# Semi-supervised Deep Learning via Transformation Consistency Regularization for Remote Sensing Image Semantic Segmentation

Bin Zhang, Yongjun Zhang, Yansheng Li, Yi Wan, Haoyu Guo, Zhi Zheng, and Kun Yang

***Abstract*— Deep convolutional neural networks have gotten a lot of press in the last several years, especially in domains like computer vision and remote sensing (RS). However, achieving superior performance with deep networks highly depends on a massive number of accurately labeled training samples. In real-world applications, gathering a large number of labeled samples is time consuming and labor intensive, especially for pixel-level data annotation. This dearth of labels in land-cover classification is especially pressing in the RS domain because high-precision high-quality labeled samples are extremely difficult to acquire, but unlabeled data are readily available. In this study, we offer a new semisupervised deep semantic labeling framework for the semantic segmentation of high-resolution RS images to take advantage of the limited amount of labeled examples and numerous unlabeled samples. Our model uses transformation consistency regularization to encourage consistent network predictions under different random transformations or perturbations. We try three different transforms to compute the consistency loss and analyze their performance. Then, we present a deep semisupervised semantic labeling technique by using a hybrid transformation consistency regularization. A weighted sum of losses, which contains a supervised term computed on labeled samples and an unsupervised regularization term computed on unlabeled data, may be used to update the network parameters in our technique. Our comprehensive experiments on two RS datasets confirmed that the suggested approach utilized latent information from unlabeled samples to obtain more precise predictions and outperformed existing semisupervised algorithms in terms of performance. Our experiments further demonstrated that our semisupervised semantic labeling strategy has the potential to partially tackle the problem of limited labeled samples for high-resolution RS image land-cover segmentation.**

***Index Terms*—Consistency regularization, convolutional neural network (CNN), semantic segmentation, semisupervised learning (SSL), unlabeled data, remote sensing (RS) imagery.**

Manuscript received 22 March 2022; revised 14 May 2022 and 3 August 2022; accepted 25 August 2022. Date of publication 2 September 2022; date of current version 7 July 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42030102 and Grant 41971284 and in part by the Fund for Innovative Research Groups of the Hubei Natural Science Foundation under Grant 2020CFA003. *(Corresponding authors: Yongjun Zhang; Yansheng Li.)*

Bin Zhang, Yongjun Zhang, Yansheng Li, Yi Wan, Haoyu Guo, and Zhi Zheng are with the Department of Photogrammetry, School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: bin.zhang@whu.edu.cn; zhangyj@whu.edu.cn; yansheng.li@whu.edu.cn; yi.wan@whu.edu.cn; haoyu.guo@whu.edu.cn; zhengzhi@whu.edu.cn).

Kun Yang is with the Basic Geographic Information Center of Guizhou Province, Guizhou 550009, China (e-mail: 413135739@qq.com).

Digital Object Identifier 10.1109/JSTARS.2022.3203750

## I. Introduction

IN IMAGE processing, semantic segmentation is a high-level and challenging mission, whose purpose is to assign a semantic class label to every pixel [1]. Semantic segmentation has received considerable critical attention in the domain of computer vision (CV) and remote sensing (RS). In the RS domain, the classification and segmentation of RS images is one of the research hot spots, and it can get a variety of land-use and land-cover (LULC) classification maps for subsequent RS research and applications, for example, long-term series of land-use cover analysis [2], [3], [4], [5]. Traditional machine learning (ML) approaches for image classification and detection generally utilized prior knowledge of experts to select and extract hand-crafted features, but their improvement potential is limited by the ability of experts. Deep neural networks have been the dominant methodology in recent years, thanks to fast advances in deep learning techniques and hardware computing capacity [6], and deep convolutional neural networks (CNNs) have found tremendous success in CV. When massive labeled samples are available, CNN models can extract high-level and abstract feature representations and get an impressive performance on a variety of datasets [7].

However, most of the current networks are data driven and trained in a supervised way. Therefore, the performance of networks highly relies on massive labeled samples [8], which means that more large-scale datasets need to be created. Unfortunately, it is extremely lengthy and laborious for collecting numerous accurately labeled samples, especially precise pixel-level annotation [9]. Furthermore, labeling samples need certain professional knowledge, and some forms of data are difficult to obtain due to security or privacy considerations. In the domain of RS, although recent advances in sensors and earth observation techniques have given birth to explosive growth in RS data [10], a large number of labeled samples are still not available for some applications. For example, high-precision LULC data, which must be collected and annotated by RS experts, are difficult to obtain [11], [12], [13]. As a result, the widespread application and development of deep network technologies have been limited in the domain of RS to solve many practical problems due to the dearth of sufficiently large labeled datasets [2], [3], [4], [14], [15]. In this scenario, figuring out how to leverage the unlabeled data to improve the performance of existing models is a formidable challenge and the motivation.

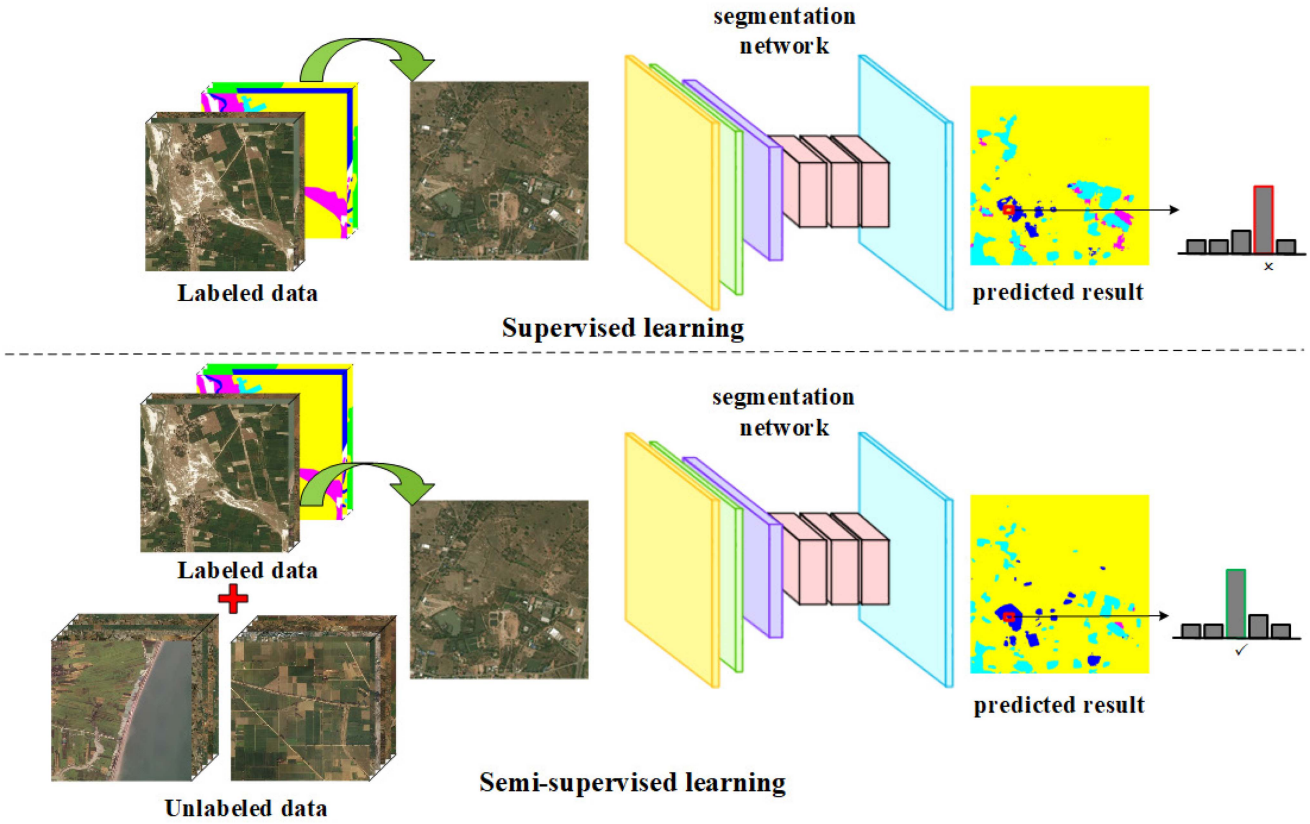



Fig. 1. Two types of ML methods: supervised learning and semisupervised learning. The probability value of each category after the softmax layer at the pixel position of the red frame in the outcome of the deep network prediction is shown by the histogram in the final column. In the supervised learning part, since only limited data are available, the classification result of the pixel in the red box is wrong. On the other hand, when more unlabeled data are supplied in the SSL section, the classification result of the pixel in the red box is correct. Combining massive unlabeled samples with limited labeled samples has been found to boost model performance [16].

Unlabeled data are relatively easy to obtain compared with labeled data. Thus, semisupervised learning (SSL) is a promising solution, which bridges the gap between unsupervised and supervised learning. The SSL method generally needs limited labeled samples and massive unlabeled samples to train a classifier [16], as shown in Fig. 1. It has been discovered that combining massive unlabeled examples with limited labeled samples considerably improves learning performance. And adding more unlabeled samples can yield more inherent information and determine more accurate decision boundaries. For supervised learning, it is costly and laborious, and it is difficult to obtain massive labeled samples. On the other hand, acquiring massive unlabeled samples is relatively inexpensive. Thus, SSL has more extensive application prospects in practical applications. Previous studies of SSL have demonstrated that consistency training by leveraging different image transformations on unlabeled data is an effective approach [17], [18], [19], [20], [21], [22].

Over the years, various semantic segmentation and SSL approaches have been developed. Fully convolutional networks (FCNs) have been widely used in RS images, such as FCN [23], SegNet [24], [25], UNet [26], and DeepLabv3 [27]. However, these methods can only perform well with a high amount of training data, and reducing the number of samples can greatly degrade the performance of the model. Some related works employing unlabeled data have been introduced in LULC, such as using multispectral data to boost the prediction performance of hyperspectral images [28] and leveraging scene-level labels for scene-level land-cover classification [29]. In addition, there are also some methods that use unlabeled data in hyperspectral image classification [30], [31], [32], SAR image classification [33], and change detection [34]. Equally important, this article focuses on the semisupervised semantic segmentation for high-resolution RS images, which is a critical topic that must be addressed.

In this study, we followed the spirit of consistency regularization widely used in SSL [19], [20] and proposed a new general framework of semisupervised semantic labeling for high-resolution RS images. Our framework employed a transformation consistency regularization (TCR) to make full use of the information provided by unlabeled samples, which enforced pixel-level consistency of the predictions by using different random transforms or perturbations. Unlike previous consistency regularisation methods, these methods mainly use a single simple image transformation as a perturbation, such as added noise, dropout, etc. Specifically, we explored three different high-level TCRs to compute the consistency loss and analyzed their performance. Then, we presented a semisupervised semantic labeling approach based on a hybrid transformation consistency

regularization (HTCR). The weighted sum of a supervised term calculated by labeled samples and a TCR loss computed by unlabeled data may be used to update the parameters of our network. On the DeepGlobe land-cover classification dataset [35] and the Inria Aerial Image Labeling dataset [36], our semisupervised segmentation framework outperformed state-of-the-art semisupervised approaches in our exhaustive experiments. This demonstrated that our semisupervised semantic segmentation approach has the potential to partially tackle the problem of limited labeled samples for land-cover classification of high-resolution RS images.

The following are the major contributions of our suggested method.

1) For limited labeled samples in high-resolution RS, we proposed a generic framework for semisupervised semantic segmentation. We employed a TCR in this framework to fully use the underlying information offered by unlabeled samples, which enforces the pixelwise consistency of the predictions through using various random transforms or perturbations.
2) We investigated three alternative transformations to compute consistency loss and analyzed their performance. We further proposed our HTCR-based semisupervised deep semantic labeling approach.
3) The experiments we present in this article show that the performance of semantic segmentation continuously improved as the number of unlabeled samples grew. We also demonstrated that our approach for semisupervised semantic labeling is a potential and promising method for addressing the issues of limited labeled samples for high-resolution RS images.

The rest of this article is organized as follows. In Section II, the recent related methods relevant to semantic segmentation and SSL are briefly presented. Our proposed method is discussed in Section III, and our experimental results are presented in Section IV. The performance of our proposed approach is discussed in Section V. Finally, Section VI concludes this article.

## II. Related Work

Recent work on semantic segmentation and SSL in the domains of CV and RS is reviewed in this section.

### A. Semantic Segmentation

In the domains of CV and RS, semantic segmentation has long been a question of great interest. For semantic segmentation, the FCN [37] was initially suggested. In pursuit of higher resolution features, encoder–decoder architecture [38] and skip connection were commonly used to recover the original size. To enhance the capability of capturing contextual information, dilated convolutions [39] and multiscale spatial pyramid pooling [40], [41] were proposed to generate more discriminative features. Recently, the self-attention mechanism [42] has become a research hot spot to extract more high-level visual features. In the RS area, Sherrah [23] proposed a network in which aerial images were used as the input for a pretrained VGG network, and digital surface model data were used for another FCN trained from scratch. The feature maps from these two networks were then concatenated to predict the label. Audebert et al. [24] proposed an encoder–decoder with a multikernel layer for fusing the predictions from multiple scales. Then, they developed a new network by using multimodal and multiscale RS data for land-cover classification [25]. Liu et al. [43] proposed a self-cascaded encoder–decoder network, which combined a coarse-to-fine refinement strategy to obtain fine predictions and a residual correction module to correct the latent fitting errors.

### B. Semisupervised Learning

SSL is a kind of ML technique that falls between supervised and unsupervised learning. Because of its capacity to incorporate labeled and unlabeled samples to train powerful models, SSL has become an exciting new research topic. With the rise of deep neural network technology, many novel SSL approaches have emerged [8]. In this subsection, we review only the SSL approaches related to deep learning, and for readers who want more details about traditional approaches, we refer readers to the literature [16].

The existing literature on SSL is extensive and focuses particularly on consistency regularization [17], [18], [19], [20], [21], [22]. In the image classification task, one of the most simple and effective methods introduced to deep neural networks is the pseudo-labeling method [17]. Its core idea was that the pseudo label of samples was generated by using the class having the maximum probability. The $\pi$-model method [18], which used the consistency regularization and input augmentation, exploited the stochastic of perturbation and penalized the difference of the results under random different augmentation to the input data [19]. The teacher–student framework was proposed to minimize the difference between the predictions [20]. The teacher network updated its parameters through the exponential moving average (EMA) of the student model parameters to obtain more stable predictions. Inspired by adversarial training, virtual adversarial training [21] did not exploit the randomness of the neural network but directly used a small noise to the input data, which can greatly change outputs of the model. Based on the mixup data augmentation technology [44], the interpolation consistency regularization method enforced consistent predictions at interpolations of unlabeled samples [22]. Recent studies (see, e.g., [45] and [46]) have shown that using both pseudo-labeling and consistency regularization achieves state-of-the-art performance on most classification datasets.

Although substantial efforts have recently been made in developing semisupervised approaches in classification problems with relatively small datasets, several of the current methods are not easily extended to real-world semantic labeling tasks. There has been an increasing amount of literature on semisupervised semantic labeling tasks. Hong et al. [47] proposed a decoupled CNN that exploited training data with image-level and pixel-level class labels to train classification and segmentation models, respectively. Souly et al. [48] proposed a method by using a generative adversarial neural network (GAN) architecture, whereby labeled samples were fed to a discriminator network to obtain class confidence score, and fake samples and unlabeled samples

were also input to the discriminator to obtain the confidence map of each class. Hung et al. [49] proposed a GAN model for semantic labeling, where the difference between their approach was that they designed a discriminator to discover reliable pixels of unlabeled samples that could promote the training process. Kalluri et al. [50] proposed a general semantic segmentation framework that can be trained on several datasets of different labels together. French et al. [51] proposed a consistency regularization using CutMix augmentation techniques. Mittal et al. [52] proposed to combine GAN and Mean Teacher in a complementary manner. Mondal et al. [53] proposed to learn a cycle-consistent regularization strategy between available labeled masks and unlabeled samples. Ke et al. [54] used two SSL constraints based on the flaw detector to learn from unlabeled data. Chen et al. [55] proposed cross consistency, which used two segmentation networks perturbed with different initialization to generate the pseudo labels to guide the other segmentation network.

### C. Semisupervised Deep Learning for Semantic Segmentation in RS

SSL also has been well explored in the RS field for automatic classification of RS images [56]. Before deep learning technology became mainstream, a novel modified transductive SVM was proposed to use unlabeled data for addressing ill-posed classification problems [57]. In recent years, a lot of research has begun looking into SSL methods on the basis of neural networks. Wu and Prasad [58] proposed a nonparametric Bayesian clustering algorithm to produce high-quality pseudo labels, resulting in improved initialization of the neural network. Zhang et al. [59] proposed a semisupervised change detection method that combined a novel multiscale feature and metric learning to strengthen the contribution of the training samples that are easy to classify and to weaken the contribution of training samples that are hard to classify. Han et al. [60] introduced a generative semisupervised solution consisting of a cotraining self-labeling strategy for learning a classifier from unlabeled samples and discriminative evaluation to enhance the classification of the confusion classes with similar texture structures and visualized features. Fang et al. [32] proposed a clustering algorithm to gather features of the network to produce pseudo targets for massive unlabeled samples in the collaborative learning framework. Hong et al. [61] used a cross-modal land-cover classification framework including three well-designed modules to carry more discriminative features from a hyperspectral image into the classification task using the multispectral data or SAR data. Recently, Zhang et al. [62] focused on semisupervised semantic labeling on the basis of consistency regularization, and their proposed method used unlabeled data to obtain promising results by encouraging the consistency of the output by using random transforms. For change detection in high-resolution RS images, Peng et al. [34] used a dual discriminator model, which enforced the consistency of features between segmentation maps and entropy maps. Kang et al. [63] proposed a SSL framework for building segmentation, which enforced pixelwise contrast and consistency constraints. Wang et al. [64] used iterative training to generate better pseudo labels to improve the segmentation performance.

Although the existing related works have made amazing progress, the majority of the available semisupervised approaches focused on change detection and hyperspectral image classification. Semisupervised semantic segmentation for high-resolution RS imagery has received insufficient attention, despite the fact that it is a significant and meaningful direction when dealing with realistic scenarios.

## III. Methods

The important principles of the SSL problem are initially described in this part, followed by an illustration of our TCR-based framework for semisupervised semantic labeling.

### A. Preliminaries

We consider a collection of training samples, which can be split to two folds: labeled samples $\mathcal{D}_L$ and unlabeled samples $\mathcal{D}_U$. The labeled data $\mathcal{D}_L$ consist of $N_L$ samples, denoted as $\mathcal{D}_L = \{(x_i^L, y_i^L)\}_{i=1}^{N_L}$, where $x_i^L \in X_L : \{x_1^L, \ldots, x_i^L, \ldots, x_{N_L}^L\}$ for all $i \in [N_L] : \{1, \ldots, N_L\}$, $y_i^L \in cardinal(C)$, and the number of categorizes is $C$. In the labeled data, each sample represents a tuple of input data $x$ and a target value $y$. All the available training samples have ground truth, and the aim is to learn a map in supervised training $f : X \to cardinal(C)$ parameterized by $\theta$ by optimizing a generic objective function

$$\mathcal{L}_s(X_L, Y_L; \theta) = \sum_{i=1}^{N_L} \ell_s(f_\theta(x_i^L), y_i^L) \tag{1}$$

where $f_\theta(\cdot)$ represents its model with parameters $\theta$ and the supervised term $\ell_s$ is commonly written as the cross-entropy (CE) function.

For SSL, one can access some labeled data $\mathcal{D}_L$ and unlabeled data $\mathcal{D}_U = \{x_i^U\}_{i=1}^{N_U}$, where the labels are unknown. The number of unlabeled samples is $N_U$, and a typical common assumption is that unlabeled samples outnumber labeled samples. The purpose of SSL is to obtain a function $f$ by leveraging all the available labeled and unlabeled samples, and the performance for the final model is superior to supervised learning trained on labeled data alone. Therefore, the objective function is commonly written as a weighted combination, that is, a supervised term $\mathcal{L}_s$ computed by labeled samples and a regularization term $R$ computed by unlabeled samples (or both):

$$\mathcal{L} = \mathcal{L}_s + \lambda R(\theta, \mathcal{D}_L, \mathcal{D}_U) \tag{2}$$

where $\lambda$ is a nonnegative hyperparameter.

Because only a tiny percentage of training samples are labeled, the regularization term is critical in determining how to employ unlabeled samples for training in order to improve the performance of the model. SSL must rely on some assumptions in order to make more accurate predictions while using unlabeled data. Smoothness assumption, cluster assumption, and manifold assumption are the three assumptions in general [16].

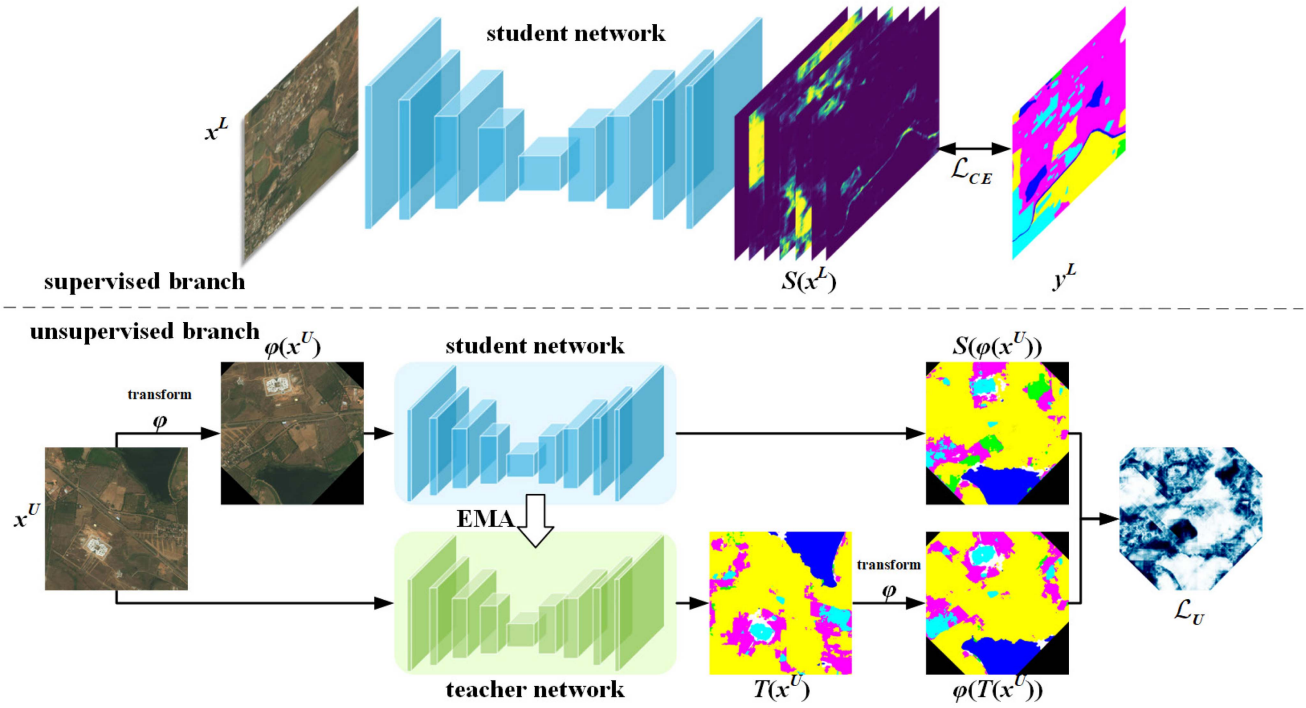



Fig. 2. Framework for semisupervised semantic segmentation. For the supervised branch, limited labeled images $x^L$ are fed into the student network that is trained by using CE loss. To leverage the information supplied by the unlabeled samples, we apply a TCR between the student and teacher networks in the unsupervised branch, which encourages consistent network predictions by utilizing distinct random transformations. The student network and teacher network are denoted as $\mathbf{S}$ and $\mathbf{T}$, respectively, where $\mathbf{S} = f_{\theta_s}$ and $\mathbf{T} = f_{\theta_t}$.

### B. Proposed Method

In this part, our general framework of semisupervised semantic segmentation for high-resolution RS images is described first.

In this framework, we use a TCR to utilize the information provided by unlabeled samples, which enforces the consistent predictions under different random perturbations. Specifically, three different transforms are used to compute consistency loss in accordance with the characteristics of the RS images. Then, we introduce our semisupervised semantic labeling method on the basis of an HTCR. The weighted sum of a supervised term and a regularization term can be used to update the parameters of the network.

Our proposed framework uses a teacher–student network, which has been widely used in network compression, knowledge distillation, and SSL. The network structure of the student and instructor networks is the same. The advantage of using a dual network is that aggregating parameters of the model throughout training stages leads to yield a more precise estimation, allowing for better unlabeled data target construction. The student network and the teacher network are denoted as $\mathbf{S}$ and $\mathbf{T}$, respectively. And their corresponding parameters are represented as $\theta_s$ and $\theta_t$, respectively. Since both labeled and unlabeled data are provided, the entire framework can be divided into two sections: the supervised branch and the unsupervised branch. The proposed framework is depicted in Fig. 2.

*1) Supervised Branch:* Limited labeled images $\mathcal{D}_L$ are fed into the student network $\mathbf{S}$ in the supervised branch, and the student network is trained in a supervised learning manner. Like other semantic segmentation methods, we use pixelwise CE as the loss function, denoted as $\mathcal{L}_{\text{CE}}$

$$\mathcal{L}_{\text{CE}}(X_L, Y_L; \theta_s) = -\sum_{h,w,c} y^L \log S(x^L). \tag{3}$$

*2) Unsupervised Branch:* For the unsupervised branch, we use a TCR between the student and teacher networks to exploit the information provided by the unlabeled samples, which encourages consistent network predictions by using different random transforms or perturbations. We obtain unlabeled data samples $\mathcal{D}_U = \{x_i^U\}_{i=1}^{N_U}$ and a random transform or perturbation $\varphi$. Then, we carry out a transform on the unlabeled samples $\tilde{x}^U = \varphi(x^U)$ and obtain the output feature map $f_{\theta_s}(\tilde{x}^U)$ by using disturbance samples as input to the student network. In the meantime, we obtain another output feature map $f_{\theta_t}(x^U)$ by feeding the original samples to the teacher network and performing the same perturbation on another output feature map $\varphi(f_{\theta_t}(x^U))$. Now, there are two output feature maps: $f_{\theta_s}(\varphi(x^U))$ and $\varphi(f_{\theta_t}(x^U))$. Finally, the discrepancy between the two output feature maps can be calculated by some distance function $d(\cdot,\cdot)$ using (4)

$$\mathcal{L}_u(X_U; \theta_s) = d(f_{\theta_s}(\varphi(x^U)), \varphi(f_{\theta_t}(x^U))) \tag{4}$$

where the distance function $d(\cdot,\cdot)$ can use mean squared error or the Kullback–Leibler (KL) divergence. In our article, we use mean squared error as the distance function.

(a) (b) (c)

Fig. 3. Examples for affine transformation, grid shuffle, and cutmix. (a) Affine transformation. (b) Grid shuffle. (c) Cutmix.

The optimization objective is to minimize this difference under random perturbations. Intuitively, the student network is encouraged to create the same feature maps as the teacher network. As the training continues, the performance of the teacher network consistently improves, and the prediction results of the student network become more reliable. Gradient descent would be used to adjust the parameters of the student model, while the EMA is used to tune the parameters of the teacher model. Since the weights of the teacher network average improve all layer features, the teacher model has better intermediate feature representations

$$\theta_t' = \alpha_{\text{EMA}}\theta_t + (1 - \alpha_{\text{EMA}})\theta_s \tag{5}$$

where $\alpha_{\text{EMA}}$ is a smoothing coefficient hyperparameter.

*a) Transformation in our framework:* Our framework uses some common operations in image processing as random disturbances, including pixel-level transformations and space-level transformations. Pixel-level transforms mainly include image filtering and image enhancement methods, such as blur, color jitter, adding noise, histogram equalization, etc. Spatial-level transforms mainly include flip, rotate, affine transform, perspective transform, elastic transformation, Cutout, CutMix, grid shuffle, etc. Since space-level transformations can cause higher dimensional disturbances, three spatial-level transformations are used in this article. Next, we introduce three effective transforms in our framework (see Fig. 3).

*Affine transformation:* Affine transformation is a type of 2-D geometric transformation used in CV, which translates pixels to new places utilizing a linear combining of translation, rotation, scaling, and shearing operations. As a result, the affine transformation is frequently employed to rectify alignment that has been twisted or deformed owing to camera flaws. Since convolution is not invariant to scaling and rotation, it has also been used recently for data augmentation during deep neural network training. The affine transformation is used here as a random perturbation on unlabeled data. In order to perturb sufficiently randomly, we use the translation in the range $(-0.2, 0.2)$ of the image size, zoom inside the range $(0.5, 1.5)$, and rotating in $(-180°, 180°)$.

*Grid shuffle:* Grid shuffle is a data augmentation method that splits an image into many small grid cells and shuffles them randomly to generate a new image, as shown in Fig. 3(b). This operation is often used to learn unsupervised feature representations by solving jigsaw puzzles [65]. Let $x \in \mathbb{R}^{W\times H\times C}$ denote an image and $P = \{x_1, x_2, \ldots, x_{M\times N}\}$ denote the set of small local patches. The grid shuffle operation first divides the image into $M \times N$ grids and randomly shuffles the positions of these small patches. The new image $x'$ is generated by merging these small patches. The grid shuffle operation is formulated as

$$\begin{aligned} P &= \{x_1, x_2, \ldots, x_{M\times N}\} = split(x) \\ x' &= merge(shuffle(P)). \end{aligned} \tag{6}$$

In practice, we cut an image into nine patches, shuffle them, and train the semantic network on the permuted images. This operation introduces a strong perturbation into the framework, allowing the model to acquire reliable feature embeddings from local image regions.

*Cutmix:* Cutmix is another data augmentation strategy and has outperformed other advanced data augmentation strategies on the ImageNet classification task. It cuts and pastes small local patches among training images according to a mask image to generate a new image [66], as shown in Fig. 3(c). Specifically, here, we use two sample images $x_A, x_B \in \mathbb{R}^{W\times H\times C}$ and their labels $y_A, y_B$. The cutmix operation is defined as

$$\begin{aligned} \tilde{x} &= (1 - \mathbf{M}) \odot x_A + \mathbf{M} \odot x_B \\ \tilde{y} &= (1 - \lambda)y_A + \lambda y_B \end{aligned} \tag{7}$$

where $\mathbf{M} \in \{0, 1\}^{W\times H}$ represents a binary array and $\odot$ is elementwise multiplication. The hyperparameter $\lambda$ is a combination ratio derived from the $\beta$-distribution $\beta(\alpha, \alpha)$. $\alpha$ is set to 1 in our implementation.

The binary mask $\mathbf{M}$ is sampled by obtaining the bounding box coordinates $\mathbf{B}$ from both the images, which indicates the area to be cropped from both images. The area in $x_A$ is filled with the small patch cropped from area in $x_B$. The aspect ratio for mask $\mathbf{M}$ is proportionate to the original image. The width and height of box $\mathbf{B}$ are computed according to the following formula:

$$\begin{aligned} r_w &= W\sqrt{1 - \lambda}, \text{where } r_x \sim \text{uniform}(0, W) \\ r_h &= H\sqrt{1 - \lambda}, \text{where } r_y \sim \text{uniform}(0, H) \end{aligned} \tag{8}$$

where $r_x$ and $r_y$ is width and height of the box, respectively. Finally, the binary mask $\mathbf{M}$ is determined by filling with zero in the bounding box $\mathbf{B}$; otherwise, it is 1. Two unlabeled images are required when this transformation is embedded in our framework. If the task is semantic segmentation, the cutmix operation

**Algorithm 1:** Training Process for TCR.

**Data:** labeled samples $\mathcal{D}_L = \{(x_i^L, y_i^L)\}_{i=1}^{N_L}$, unlabeled samples $\mathcal{D}_U = \{x_i^U\}_{i=1}^{N_U}$

**for** *i in [1, number of iterations]* **do**
- obtain a minibatch including labeled samples $x^L$ and unlabeled samples $x^U$
- /* supervised part */
- calculate CE loss $\mathcal{L}_s$ according to Equation (3)
- backward and update $\theta_s$ using optimizer
- /* semi-supervised part */
- get a random perturbation transformation $\varphi$
- carry out a random perturbation $\varphi$ on unlabeled samples $x^U$, that is $\tilde{x}^U = \varphi(x^U)$
- get the output feature map $f_{\theta_s}(\tilde{x}^U)$ by using perturbation samples $\tilde{x}^U$ as input to student network
- get the another output feature map $f_{\theta_t}(x^U)$ by using original samples $x^U$ as input to teacher network
- do same perturbation $\varphi$ on another output feature map $f_{\theta_t}(x^U)$, that is $\varphi(f_{\theta_t}(x^U))$
- calculate the transformation consistency regularization loss $\mathcal{L}_u$ using Equation (4)
- backward and update $\theta_s$ using optimizer
- update $\theta_t$ using EMA

**end**

for the label image is the same as the sample image

$$\tilde{y} = \mathbf{M} \odot y_A + (1 - \mathbf{M}) \odot y_B. \tag{9}$$

*b) Hybrid transformation consistency regularization:* When compared to single perturbation, such as the above three transformations, we believe that combining them can serve as more effective and diverse perturbations. Hence, we proposed an HTCR. Different combinations of the above three transformations were explored, and we verified their effectiveness on our datasets. When three transformations were embedded in our framework, the final training target $\mathcal{L}$ comprises four terms, i.e.,

$$\mathcal{L} = \mathcal{L}_s + \lambda_a \mathcal{L}_a + \lambda_{gs} \mathcal{L}_{gs} + \lambda_c \mathcal{L}_c \tag{10}$$

where $\mathcal{L}_a$, $\mathcal{L}_{gs}$, and $\mathcal{L}_c$ are the corresponding consistency regularization terms, and $\lambda_a$, $\lambda_{gs}$, and $\lambda_c$ are the weights that influence the size of the corresponding loss term.

In summary, the proposed entire framework of deep semantic semisupervised labeling can be described as Algorithm 1.

## IV. Experimental Results

### A. Experimental Dataset

We performed extensive experiments to verify our proposed semisupervised semantic segmentation approach on two datasets: the DeepGlobe Challenge land-cover classification dataset [35] and the Inria Aerial Image Labeling dataset [36].

*1) DeepGlobe Land-Cover Classification Dataset:* There are 1146 high-resolution RS images from rural areas in this dataset. The images have a pixel resolution of 50 cm and a shape of 2448 × 2448 pixels. All the images are separated into three parts: training set, validation set, and test set with 803, 171, and 172 images, respectively. As shown in Fig. 4, the ground truth is classified into seven categories: urban land, farm land, range land, forest land, water, barren land, and unknown (such as clouds). The unknown class is ignored when calculating accuracy. The DeepGlobe dataset is challenging because of the variety and diversity of its land-cover types. Because the ground truth for validation and the test set were not accessible, only the training set was used as experimental data. As a result, we picked 100 images at random as the validation dataset, 100 images as the test dataset, and the rest as the training dataset. In the training set, training samples were divided into labeled samples and unlabeled samples. Three distinct split methods were applied, with the labeled samples partitioned into 20, 50, and 100 images and the remaining training images regarded as unlabeled samples.

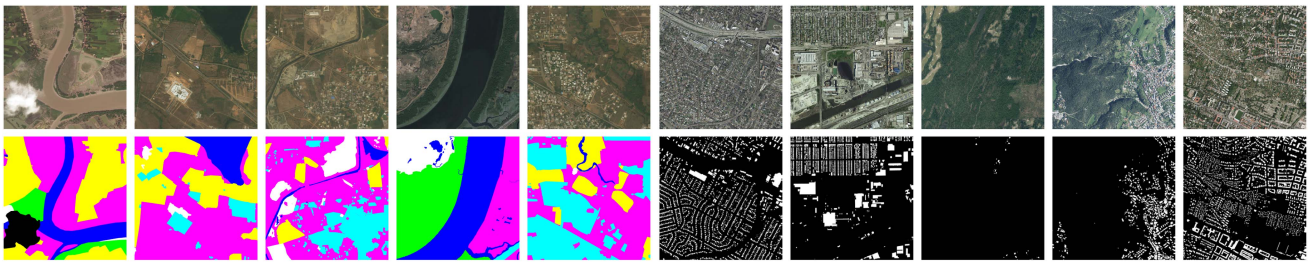
*2) Inria Aerial Image Labeling Dataset:* This benchmark dataset is composed of 360 orthorectified aerial photos, each of which is 5000 × 5000 in size. The data were acquired from ten cities over different urban areas. There are 36 images in each city. The ground-truth data were created by rasterizing the vectors of public domain building footprints. In our experiments, we used the public images from five cities as the experimental data because the ground truths for the other five cities were unavailable. As with the previous dataset, we randomly selected images from one city as the test dataset and divided them into 9, 18, and 36 images as the labeled data from the other four cities, and the remaining images were treated as unlabeled data.

### B. Implementation Details

We adopted the DeepLab-V2 [67] as the student and teacher network, and the ResNet-50 model pretrained on the ImageNet dataset [7] served as the backbone. Our model was trained using the Adam optimizer with a mini-batch size of 4. The learning rate was $3 \times 10^{-5}$ and the total iterations are 40 000. The random flip was employed for the data augmentation approach. the small patches were cropped to 513 × 513 in size finally. The weights of the regularization loss component $\mathcal{L}_a$, $\mathcal{L}_{gs}$, and $\mathcal{L}_c$ were set at 0.1, 1.0, and 1.0, respectively. The smoothing coefficient hyperparameter $\alpha_{\text{EMA}}$ was set at 0.99. In order to evaluate our method accurately, each experiment was conducted three times to calculate the mean values.

We compared our proposed method with the model trained using just labeled samples (denoted as baseline) and the model trained by all the labeled samples (denoted as full), which served as the lowest and upper bounds for supervised learning performance, respectively. To verify the effectiveness of our approach, we compared it to the following six state-of-the-art SSL algorithms.

1) *AdvSSL [49]:* This method proposed an adversarial model for semisupervised semantic labeling. For the labeled data, the higher structural information was learned from the adversarial loss. The trustworthy regions of the unlabeled samples generated by the discriminator model for the



(a) (b)

Fig. 4. Some examples of DeepGlobe land-cover classification dataset and Inria Aerial Image Labeling dataset. (a) Examples from the DeepGlobe land-cover classification dataset. (b) Examples from the Inria Aerial Image Labeling dataset.

unlabeled data were used to refine the results of the segmentation model.

2) *CutMix [51]:* This method was based on consistency regularization, which encouraged similar predictions under strong varied perturbations. Specifically, the authors proposed to mix the unlabeled samples and corresponding predictions and enforce consistency between the network outputs of the mixed unlabeled samples and the mixed predictions of unlabeled samples.
3) *s4GAN [52]:* This method combined GAN and Mean Teacher in a complementary manner to enhance the consistency of outputs. A segmentation network and a discriminator network were incorporated in the s4GAN branch, which may leverage unlabeled inputs to enhance the prediction quality. The image-level category labels were predicted from the Mean Teacher branch to filter the outputs of the S4GAN branch, which effectively removed incorrect predictions of the segmentation network.
4) *CCT [68]:* This method proposed adding different perturbations to the outputs of the encoder to enforce consistency over the main decoder and multiple auxiliary decoders. The shared encoder is enhanced by using the additional training loss calculated from multiple auxiliary decoders.
5) *CPS [55]:* This technique offered a simple semisupervised segmentation strategy that enforces the consistency between two networks with the same structure but different initialization, by supervising the other network with the one-hot pseudo segmentation map acquired from the first.
6) *S4Net [62]:* This method utilized unlabeled samples by consistency regularization, which ensures the pixelwise consistency of predictions under random various transformations.

The overall accuracy, mean intersection over union (mIoU), and F1-score were used as assessment metrics in this study. The F1-score is defined as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{12}$$

### C. Ablation Study

Ablation studies were performed on the DeepGlobe validation dataset to verify the performance of our proposed semisupervised semantic segmentation approach. The baseline results used 20 labeled images only. All of the semisupervised semantic segmentation results below were obtained using 20 images as labeled samples and the remaining images as unlabeled samples.

*1) Impact of the Hyperparameter λ on Accuracy:* In this part, the hyperparameters of consistency regularization loss $\lambda_a$, $\lambda_{gs}$, and $\lambda_c$ were analyzed. Because it was impossible to try all the possible values, six different values were used here: 0.1, 0.5, 1.0, 2.0, 5.0, and 10.0. The validation dataset was used to evaluate the results of the various values, and the settings and implementation details were the same as in the previous experiments. As shown in Fig. 5, the three random transforms improved the result remarkably. Compared with the baseline result, the results of using affine transformation as the random transformation were better than the baseline method. With an increase in $\lambda_a$, the performance displayed a downward trend, and this transform increased to 2.01 mIoU when $\lambda_a = 0.1$. Meanwhile, employing grid shuffle as the random perturbation outperformed the baseline with a large margin. When $\lambda_{gs} = 1.0$, the method yielded results of 76.74 in accuracy and 53.43 in mIoU. Likewise, the cutmix perturbation also achieved similar experimental results. When we set $\lambda_c = 1.0$, the performance improved to 52.60 mIoU. In the experiment below, if there was no additional explanation, we set $\lambda_a = 0.1$, $\lambda_{gs} = 1.0$, and $\lambda_c = 1.0$.

*2) Impact of Combination of Different Transform:* Compared with single perturbation, we contend that combinations of the above three transforms can serve as a more effective and diverse perturbation. To verify our proposed HTCR, we tried different combinations of the above three transforms. The results were shown in Table I. When using 20 labeled images only, the baseline method yielded 72.28, 46.36, and 59.56 in overall accuracy, mIoU, and mean F1-score, respectively. Furthermore, when we adopted single perturbation, the performance of using affine transform, grid shuffle, and cutmix improved to 48.37 mIoU, 53.43 mIoU, and 52.60 mIoU, respectively. When we combined two transforms, the combination of grid shuffle and cutmix resulted in 54.18 mIoU, but the performance increase

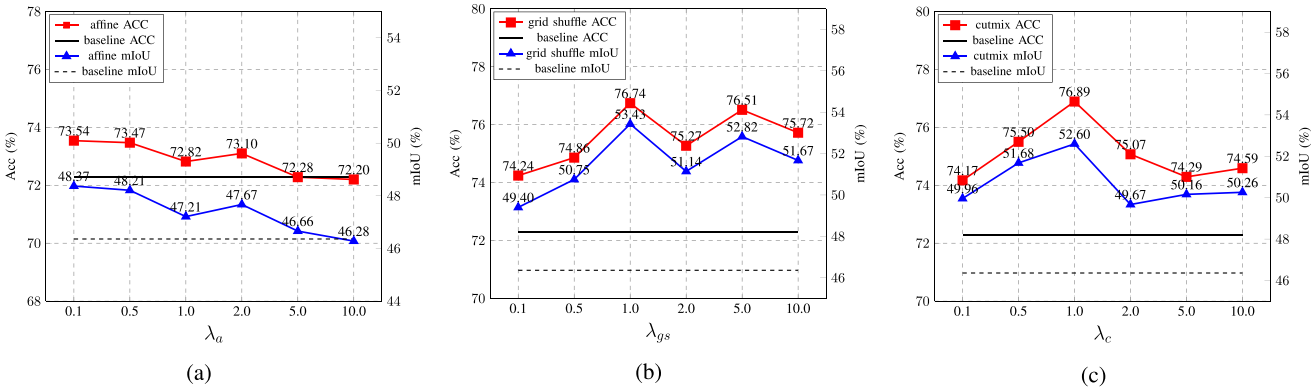



Fig. 5. Impact of the hyperparameter λ on accuracy for affine transform, grid shuffle, and cutmix. (a) Affine. (b) Grid shuffle. (c) Cutmix.

TABLE I
IMPACT OF COMBINATION OF DIFFERENT TRANSFORM ON ACCURACY

| affine | grid shuffle | cutmix | Acc | mIoU | mF1 |
|---|---|---|---|---|---|
| | | | 72.28±0.21 | 46.36±0.32 | 59.56±0.35 |
| ✓ | | | 73.54±0.33 | 48.37±0.16 | 61.74±0.40 |
| | ✓ | | 76.74±0.70 | 53.43±1.69 | 66.91±0.90 |
| | | ✓ | 76.89±0.95 | 52.60±1.32 | 66.44±1.36 |
| ✓ | ✓ | | 74.91±0.45 | 50.08±0.94 | 63.40±1.93 |
| ✓ | | ✓ | 73.89±0.14 | 49.23±0.30 | 62.45±0.18 |
| | ✓ | ✓ | **77.86±2.12** | **54.18±1.72** | **67.80±1.62** |
| ✓ | ✓ | ✓ | 74.26±1.85 | 49.48±2.80 | 62.73±3.15 |

The bold values denote the top-performing combination of different transform.

of the other two combinations was not apparent. Moreover, integrating all three transforms also did not significantly enhance the results, which indicated that affine transformation was not compatible with other complex transforms, and using multiple transforms at the same time did not always improve the results. However, the combination of grid shuffle and cutmix improved the baseline results remarkably with 7.82 mIoU. This shows that the combination of grid shuffle and cutmix can serve as a more diverse perturbation and is an effective way to apply consistency regularization in SSL. If not explicitly stated below, our method here forward uses a combination of these two transforms by default.

*3) Impact of the EMA Hyperparameter:* Sensitivity experiments were also conducted for EMA hyperparameters $\alpha_{\text{EMA}}$. The progressive improvement in performance is plainly seen in Fig. 6. From the graph above, we can see that the mIoU of the proposed method is 50.65 when $\alpha_{\text{EMA}}$ is equal to 0. At this point, the parameter of the teacher network is exactly the same as that of the student network, and the teacher network does not provide guidance to the student network. Therefore, the performance improvement mainly comes from the consistency of the transformation. Averaging the parameters of the model across the training phase, rather than utilizing the final weights directly, produces a more accurate model as $\alpha_{\text{EMA}}$ grows. It can be seen that the highest mIoU is reached when $\alpha_{\text{EMA}}$ is equal to 0.99. Therefore, in the experiments of this article, if not explicitly stated, we set $\alpha_{\text{EMA}} = 0.99$.

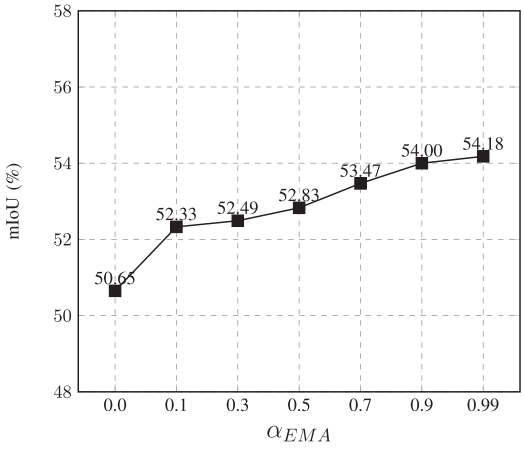



Fig. 6. Impact of different EMA hyperparameter values on performance.

TABLE II
IMPACT OF THE CONSISTENCY DISTANCE FUNCTION

| distance function | Acc | mIoU | mF1 |
|---|---|---|---|
| KL divergence | 77.53±0.59 | 53.35±1.08 | 66.97±1.05 |
| MSE | 77.86±2.12 | 54.18±1.72 | 67.80±1.62 |

*4) Impact of the Consistency Distance Function:* Table II compared the results obtained from the different consistency distance function. Here, we used two distance functions: KL divergence and MSE. When MSE was utilized as the distance function, the suggested approach produced the best results. Therefore, in the experiments of this article, if not explicitly stated, we used MSE as the consistency distance function.

### D. Comparison With State-of-the-Art Methods

On the test sets of the DeepGlobe dataset and the Inria dataset, we compared the proposed method to current semisupervised semantic labeling methods in this part.

*1) Results for DeepGlobe Dataset:* Following common practice, we compared the results of our proposed semisupervised

TABLE III
EXPERIMENTAL RESULTS FOR THE DEEPGLOBE LAND-COVER CLASSIFICATION DATASET

| Method | Number of labeled data | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 20 ($\frac{1}{30}$) | | | 50 ($\frac{1}{12}$) | | | 100 ($\frac{1}{6}$) | | | 603 (100%) | | |
| | Acc | mIoU | mF1 | Acc | mIoU | mF1 | Acc | mIoU | mF1 | Acc | mIoU | mF1 |
| AdvSSL [49] | 76.56±0.14 | 52.31±0.79 | 65.52±1.12 | 77.60±0.44 | 55.23±0.39 | 69.32±0.38 | 83.37±0.53 | 62.77±0.34 | 75.48±0.12 | 84.28±1.00 | 64.14±0.83 | 76.75±0.57 |
| CutMix [51] | 79.17±0.44 | 55.55±0.20 | 68.59±0.41 | 80.80±0.37 | 59.68±0.45 | 72.48±0.37 | 85.10±0.15 | 65.13±0.27 | 76.96±0.19 | 88.34±0.25 | 72.68±0.42 | 83.30±0.35 |
| s4GAN [52] | 77.32±0.55 | 53.97±0.33 | 67.68±0.14 | 79.95±0.49 | 58.05±0.50 | 71.61±0.12 | 83.83±0.27 | 63.25±0.54 | 75.95±0.45 | 85.61±0.23 | 67.70±0.68 | 79.77±0.49 |
| CCT [68] | 76.28±0.30 | 53.03±0.73 | 66.43±0.66 | 77.20±0.82 | 54.80±0.45 | 68.02±0.36 | 83.76±0.28 | 62.75±0.36 | 75.19±0.27 | 85.19±0.05 | 65.52±0.33 | 77.71±0.20 |
| CPS [55] | 76.47±0.39 | 52.25±0.31 | 65.43±0.18 | 79.94±0.29 | 57.85±0.61 | 70.59±0.88 | 84.76±0.09 | 64.84±0.09 | 76.94±0.10 | 86.44±0.55 | 69.28±0.10 | 80.89±0.12 |
| S4Net [62] | 76.75±0.61 | 53.28±0.78 | 67.13±0.73 | 78.31±0.56 | 54.77±1.32 | 69.03±1.03 | 84.68±0.15 | 64.26±0.40 | 76.41±0.41 | 87.47±0.15 | 71.06±0.27 | 82.18±0.23 |
| Baseline | 76.32±0.04 | 51.83±0.08 | 65.72±0.02 | 77.56±0.88 | 54.68±0.30 | 68.44±0.05 | 83.73±0.17 | 63.12±0.21 | 75.64±0.17 | 85.62±0.28 | 67.48±0.36 | 79.43±0.38 |
| HTCR | **80.78±0.31** | **57.81±0.49** | **70.76±0.29** | **81.57±0.5** | **60.56±0.43** | **73.19±0.25** | **85.38±0.26** | **65.86±0.30** | **77.50±0.33** | **88.67±0.05** | **72.95±0.22** | **83.44±0.21** |

The bold values denote the top-performing method for each of these metrics.

TABLE IV
EXPERIMENTAL RESULTS FOR THE INRIA AERIAL IMAGE LABELING DATASET

| Method | Number of labeled data | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 4 ($\frac{1}{36}$) | | | 9 ($\frac{1}{16}$) | | | 18 ($\frac{1}{8}$) | | | 144 (100%) | | |
| | Acc | IoU | F1 | Acc | IoU | F1 | Acc | IoU | F1 | Acc | IoU | F1 |
| AdvSSL [49] | 81.51±0.13 | 40.32±0.44 | 57.47±0.44 | 88.01±0.11 | 64.03±0.95 | 78.07±0.71 | 89.22±0.24 | 67.22±1.12 | 80.40±0.79 | 89.72±0.24 | 68.56±1.22 | 81.34±0.86 |
| CutMix [51] | 86.77±0.32 | 57.81±1.27 | 73.26±1.02 | 90.32±0.24 | 70.74±1.00 | 82.86±0.68 | 91.00±0.04 | 72.87±0.17 | 84.31±0.12 | **91.73±0.02** | **75.39±0.12** | **85.97±0.08** |
| s4GAN [52] | 82.68±1.71 | 49.56±1.12 | 66.01±1.25 | 87.74±0.22 | 65.31±0.97 | 79.02±0.70 | 89.34±0.29 | 69.30±0.72 | 82.86±1.42 | 90.33±0.07 | 71.89±0.51 | 83.64±0.35 |
| CCT [68] | 84.55±0.29 | 54.35±0.27 | 70.43±0.23 | 87.24±0.12 | 65.15±0.08 | 78.90±0.06 | 87.96±0.09 | 65.42±0.41 | 79.09±0.30 | 87.96±0.04 | 66.05±1.00 | 79.55±0.07 |
| CPS [55] | 77.13±0.15 | 25.37±1.89 | 40.44±2.41 | 85.94±0.41 | 57.48±0.56 | 73.00±0.45 | 88.61±0.09 | 66.03±0.16 | 79.54±0.11 | 89.26±0.11 | 68.07±0.45 | 81.00±0.32 |
| S4Net [62] | 86.39±0.16 | 61.23±0.66 | 75.95±0.51 | 89.28±0.10 | 68.99±0.54 | 81.65±0.38 | 90.23±0.26 | 71.46±0.10 | 83.35±0.06 | 91.06±0.03 | 73.94±0.12 | 85.02±0.08 |
| Baseline | 85.22±0.21 | 57.78±1.05 | 73.24±0.85 | 87.75±0.05 | 64.31±0.45 | 78.27±0.33 | 88.82±0.03 | 67.19±0.40 | 80.38±0.29 | 89.15±0.08 | 68.07±0.19 | 81.00±0.14 |
| HTCR | **87.98±0.22** | **62.30±1.11** | **76.77±0.84** | **90.98±0.04** | **72.69±0.10** | **84.18±0.07** | **91.14±0.04** | **73.26±0.08** | **84.57±0.05** | 91.65±0.06 | 75.19±0.12 | 85.84±0.08 |

semantic segmentation method HTCR to those of other methods in the case of four different data partitions. The results of our proposed method were superior than the baseline method by a considerable margin in four distinct scenarios, as shown in Table III. Especially, when there were only 20 labeled images, our method improved by 5.98 mIoU compared to the baseline method, while the existing methods AdvSSL and s4GAN have limited improvement. Both of these two methods were based on the GAN model. We believed that when labeled data were sparse, the GAN model lacked sufficient supervision signals, which led to unstable training. Although the CCT method employed consistency regularization in the higher dimensional space, its performance on RS datasets was insufficient. Compared with the CutMix method and S4Net, our proposed method combined multiple transformations, so it is a more diverse perturbation. Compared with the baseline method, the CPS method only improved by 0.42 mIoU, which may be because in the early stage of training, the prediction results of the two models were poor, and mutual supervision provided more invalid gradient updates. Furthermore, our results were 8.11% higher than the results of supervised learning when all the images were handled as labeled data and unlabeled data at the same time, proving the usefulness of our proposed HTCR. We also discovered that the results of our method using 20 labeled images were superior to the results of supervised learning using 50 labeled images, demonstrating that our SSL method is very effective when the number of labeled samples is small and that good results can be achieved only by adding unlabeled images.

Fig. 7 depicts a visual comparisons between the results provided by our proposed approach and the other methods. The graphic shows that our approach obtained better visual results than the other methods, demonstrating that the overall semantic content of an image can be better predicted by our model with fewer false detection compared to the other methods. In addition, our SSL method produced better details, integrity, and correctness, while the results of other methods often had checkerboard artifacts.

*2) Results for Inria Dataset:* To validate our method further, we also compared our method to the current methods on the Inria dataset. Using this dataset to extract buildings is more difficult than using the DeepGlobe dataset since it is a binary classification issue with significantly imbalanced positive and negative samples. When just a few annotated images were available, our results were superior to the other approaches, as shown in Table IV. For example, when there were only four labeled images, our accuracy was 4.52% higher than the results of supervised learning. With the exception of Cutmix and S4Net, the results of the existing methods were even worse than the baseline results. It is noteworthy that the CPS method obtained poor results when only a small number of labeled samples were available. When all the images were considered as annotated samples and unlabeled samples at the same time, our results were 7.12% higher than the results of supervised learning. Similarly, we also found that the results of our method using nine labeled images were better than the results of supervised learning using 18 labeled images, which also indicates that our method may be effective even when there is a limited amount of labeled data.

Fig. 8 depicts a visual comparison of the results generated by our method and the other methods. The ground truth and corresponding segmentation results for five original large images from the test set are displayed. Our method achieved better visual results compared with the other methods. Our results were better than other methods in both small dense buildings and large buildings, with fewer false detections, better integrity, and more accurate edges of the buildings, while CCT easily connected small dense buildings together and CPS basically gives incorrect classification results. To better show the details of the results, we

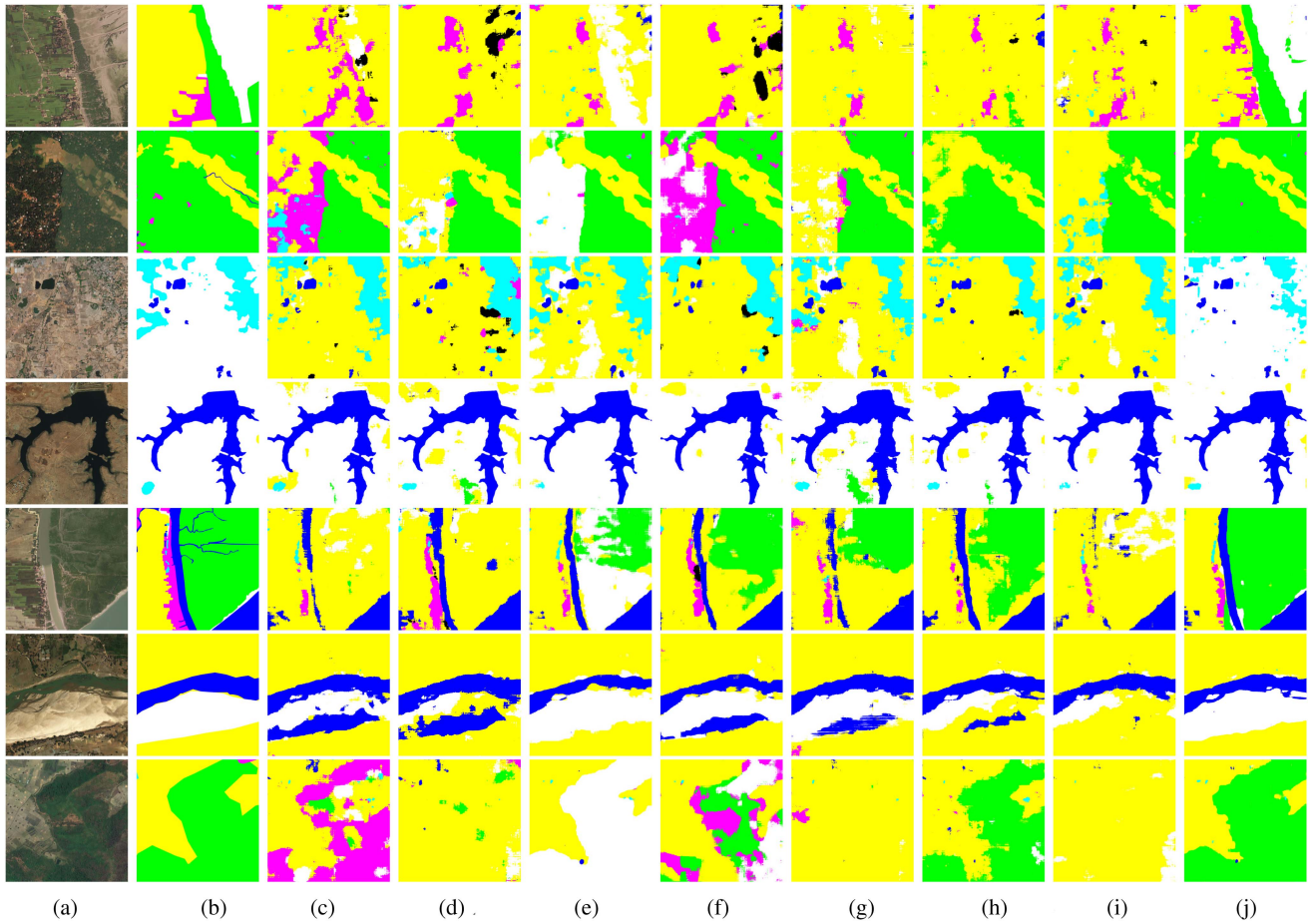



Fig. 7. Visual comparison of segmentation result on the DeepGlobe land-cover classification dataset. (a) Images. (b) GT. (c) Baseline. (d) AdvSSL. (e) CutMix. (f) s4GAN. (g) CCT. (h) CPS. (i) S4Net. (j) HTCR.

also compared the results of local patches (shown in red boxes) in the next row of each large image.

## V. Discussion

In this part, in order to further confirm the effectiveness of the proposed approach, we explored the sensitivity of the proposed method to different amounts of unlabeled data, the sensitivity of the semantic segmentation network used, and qualitatively visualized the assumptions contained in the proposed method. All these experiments were still performed on the DeepGlobe dataset, and all the semisupervised semantic segmentation results below were obtained using 20 images as labeled samples and the remaining images as unlabeled samples.

### A. Sensitivity Analysis of the Amount of Unlabeled Data

As can be observed from the preceding experiments, our approach outperformed the baseline model by a significant margin. However, previous experiments did not carefully study the influence of the number of unlabeled images on SSL. These experiments conducted in this subsection show that a key factor for the success or failure of SSL may be the number of unlabeled images. It is conceivable that accuracy was scarcely improved with only a few unlabeled images, while it was higher when the number of unlabeled images was large. We, therefore, further explored the impact of the amount of unlabeled data. In this experiment, we trained the baseline model using 20 labeled images. Fig. 9 depicts the model performance as a function of the amount of unlabeled samples, demonstrating that the accuracy of the proposed SSL method improves as the number of unlabeled images grows. Furthermore, when we used all the remaining images as unlabeled data, our model then achieved the highest performance. Therefore, we believe that using more unlabeled images likely will further improve accuracy.

### B. Robust Analysis of the Proposed Method

Because our framework does not rely on a specific network model, we replace the DeepLab-V2 with three current semantic segmentation networks (FCN, PSPNet, and DeepLab-V3) to confirm the stability of our proposed approach. Other segmentation networks in our architecture also improved accuracy, as seen in Table V. In particular, using DeepLab-V3 in our framework resulted in a mIoU that was 5.07% higher than the results of the supervised learning model, demonstrating the outstanding performance and robustness of our proposed approach.

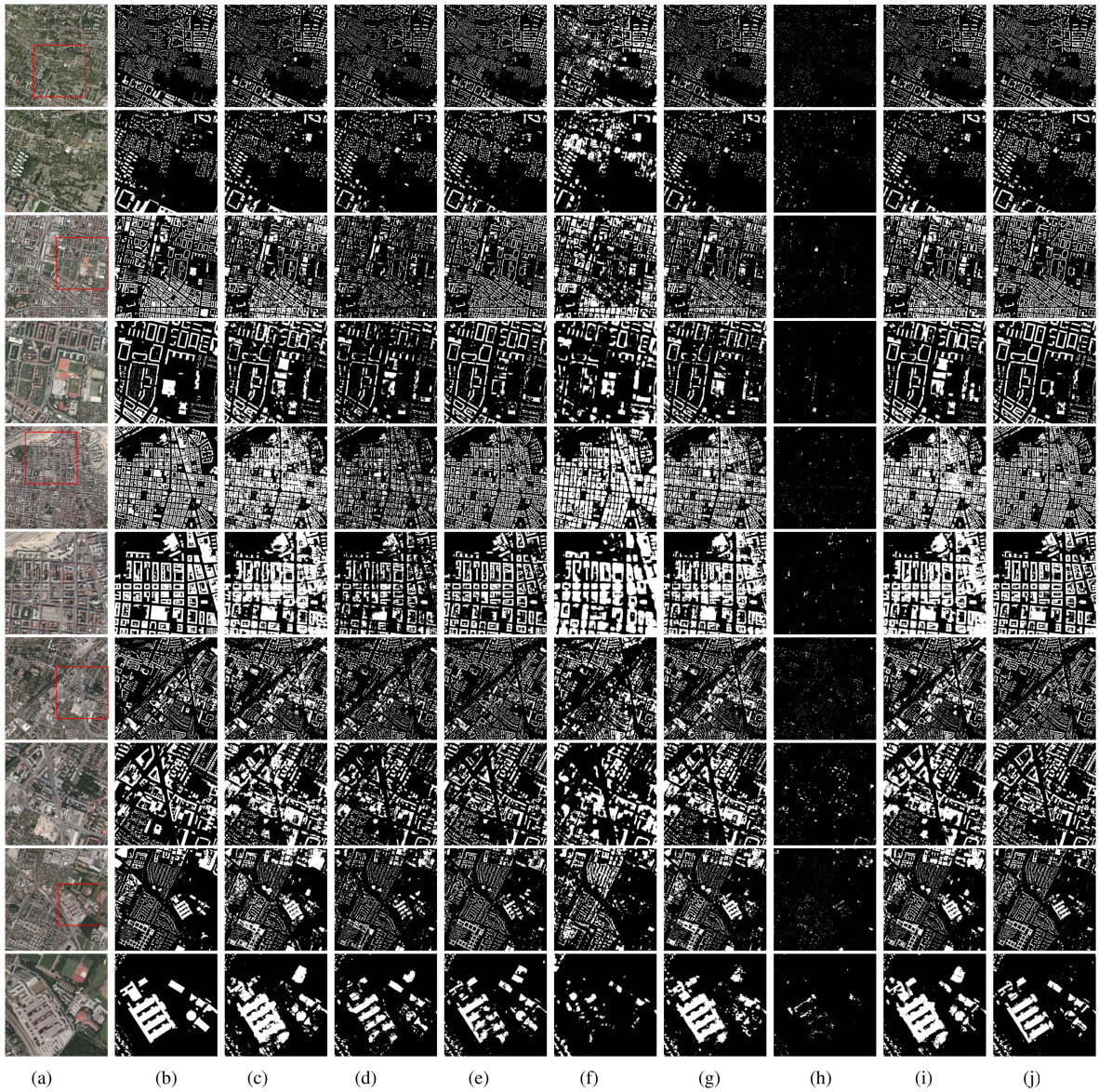



Fig. 8. Visual comparison of segmentation result on the Inria Aerial Image Labeling dataset. (a) Images. (b) GT. (c) Baseline. (d) AdvSSL. (e) CutMix. (f) s4GAN. (g) CCT. (h) CPS. (i) S4Net. (j) HTCR.

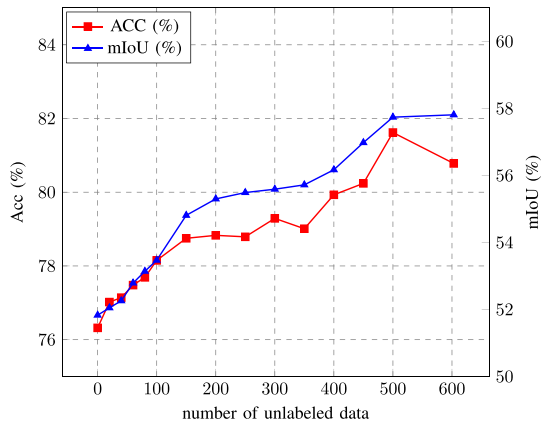



Fig. 9. Sensitivity of the number of unlabeled data on accuracy.

TABLE V
PERFORMANCE BY USING OTHER SEGMENTATION NETWORKS

| model | Acc | mIoU | mF1 |
|---|---|---|---|
| FCN (sup) | 76.89±0.20 | 52.42±0.22 | 66.61±0.21 |
| FCN (semi) | 77.78±0.30 | 54.75±0.59 | 68.27±0.48 |
| FCN(resnet50) (sup) | 77.28±0.18 | 52.76±0.62 | 66.67±0.41 |
| FCN(resnet50) (semi) | 78.16±0.13 | 55.22±0.32 | 68.63±0.34 |
| PSPNet (sup) | 78.77±0.25 | 55.95±0.36 | 69.30±0.14 |
| PSPNet (semi) | 79.83±0.54 | 57.83±0.75 | 70.72±0.59 |
| DeepLab V3 (sup) | 78.90±0.56 | 55.87±0.98 | 69.19±0.84 |
| DeepLab V3 (semi) | 80.37±0.15 | 58.70±0.12 | 71.50±0.20 |

### C. Assumption in SSL

When we compared our SSL method to supervised learning in previous trials, we obtained more precise predictions by considering the unlabeled samples. However, an important prerequisite

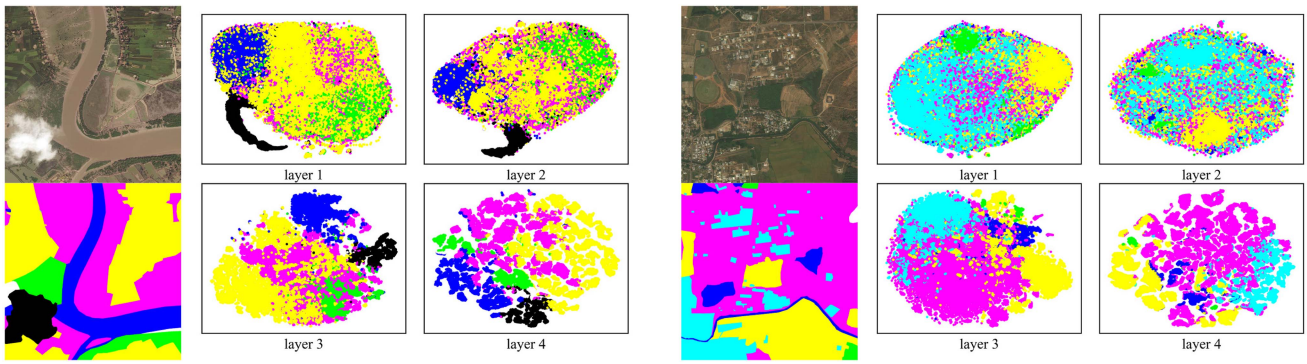



Fig. 10. Using t-SNE to visualize the features of the intermediate layer.

is that the unlabeled data must be able to help elucidate the distribution of examples [16]. There were two assumptions in this process. The first was the smoothness assumption, which states that if two points are close, then the corresponding outputs also should be close. The manifold assumption stipulates that high-dimensional data lie approximately on a low-dimensional manifold. Based on the above two assumptions, our method followed a modification of the smoothness assumption in deep learning, supposing that if two example points are close in a high-density region or low-dimensional manifold, then the two examples should have corresponding outputs. To validate this, we employed t-SNE [69] to plot four intermediate features of our deep network, as shown in Fig. 10. Our results show that the deeper features of the same category were closer compared to the shallow features. Every pixel in the image may be thought of as a point in a low-dimensional manifold, with the deep network attempting to learn its function by mapping data samples to high-dimensional features in the manifold. The distribution of the sample was not clearly and completely depicted when just a limited number of labeled samples were available. We employed a TCR to fully use the underlying information offered by unlabeled data, allowing the model to learn more precise decision boundaries by exploring a more realistic and accurate distribution of instances.

### *D. Difference Between the Proposed Method and Existing Methods*

The main differences between the proposed method and existing methods are as follows.

1) Compared with GAN-based methods, such as AdvSSL, s4GAN, etc., the proposed method used the teacher–student framework to avoid training instability due to adversarial training, and the model converged faster than GAN-based methods.
2) Compared with consistency-based methods, such as CutMix, CCT, CPS, etc., the proposed method explored three different transforms to compute consistency loss and analyzed their performance. To make perturbations more effective and diverse, we further proposed our semisupervised semantic segmentation method based on an HTCR. Different combinations of the above three transformations were explored, and we verified its effectiveness on our datasets. However, methods in CV do not perform optimally on RS data, especially the CPS method performs poorly on binary classification problems with imbalanced samples.
3) Although the existing related works have made amazing progress, most of the existing semisupervised methods focus on change detection and hyperspectral image classification. Insufficient research has been conducted on semisupervised semantic segmentation for high-resolution RS imagery, which is also an important and meaningful direction when facing realistic scenes. The proposed method has been tested on two high-resolution datasets to verify its effectiveness.

Compared with S4Net, although both the methods use consistent regularization, the previously proposed network is different from the one proposed in this method. The previous framework uses UNet as the segmentation network, and the labeled and unlabeled samples are input to the same segmentation network, and the weights of the two networks are the same, whereas in the framework proposed in this method, a dual network is used, i.e., a student network and a teacher network, and the weights of the two networks are different. In the previous framework, the unlabeled samples are transformed twice, and then, both are input to the network for consistency regularization comparison, while in the proposed framework, the unlabeled samples are transformed once and then both are input to the network with the original samples, then the same transformation is done on the output of the original samples, and finally the consistency regularization comparison is done. The previous framework, after getting the output of the network, needs to do the inverse transform to restore the position of each pixel first, which is troublesome when using some other transforms and makes the whole framework inflexible. The advantage of the proposed framework in this method is that only one transformation is done without inverse transformations, which can improve the flexibility of the framework by embedding arbitrary transformations.

## VI. Conclusion

In this article, a novel general framework was introduced for semisupervised semantic segmentation of high-resolution RS

images. In this framework, three different TCRs were used for encouraging consistent network predictions by using different random transforms or perturbations. Then, we proposed an HTCR method that combined different transforms. We conducted extensive experiments on the DeepGlobe dataset and the Inria dataset and found that our semisupervised semantic segmentation framework for high-resolution RS images outperformed other state-of-the-art semisupervised semantic segmentation approaches, especially when the labeled samples were scarce. Our experiments further demonstrated that our method for land-cover classification is promising to address the issue of scarce labeled samples. This feature of the method in this article is suitable for most RS image users. With the accumulation of annual or quarterly RS data, the method in this article can continuously improve the accuracy of the classifier without supplementing the labeled data.

A limitation of this study is that the two datasets used in this article are not dedicated to semisupervised RS semantic segmentation. More research using controlled trials is needed to use larger and especially designed for SSL dataset, such as MiniFrance dataset [70]. In spite of its limitations, the study certainly adds to our understanding of the SSL in RS. We believe future work can explore the following topics. It can embed more effective transforms into our framework to enforce pixel-level consistency. An interesting future line of research is using pseudo labels. It can use a model to generate pseudo labels, then use pseudo labels and real labels to train a new model, and iterate repeatedly. This is complementary to our method. In addition, unsupervised pretraining can also be used. It can train the model in an unsupervised fashion by using unlabeled data, such as self-supervised pretraining and fine-tuning it with labeled data.

## REFERENCES


[1] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Appl. Soft Comput.*, vol. 70, pp. 41–65, 2018.

[2] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[3] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[4] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, 2019.

[5] N. Zang, Y. Cao, Y. Wang, B. Huang, L. Zhang, and P. T. Mathiopoulos, "Land-use mapping for high spatial resolution remote sensing image via deep learning: A review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5372–5391, 2021.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

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

[8] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," 2020, *arXiv:2006.05278*.

[9] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-RNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5230–5238.

[10] Y. Ma et al., "Remote sensing big data computing: Challenges and opportunities," *Future Gener. Comput. Syst.*, vol. 51, pp. 47–60, 2015.

[11] Y. Li, Y. Zhang, and Z. Zhu, "Error-tolerant deep learning for remote sensing image scene classification," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 1756–1768, Apr. 2021.

[12] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 20–33, 2021.

[13] Y. Li, Z. Zhu, J.-G. Yu, and Y. Zhang, "Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10590–10603, Dec. 2021.

[14] F. Sharifzadeh, G. Akbarizadeh, and Y. Seifi Kavian, "Ship classification in SAR images using a new hybrid CNN–MLP classifier," *J. Indian Soc. Remote Sens.*, vol. 47, no. 4, pp. 551–562, 2019.

[15] N. Davari, G. Akbarizadeh, and E. Mashhour, "Corona detection and power equipment classification based on GoogleNet-AlexNet: An accurate and intelligent defect detection model based on deep learning for power distribution lines," *IEEE Trans. Power Del.*, vol. 37, no. 4, pp. 2766–2774, Aug. 2022.

[16] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning," *IEEE Trans. Neural Netw.*, vol. 20, no. 3, pp. 542–542, Mar. 2009.

[17] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Represent. Learn.*, 2013, Art. no. 2.

[18] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. 5th Int. Conf. Learn. Represent.*, 2017.

[19] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1163–1171.

[20] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[21] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2018.

[22] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 3635–3641.

[23] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*.

[24] N. Audebert, B. Le Saux, and S. Lefévre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 180–196.

[25] N. Audebert, B. Le Saux, and S. Lefévre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 140, pp. 20–32, 2018.

[26] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020.

[27] T.-S. Kuo, K.-S. Tseng, J.-W. Yan, Y.-C. Liu, and Y.-C. F. Wang, "Deep aggregation net for land cover classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 252–256.

[28] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 147, pp. 193–205, 2019.

[29] R. Zhu, L. Yan, N. Mo, and Y. Liu, "Semi-supervised center-based discriminative adversarial learning for cross-domain scene-level land-cover classification of aerial images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 155, pp. 72–89, 2019.

[30] G. Camps-Valls, T. V. B. Marsheva, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.

[31] S. Zhou, Z. Xue, and P. Du, "Semisupervised stacked autoencoder with cotraining for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3813–3826, Jun. 2019.

[32] B. Fang, Y. Li, H. Zhang, and J. C.-W. Chan, "Collaborative learning of lightweight convolutional neural network and deep clustering for hyperspectral image semi-supervised classification with limited training samples," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 164–178, 2020.

[33] H. Bi, J. Sun, and Z. Xu, "A graph-based semisupervised deep learning model for PoLSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2116–2132, Apr. 2019.

[34] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.

[35] I. Demir et al., "DeepGlobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 172–181.

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

[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.

[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Represent.*, 2015.

[40] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[41] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[42] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[43] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 78–95, 2018.

[44] H. Zhang, C. Moustapha, N. D. Yann, and L.-P. David, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[45] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 5050–5060.

[46] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 596–608.

[47] S. Hong, H. Noh, and B. Han, "Decoupled deep neural network for semi-supervised semantic segmentation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1495–1503.

[48] N. Souly, C. Spampinato, and M. Shah, "Semi supervised semantic segmentation using generative adversarial network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5688–5696.

[49] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2018.

[50] T. Kalluri, G. Varma, M. Chandraker, and C. Jawahar, "Universal semi-supervised semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5259–5270.

[51] G. French, T. Aila, S. Laine, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, high-dimensional perturbations," in *Proc. Brit. Mach. Vis. Conf.*, 2020.

[52] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-supervised semantic segmentation with high-and low-level consistency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1369–1379, Apr. 2021.

[53] A. K. Mondal, A. Agarwal, J. Dolz, and C. Desrosiers, "Revisiting cycle-GAN for semi-supervised segmentation," 2019, *arXiv:1908.11569*.

[54] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wisesemi-supervised learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 429–445.

[55] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2613–2622.

[56] L. Bruzzone and C. Persello, "Recent trends in classification of remote sensing data: Active and semisupervised machine learning paradigms," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2010, pp. 3720–3723.

[57] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3363–3373, Nov. 2006.

[58] H. Wu and S. Prasad, "Semi-supervised deep learning using pseudo labels for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1259–1270, Mar. 2017.

[59] W. Zhang, X. Lu, and X. Li, "A coarse-to-fine semi-supervised change detection for multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3587–3599, Jun. 2018.

[60] W. Han, R. Feng, L. Wang, and Y. Cheng, "A semi-supervised generative framework with deep learning features for high-resolution remote sensing image scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 23–43, 2018.

[61] D. Hong, N. Yokoya, G.-S. Xia, J. Chanussot, and X. X. Zhu, "X-ModalMet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 167, pp. 12–23, 2020.

[62] B. Zhang, Y. Zhang, Y. Li, Y. Wan, and F. Wen, "Semi-supervised semantic segmentation network via learning consistency for remote sensing land-cover classification," *ISPRS Ann. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 2, pp. 609–615, 2020.

[63] J. Kang, Z. Wang, R. Zhu, X. Sun, R. Fernandez-Beltran, and A. Plaza, "PiCoCo: Pixelwise contrast and consistency learning for semisupervised building footprint segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10548–10559, 2021.

[64] J.-X. Wang, S.-B. Chen, C. H. Ding, J. Tang, and B. Luo, "Semi-supervised semantic segmentation of remote sensing images with iterative contrastive network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2504005.

[65] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 69–84.

[66] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6023–6032.

[67] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[68] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12674–12684.

[69] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.

[70] J. Castillo-Navarro, B. Le Saux, A. Boulch, N. Audebert, and S. Lefévre, "Semi-supervised semantic segmentation in earth observation: The minifrance suite, dataset analysis and multi-task network study," *Mach. Learn.*, vol. 111, pp. 3125–3160, 2022.