentation results as feedback to the generator. Through the tripartite cooperation of the conductor, generator, and discriminator, the fidelity and performance improvement brought by synthetic data can be boosted.

C. Loss Functions

1) Feature matching loss: Let L_D be the number of layers of the discriminator, $D^{(i)}(*)$ be the i-th feature map of the discriminator, and N_i be the number of elements of $D^{(i)}$. We exclude the final prediction from the last layer of the discriminator and define the feature-matching loss as follows:

$$\mathcal{L}_{fm} = \mathbb{E} \sum_{i=1}^{L_D - 1} \frac{1}{N_i} [||D^{(i)}(\boldsymbol{o}, \boldsymbol{m}) - D^{(i)}(G(\boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\beta}), \boldsymbol{m})||_1]. \tag{6}$$

$\mathbb{E}$ represents the mathematical expectation, which is the mean value of all samples.

2) Perceptual loss: We adopt the pre-trained VGG network [39] to compute perceptual loss. Let L_F be the number of layers of the VGG network, $F^{(i)}(*)$ be the i-th feature map of the VGG network, and M_i be the number of elements of $F^{(i)}$. The perceptual loss is denoted as follows:

$$\mathcal{L}_{perc} = \mathbb{E} \sum_{i=1}^{L_F} \frac{1}{M_i} [||F^{(i)}(\boldsymbol{o}) - F^{(i)}(G(\boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\beta}))||_1]. \tag{7}$$

3) Conductor loss: The loss function of the conductor $\mathcal{L}_c$ is defined as follows:

$$\mathcal{L}_c = \mathcal{L}_c^{real} + \mathcal{L}_c^{fake}. \tag{8}$$

where $\mathcal{L}_c^{real}$ is the loss function on real images and $\mathcal{L}_c^{fake}$ is the loss function on the fake images. The loss terms $\mathcal{L}_c^{real}$ and $\mathcal{L}_c^{fake}$ are defined as follows:

$$\mathcal{L}_c^{real} = -\mathbb{E} \sum_{k=1}^{K} \log \frac{\exp(C_k(\boldsymbol{o}))}{\sum_{k=1}^{K} \exp(C_i(\boldsymbol{o}))} \boldsymbol{m}, \tag{9}$$

where K is the total number of semantic classes, $\boldsymbol{o}$ is the real image, $C_k(*)$ is the k-th channel output of conductor and $\boldsymbol{m}$ is the ground truth segmentation mask of $\boldsymbol{o}$.

$$\mathcal{L}_c^{fake} = -\mathbb{E} \sum_{k=1}^{K} \log \frac{\exp(C_k(G(\boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\beta})))}{\sum_{i=1}^{K} \exp(C_i(G(\boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\beta})))} \boldsymbol{m}, \tag{10}$$

where $\boldsymbol{\gamma}, \boldsymbol{\beta}$ are Semantic Embeddings and $G(\boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\beta})$ is the synthetic image of generator. Although $\mathcal{L}_c^{real}$ and $\mathcal{L}_c^{fake}$ are very similar, they bring different effects. For $\mathcal{L}_c^{real}$, the real images fit the semantic mask, so the conductor can be trained to learn meaningful information. For $\mathcal{L}_c^{fake}$, the conductor provides the segmentation results, which are fed back to the generator via back-propagation and thus guide the image generation.

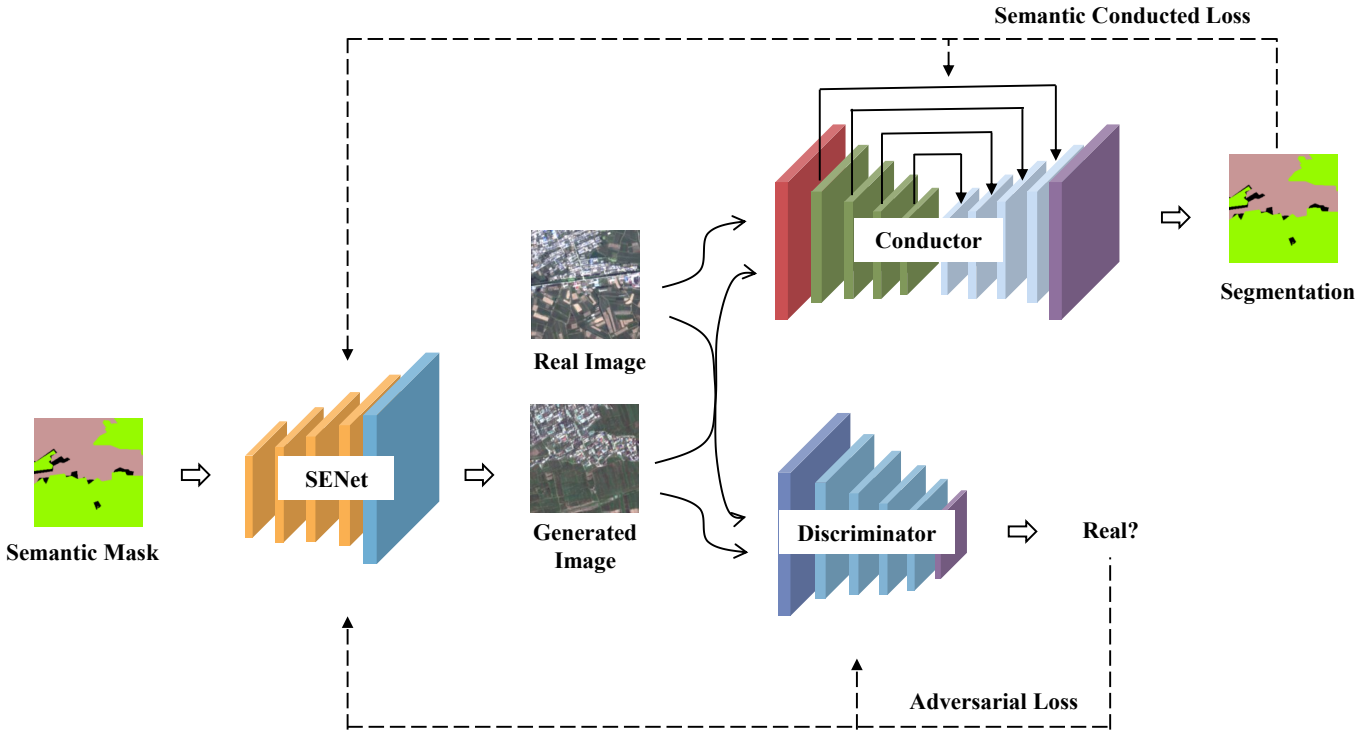

Fig. 4. The proposed tripartite cooperation training scheme. The discriminator provides coarse-grained image generation feedback while the conductor provides fine-grained semantic feedback.

4) Conditional adversarial loss: We adopt hinge-based adversarial loss and therefore the optimization of the generator and discriminator can be written as follows:

$$\mathcal{L}_{adv}^{G} = -\mathbb{E}D(G(\boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\beta}), \boldsymbol{m})), \tag{11}$$

$$\begin{aligned}\mathcal{L}_{adv}^{D} = &-\mathbb{E}[\min(0, -1 + D(\boldsymbol{o}, \boldsymbol{m}))] \\ &-\mathbb{E}[\min(0, -1 - D(G(\boldsymbol{m}, \boldsymbol{\gamma}, \boldsymbol{\beta}), \boldsymbol{m}))],\end{aligned} \tag{12}$$

where $D(*)$ is the prediction of discriminator.

5) Total loss: The generator is trained with conditional adversarial loss $\mathcal{L}_{adv}$, feature matching loss $\mathcal{L}_{fm}$, perceptual loss $\mathcal{L}_{perc}$ and conductor loss $\mathcal{L}_{c}$. The total loss is defined as follows:

$$\min_{G,C}(\max_{D}(\mathcal{L}_{adv}^{D}) + \mathcal{L}_{adv}^{G} + \lambda_{fm}\mathcal{L}_{fm} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{c}\mathcal{L}_{c}), \tag{13}$$

where we set $\lambda_{fm} = 10$, $\lambda_{perc} = 10$ and $\lambda_{c} = 1$ in our experiments. Except for the conductor loss, other objectives follow conventional image translation methods [11], [15].

IV. Experiments

A. Experimental Setup

1) Datasets: We conduct the experiments on GID [40], a large-scale land-cover classification dataset composed of Gaofen-2 (GF-2) satellite images. We utilize the Fine Land-cover Classification part GID-15 which consists of ten 6800×7200 remote sensing images and corresponding semantic masks. GID-15 includes 15 semantic classes, i.e. industrial land, urban residential, rural residential, traffic land, paddy field, irrigated land, dry cropland, garden plot, arbor woodland, shrub land, natural grassland, artificial grassland, river, lake, and pond. We cut the images into 256×256 image patches without overlap and filter out those images with more than eighty percent of pixels of unknown class. We divide the remaining 1680 paired images into the training set, validation set, and test set according to a 4:1:1 ratio and reserve an additional 1120 pairs for the ablation experiments.

2) Implement Details: We adopt Adam optimizer [41] and set $\beta_1 = 0$, $\beta_2 = 0.9$. We utilize TTUR strategy [42] for training. The initial learning rates for the generator, discriminator, and conductor are set to 0.0001, 0.0004, and 0.0001, respectively. The models are trained for 200 epochs and the learning rates decayed to zero in the last 100 epochs. The whole framework is implemented by Pytorch and the experiments are performed on a single GeForce RTX 3090 GPU.

3) Fidelity Metrics: We adopt Fréchet Inception Distance(FID) [42] and segmentation-based metrics to evaluate generative quality. FID measures the distance of the feature distributions between generated images and real ones. Lower FID indicates higher generative quality. As for the segmentation-based metrics, we evaluate the Intersection-over-Union (mIoU), Frequency Weighted Intersection-over-Union (FWIoU), and pixel accuracy (acc) of a pre-trained model on generated images. The higher the fidelity is, the better performance of a well-trained segmentation model will be. We adopt U-Net [43] as the segmentation model and train it for 200 epochs.

4) Diversity Metrics: We adopt LPIPS [44] to evaluate global diversity. LPIPS measures the L_1 distance of weighted

TABLE I
QUANTITATIVE COMPARISON WITH THE STATE-OF-THE-ART METHODS.

Methods	Fidelity				Diversity			Downstream Improvement		
	FID ↓	mIoU ↑	FWIoU ↑	acc ↑	LPIPS ↑	mCSD ↑	mOCD ↓	d-mIoU ↑	d-FWIoU ↑	d-acc ↑
Pix2PixHD	219.91	49.71	74.16	83.75	0	-	-	-10.84%	-5.18%	-3.02%
SPADE	191.27	**68.69**	**89.80**	**94.50**	0	-	-	+4.58%	+2.35%	+1.17%
VGanGAN	<u>166.75</u>	64.04	87.09	92.74	0.1839	0.0445	0.0982	-0.06%	-0.12%	-0.40%
GroupDNet	231.55	45.94	69.02	80.35	0.0914	0.0038	0.0141	+3.63%	+2.05%	+1.06%
INADE	167.05	62.56	85.75	92.02	<u>0.3087</u>	<u>0.0714</u>	**0.0104**	<u>+5.85%</u>	<u>+2.52%</u>	<u>+1.29%</u>
Ours	**155.29**	<u>67.50</u>	<u>89.39</u>	<u>94.10</u>	**0.3119**	**0.0743**	<u>0.0134</u>	**+6.59%**	**+3.67%**	**+1.97%**

perceptual similarity between two images. To evaluate diversity, we generate 10 images for each semantic mask and randomly sample 10 pairs per mask to compute mean LPIPS as the final results. A higher LPIPS score indicates better global diversity. To evaluate the semantic-level diversity, we adopted mCSD and mOCD by following [1], [9]. mCSD measures how well the diversity over the specific semantic classes while mOCD measures how well other classes are kept unchanged. For mCSD, higher scores are better. For mOCD, lower scores are better.

5) Downstream Performance Improvement: To evaluate the performance improvement brought by the generated images to downstream segmentation tasks, we first train a baseline model on the original training set and then compare its performance with models trained from both the training set and generated images. We adopt U-Net [43] as the structure without deliberate selection. The model consists of a contracting path and an expansive path. The contracting path downsamples the image and the expansive path upsamples the features. The output of each stage in the contracting path is copied to the corresponding stage in the expansive path using the skip connection. The model is trained with Adam optimizer and is trained for the same iterations both under the augmentation and non-augmentation settings. We measure the percentage improvement on mIoU, FWIoU, and acc. The corresponding metrics are termed d-mIoU, d-FWIoU, and d-acc.

B. Comparison with other methods

We compare our method with several state-of-the-art semantic image synthesis methods, including Pix2PixHD [11], SPADE [15], GroupDNet [1], and INADE [9]. The Pix2PixHD are quality-oriented while the GroupDNet and INADE are diversity-oriented. We also trained SPADE with its diversity strategy which is termed VSPADE. All the models are retrained using the codes provided by the authors. The quantitative results are summarized in Tab. I. Bold and underlined numbers are the best and the second best of each metric respectively. We do not evaluate mCSD and mOCS for the methods that do not support diversity generation and the corresponding numbers are replaced by '-'.

In the evaluation dimension of fidelity, our method achieves better generative quality than other methods in terms of FID. Although our method is slightly inferior on segmentation-based metrics with the difference of 1.7%, 0.5% and 0.4% on mIoU, FWIoU, and acc respectively compared to SPADE, our method achieves 18% superior FID. Compared to methods with semantic-level diversity, our method achieves the best performance over all the fidelity metrics.

As for diversity, our method outperforms all other methods on a global level, achieving the highest LPIPS. When it comes to semantic-level diversity metric, our method achieves competitive results with INADE and outperforms other methods. Our method is superior to INADE in terms of the diversity of specific semantic regions but inferior in terms of keeping other regions unchanged. However, when specific semantics are changed, other areas will also appear to be slightly altered to ensure the harmony of the whole image. A slightly high mOCD does not necessarily mean a worse performance.

The key advantage of our method is the performance improvement in downstream segmentation tasks. Both SPADE, which is optimal in terms of fidelity, and INADE, which is optimal in terms of diversity, do not perform as well as our method in terms of bringing gains to the downstream segmentation tasks. This is because generative models often trade off between fidelity and diversity due to model capacity constraints, just as SPADE outperforms VSPADE in terms of fidelity while SPADE outperforms SPADE in terms of diversity. Our method achieves competitive fidelity and diversity at the same time and therefore achieves superior downstream task gain. Our method improves the feasibility of adopting semantic image synthesis as data augmentation. Fig. 5 shows the qualitative results between our method and others. Fig. 6 shows more generative results of our method. Fig. 7 gives examples of semantic-level diversity generation. The above experiments show that our method achieves comparable fidelity and diversity and superior result in downstream improvement.

C. Ablation Study

Tab. II summarizes ablation results on CSEBGAN. We performed relevant ablation experiments to validate the effectiveness of the Semantic Embedding and tripartite cooperation training scheme.

For Semantic Embedding, a key point of our design is that each layer of the generative network shares the Semantic Embeddings, so we designed ablation experiments to verify the effectiveness of this design. The w/o sharing item represents that each layer of the generative network does not share the Semantic Embeddings but learns its own set of Semantic Embeddings, which is similar to CLADE and INADE. The

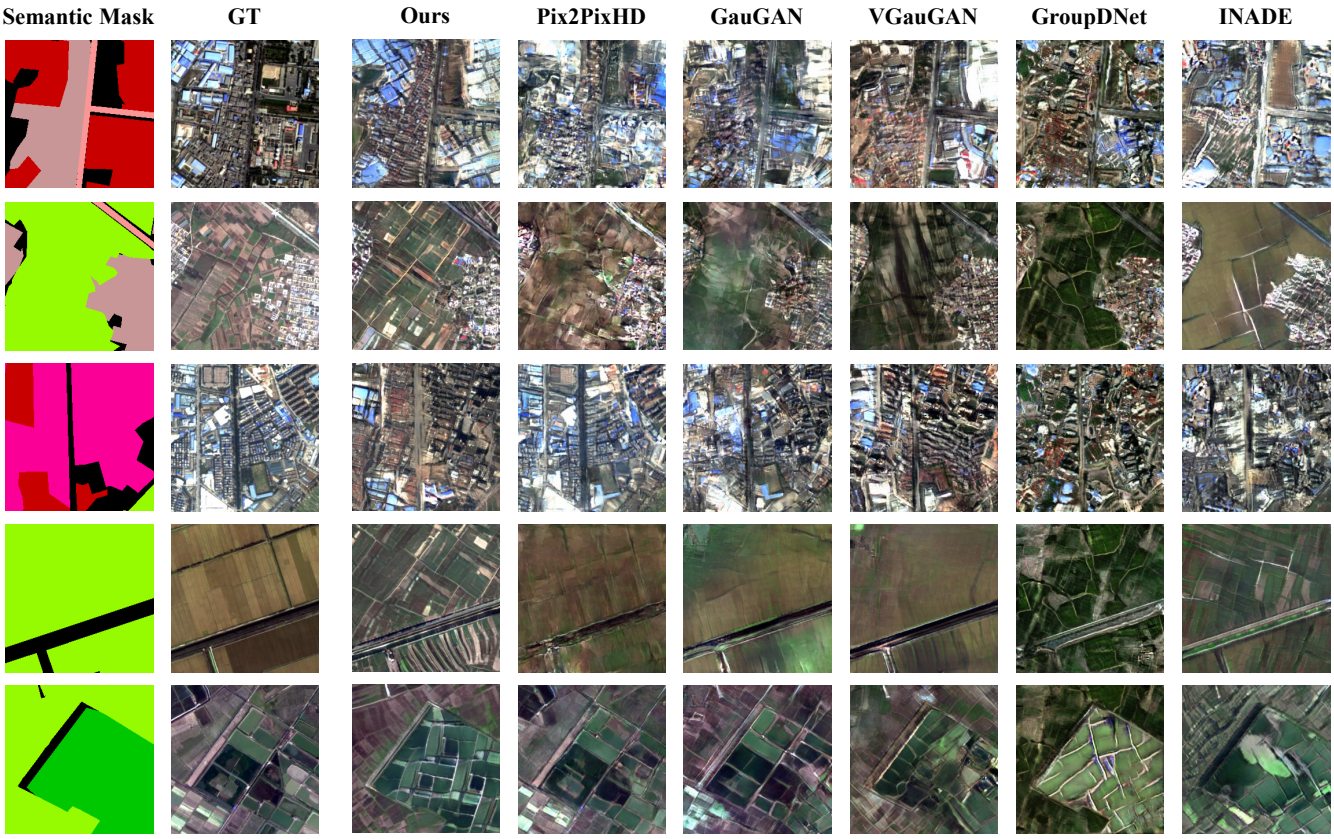

Fig. 5. Qualitative comparison with the SOTA semantic image synthesis methods. The first column represents semantic masks and the second column represents ground truth. From left to right are generated results of our method, Pix2PixHD, SPADE, VSPADE, GroupDNet, and INADE.

TABLE II
ABLATION ANALYSIS OF OUR METHOD.

	Models	Fidelity				Diversity			Downstream Improvement		
		FID ↓	mIoU ↑	FWIoU ↑	acc ↑	LPIPS ↑	mCSD ↑	mOCD ↓	d-mIoU ↑	d-FWIoU ↑	d-acc ↑
	CSEBGAN	**155.29**	**67.50**	**89.39**	94.10	0.3119	**0.0743**	0.0134	**+6.59%**	**+3.67%**	**+1.97%**
Semantic Embedding	w/o sharing	156.37	66.32	88.35	93.40	0.3035	0.0554	0.1342	+6.40%	+3.49%	+1.92%
	w/o C	165.92	62.45	86.51	92.53	**0.3440**	0.0338	0.0045	+3.39%	+2.27%	+1.23%
Conductor	frozen C	156.66	66.83	89.03	**94.13**	0.3360	0.0720	0.0170	+5.38%	+1.66%	+0.85%
	w/o real	159.29	64.43	86.41	92.57	0.3168	0.0709	0.0175	+6.38%	+2.64%	+1.47%
Discriminator	w/o D	235.71	5.87	25.16	45.38	0.0067	0.0014	**0.0004**	+4.26%	+0.23%	-0.10%

inconsistency of each layer causes defects in fidelity, diversity, and downstream improvement. The experimental results indicate the superiority of sharing Semantic Embeddings to ensure the consistency of the same region while adopting unique SMBlk in each layer to obtain the multi-scale control parameters.

Since we implement the tripartite cooperative training scheme with the help of the conductor and the discriminator, we focus on the changes to the conductor and the discriminator.

For the conductor, we first verified the effect of its presence or absence on image generation performance. We found that excluding the conductor in the training framework (w/o C) hurts fidelity and downstream improvement. We then discuss the question of whether it is possible to use a pre-trained semantic segmentation network (frozen C) in the training process directly, which is equivalent to the utilization of segmentation loss in [19]. Compared to the frozen conductor, our method obtains a lower FID and higher mIoU, FWIoU. In general, our method obtains higher generation quality compared to directly adopting segmentation loss. Finally, we experimented with the need to constrain the Conductor utilizing real images. Compared with not utilizing real images to constrain the conductor (w/o real) during the training process, our method achieves superior results, indicating the necessity of using real images to constrain the conductor.

For the discriminator, we verified the impact of its presence or absence on the performance of image generation, and the experimental results show that excluding the discriminator in the training framework (w/o D) hurts fidelity and downstream improvement. Experiments on both conductor and discriminator indicated the superiority of the tripartite cooperation

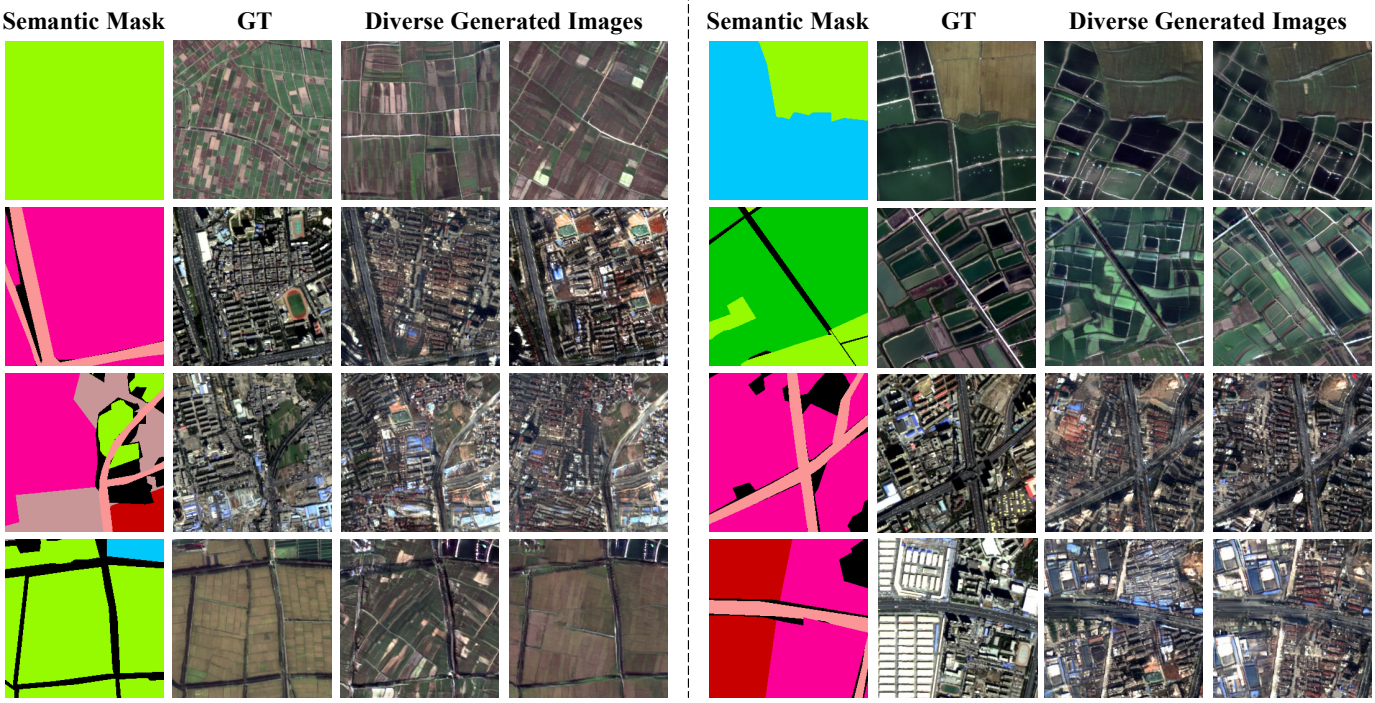

Fig. 6. Generated images of our method. The first left column shows the semantic masks and the second left column shows the ground truth. The right two columns are the generated images of our method.

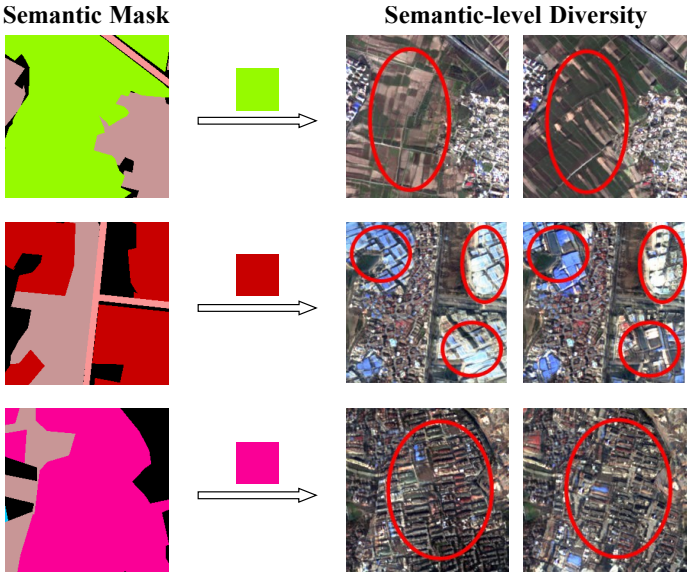

Fig. 7. Illustration of semantic-level generation. The first column shows the input semantic masks. The second column shows the target semantic class. The right two columns are the generated results. The target areas are denoted by red circles.

adversarial training scheme.

D. Downstream Improvement Study

The proposed method also enables superior data augmentation and thus alleviates the pressure of collecting real data [4]. Large and high-quality datasets usually serve as an essential tonic in deep learning methods [23], [43], [45]. However, dataset-collecting can be extremely laborious and expensive, especially for pixel-wise labels data in semantic segmentation [46]. An ideal semantic image synthesis generator can capture the conditional distribution of real images given semantic masks and generate data not existing in the original data set. Then the paired generated images and semantic masks can be used to train models together with the original training data and consequently improve the segmentation performance as a superior data augmentation. To the best of our knowledge, we are the first to adopt semantic image synthesis as data augmentation to improve downstream segmentation performance. We conduct extensive experiments in this area.

To better analyze the performance improvement of each semantic class, we first counted the percentage of pixels in the training set for each semantic category out of the total number of pixels, and the result is shown in Fig. 8. We use our method to double the original data for each semantic class. The visualization of the IoU boost brought on each semantic class by training the segmentation model with the augmented data is shown in Fig. 9. It can be seen that adopting our method as data augmentation has a significant improvement for semantic classes that originally accounted for a small proportion and had poor segmentation results. For example, the IoU of garden plot is increased by 219.85%, and the IoU of shrub land is increased by 336.54%. Even for semantic categories that originally had good segmentation results, such as arbor woodland and irrigated land, our method also improves the performance. However, for some specific semantic classes, such as rural residential in this experiment, the result decreases slightly after adding generated data. We speculate that the reason for this is that the texture details

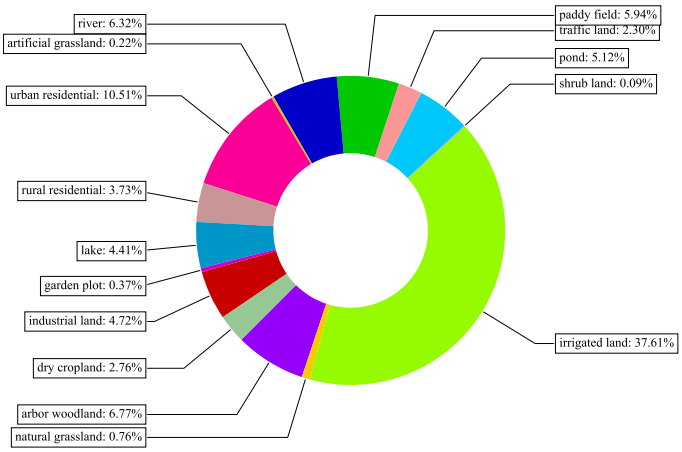

Fig. 8. Percentage of pixels in the training set for each semantic class out of the total number of pixels.

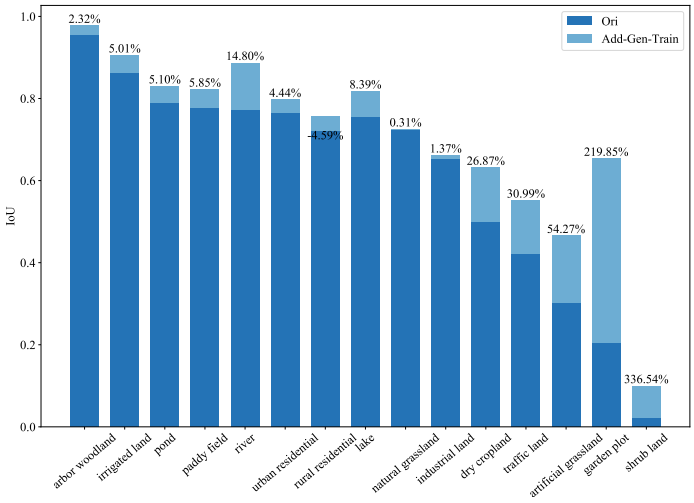

Fig. 9. The horizontal coordinate are the semantic classes in decreasing order of original IoU performance. Ori represents the IoU of the performance of the model trained with the original training set. Add-Gen-Train represents the boosted performance of the model trained on the augmented dataset generated by our method. The numbers give percentage boosts for each class.

of rural residential are more complex compared to classes such as irrigated land and occupy a smaller portion of the dataset, which in turn leads to a bias in the estimation of such distributions. Fig. 10 visualizes quantity comparisons of segmentation results before and after data augmentation. It can be seen that the semantic segmentation model obtains better results after data augmentation.

We also conduct experiments on original training sets of different sizes. The results are illustrated in Tab. III. We first randomly select training subsets of different sizes as 280, 560, 840, and 1120 respectively in the original training set. After that, to guarantee the fairness of experiments, we retrain all the generative models on each training subset and augment each training subset with twice as much data using the trained generative models. Finally, we train each segmentation model on the augmented training subsets and evaluate the performance improvement due to data augmentation by using

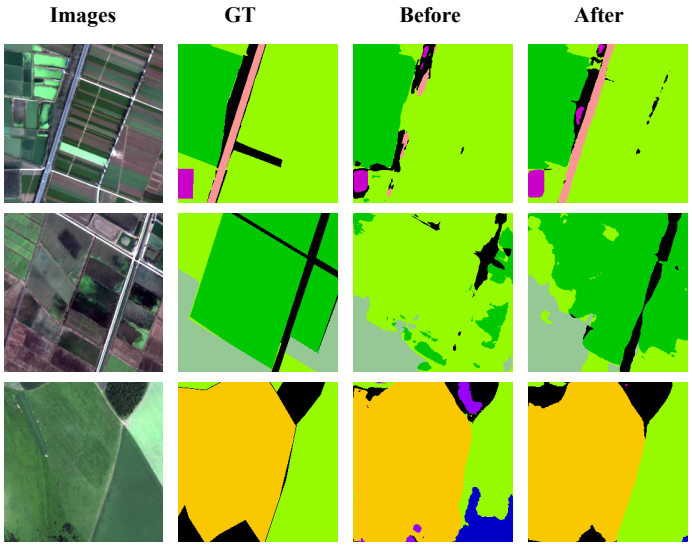

Fig. 10. The left two columns are the images and corresponding ground truth semantic masks. The right two columns are the segmentation results before and after augmentation.

TABLE III
DOWNSTREAM IMPROVEMENT ON DIFFERENT SIZES OF DATASETS.

sizes of datasets	d-mIoU	d-FWIoU	d-acc
280	+7.81%	+3.73%	+2.30%
560	+4.12%	+2.82%	+1.55%
840	+3.46%	+4.52%	+2.55%
1120	+6.59%	+3.67%	+1.97%

our method. The experimental results demonstrate that our method can bring performance improvement regardless of the size of the original training sets, proving the generalizability of our method.

V. CONCLUSION

In this paper, we propose a novel method for remote sensing image synthesis, which achieves both competitive fidelity and diversity. The proposed method decouples semantic classes into independent Semantic Embeddings to boost fine-grained diversity and takes advantage of regularities between semantic classes to improve fidelity. The proposed method further introduces a conductor network in training to provide pixel-wise semantic feedback. To our best knowledge, we are also the first to utilize semantic image synthesis as data augmentation to improve the downstream performance of remote sensing image segmentation. Compared with the state-of-the-art methods, the proposed method shows significant improvement when serving as data augmentation.

REFERENCES

[1] Z. Zhu, Z. Xu, A. You, and X. Bai, "Semantically multi-modal image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5467–5476.

[2] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, "Semantic image synthesis via diffusion models," *arXiv preprint arXiv:2207.00050*, 2022.

[3] B. Zhao, S. Zhang, C. Xu, Y. Sun, and C. Deng, "Deep fake geography? when geospatial data encounter artificial intelligence," *Cartography and Geographic Information Science*, vol. 48, no. 4, pp. 338–352, 2021.

[4] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A comprehensive survey of image augmentation techniques for deep learning," *arXiv preprint arXiv:2205.01491*, 2022.

[5] X. Liu, G. Yin, J. Shao, X. Wang *et al.*, "Learning to predict layout-to-image conditional convolutions for semantic image synthesis," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[6] Z. Tan, D. Chen, Q. Chu, M. Chai, J. Liao, M. He, L. Yuan, G. Hua, and N. Yu, "Efficient semantic image synthesis via class-adaptive normalization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[7] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] D. Yang, S. Hong, Y. Jang, T. Zhao, and H. Lee, "Diversity-sensitive conditional generative adversarial networks," *arXiv preprint arXiv:1901.09024*, 2019.

[9] Z. Tan, M. Chai, D. Chen, J. Liao, Q. Chu, B. Liu, G. Hua, and N. Yu, "Diverse semantic image synthesis via probability distribution modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7962–7971.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[11] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.

[12] 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 International Conference on Computer Vision*, 2017, pp. 2223–2232.

[13] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "Sean: Image synthesis with semantic region-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5104–5113.

[14] H. Tang, D. Xu, Y. Yan, P. H. Torr, and N. Sebe, "Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7870–7879.

[15] 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*, 2019, pp. 2337–2346.

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

[17] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, "You only need adversarial supervision for semantic image synthesis," *arXiv preprint arXiv:2012.04781*, 2020.

[18] Y. Shi, X. Liu, Y. Wei, Z. Wu, and W. Zuo, "Retrieval-based spatially adaptive normalization for semantic image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 224–11 233.

[19] Z. Lv, X. Li, Z. Niu, B. Cao, and W. Zuo, "Semantic-shape adaptive feature modulation for semantic image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 214–11 223.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[21] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[22] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?" in *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2016, pp. 1–6.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[24] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[25] Y. Wang, R. Huang, S. Song, Z. Huang, and G. Huang, "Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 960–11 973, 2021.

[26] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "Bagan: Data augmentation with balancing gan," *arXiv preprint arXiv:1803.09655*, 2018.

[27] A. Ali-Gombe and E. Elyan, "Mfc-gan: class-imbalanced dataset classification using multiple fake class generative adversarial network," *Neurocomputing*, vol. 361, pp. 212–221, 2019.

[28] H. Yang and Y. Zhou, "Ida-gan: a novel imbalanced data augmentation gan," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 8299–8305.

[29] F. H. K. d. S. Tanaka and C. Aranha, "Data augmentation using gans," *arXiv preprint arXiv:1904.09135*, 2019.

[30] F. Konidaris, T. Tagaris, M. Sdraka, and A. Stafylopatis, "Generative adversarial networks as an advanced data augmentation technique for mri data." in *VISIGRAPP (5: VISAPP)*, 2019, pp. 48–59.

[31] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[32] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.

[33] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2107–2116.

[34] R. Li, W. Cao, Q. Jiao, S. Wu, and H.-S. Wong, "Simplified unsupervised image translation for semantic segmentation adaptation," *Pattern Recognition*, vol. 105, p. 107343, 2020.

[35] Z. Zou, T. Shi, W. Li, Z. Zhang, and Z. Shi, "Do game data generalize well for remote sensing image segmentation?" *Remote Sensing*, vol. 12, no. 2, p. 275, 2020.

[36] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[40] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[42] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[43] 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.

[44] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[45] 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*, 2015, pp. 3431–3440.

[46] Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler, "Datasetgan: Efficient labeled data factory with minimal human effort," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 145–10 155.

Chendan Wang received the B.S. and M.S. degrees from the School of Astronautics, Beihang University, Beijing, China, in 2020 and 2023, respectively. Her research interests include image processing, machine learning, and pattern recognition.

Bowen Chen received his B.S. degree from China University of Petroleum East China, Qingdao, Shandong, China, in 2022. He is currently working toward his master's degree in the Image Processing Center, School of Astronautics, Beihang University.

His research interests include machine learning and pattern recognition.

Zhengxia Zou received his B.S. degree and his Ph.D. degree from the Image Processing Center, School of Astronautics, Beihang University in 2013 and 2018, respectively. He is currently a Professor at the School of Astronautics, Beihang University. During 2018-2021, he was a postdoc research fellow at the University of Michigan, Ann Arbor. His research interests include computer vision and related problems in remote sensing and autonomous driving. He has published more than 20 peer-reviewed papers in toptier journals and conferences, including TPAMI, TIP, TGRS, CVPR, ICCV, AAAI. His research has been featured in more than 30 global tech media outlets and adopted by multiple application platforms with over 50 million users worldwide. His personal website is https://zhengxiazou.github.io/.

Zhenwei Shi (Member, IEEE) received a Ph.D. degree in mathematics from the Dalian University of Technology, Dalian, China, in 2005.

He was a Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, China, from 2005 to 2007. He was Visiting Scholar with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA., from 2013 to 2014. He is currently a Professor and Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing. He has authored or co-authored over 200 scientific articles in refereed journals and proceedings, including the IEEE Transactions on Pattern Analysis and Machine Intelligence, the IEEE Transactions on Image Processing, the IEEE Transactions on Geoscience and Remote Sensing, the IEEE Geoscience and Remote Sensing Letters, the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) and the IEEE International Conference on Computer Vision (ICCV). His current research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Dr. Shi serves as an Editor for IEEE Transactions on Geoscience and Remote Sensing, Pattern Recognition, ISPRS Journal of Photogrammetry and Remote Sensing, Infrared Physics and Technology, etc. His personal website is http://levir.buaa.edu.cn/.