Remote Sensing Image Semantic Change Detection Boosted by Semi-Supervised Contrastive Learning of Semantic Segmentation

Xiuwei Zhang, Yizhe Yang, Lingyan Ran, Liang Chen, Kangwei Wang, Lei Yu, Peng Wang, and Yanning Zhang, *Senior Member, IEEE*

***Abstract*— Semantic change detection (SCD) is a challenging task in remote sensing image (RSI) interpretation, which adopts multitemporal images to detect, locate, and analyze pixel-level land-cover "from-to" changes. In SCD, the severe class imbalance problem and the occurrence of confusing categories are very typical, making it challenging to accurately distinguish the easily confused categories with limited semantic context information. However, previous works did not address these issues in depth. This article proposes a novel SCD method named semi-supervised contrastive learning (SSCLNet), in which a simple and effective SCD network is designed as a strong baseline, and a semi-supervised contrastive learning module of semantic segmentation (SS) is presented to enhance the distinguishability of categories. Our baseline extracts semantic context through high-resolution network (HRNet), gets change information simply through an absolute difference, and then directly performs SCD based on the fusion of semantic context and change information. To utilize the semantic context information of the unlabeled non-changed regions, we employ a self-training (ST) method for semi-supervised SS. To learn distinguishable feature representations for easily confused categories, we present contrastive learning with an adaptive sampling strategy for SS. It selects challenging negative samples for each category from the other categories that exhibit similar features or attributes. The sampling space includes both the labeled changed samples and the non-changed samples predicted by ST. The comprehensive experiments on the SECOND and the Landsat-SCD dataset demonstrate that the proposed SSCLNet achieves the state-of-the-art (SOTA) performance, with a significant improvement of 2.07% and 4.15% in the score value, respectively.**

***Index Terms*— Contrastive learning, self-training (ST), semantic change detection (SCD).**

Manuscript received 9 May 2023; revised 14 September 2023, 23 December 2023, and 21 March 2024; accepted 17 April 2024. Date of publication 29 April 2024; date of current version 23 May 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC32093042023YFC3209305, in part by the National Natural Science Foundation of China under Grant 61971356U19B2037, in part by the Natural Science Foundation of Shaanxi Province under Grant 2021KWZ-03 and Grant 2022JQ-686, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2024JC-YBQN-0719, and in part by the Natural Science Foundation of Ningbo under Grant 2023J262. *(Corresponding author: Lingyan Ran.)*

Xiuwei Zhang, Yizhe Yang, Lingyan Ran, Kangwei Wang, Lei Yu, Peng Wang, and Yanning Zhang are with Shaanxi Provincial Key Laboratory of Speech and Image Information Processing and the National Engineering Laboratory for Integrated Aerospace-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: lran@nwpu.edu.cn).

Liang Chen is with the Information Center, Yellow River Conservancy Commission, Zhengzhou 450000, China.

Digital Object Identifier 10.1109/TGRS.2024.3395135

I. Introduction

IN THE field of remote sensing, semantic change detection (SCD) is the approach to monitor, locate, detect, and analyze semantic changes on the Earth's surface by using geographically co-registered multitemporal remote sensing images (RSI) [1], which is widely applied in urban planning, environmental monitoring, and disaster assessment [2], [3], [4].

Different from conventional change detection (CD), which only predicts binary pixel-level change/non-change labels, SCD detects the subtle changes in land-cover types on the Earth's surface [5]. SCD not only detects "where changes happen" but also predicts "how changes happen" in parallel by indicating change directions (e.g., "from land to building," "from vegetation to water," etc.). That makes it a crucial and challenging image interpretation task [6], [7], [8].

Fig. 1(a) illustrates the process of SCD. A pair of dual-temporal remote sensing images captured from the same geographic location serves as input. The SCD model subsequently predicts semantic change labels for each temporal remote sensing image. Notably, SCD extends beyond pixel-level CD to predict semantic labels for changed pixels in every phase, offering insights into evolving land-cover types.

The SCD task encounters serious issues related to class imbalance and class confusion. As depicted in Fig. 1(b), the proportion of unchanged regions significantly exceeds that of changed regions. Only the changed regions have semantic annotations corresponding to their land-cover categories. And several land-cover categories show high similarity, further compounding the challenge. Consequently, distinguishing these easily confused categories accurately becomes intricate due to the scarcity of semantic context information within changed regions. These issues ultimately hamper the overall performance of SCD.

Many scholars have devoted their efforts to the SCD topic and proposed excellent methods. Typically, they can be categorized as direct classification (DC)-based methods [9], [10], [11], [12], [13], [14], [15] and post-classification comparison (PCC)-based methods [16], [17], [18], [19]. The DC-based methods directly obtain semantic change maps using the original multitemporal remote sensing images. The PCC-based methods classify multitemporal remote sensing

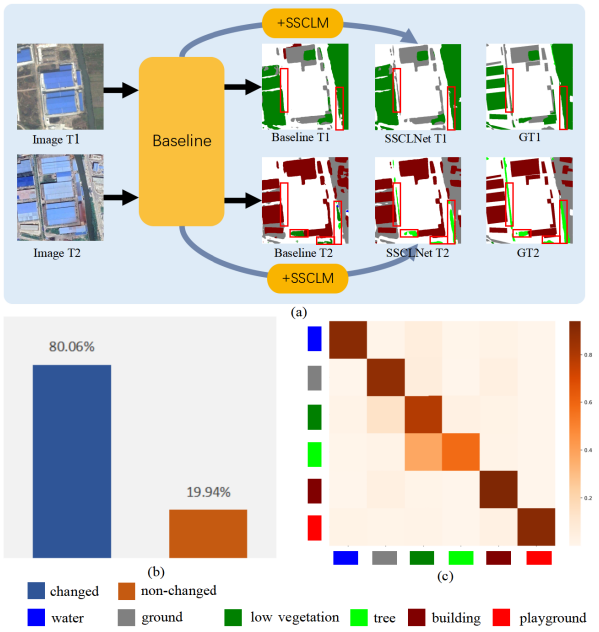

Fig. 1. Challenges of the SCD task (in SECOND dataset). (a) Process of the SCD task is illustrated and specific challenges addressed by our proposed SSCLNet (Baseline+SSCLM) are highlighted: accurately distinguish the easily confused categories with limited semantic context information. (b) Distribution of changed and non-changed categories in the dataset is depicted. Notably, the proportion of the unchanged category (80.06%) significantly outweighs that of changed categories (19.94%). (c) Similarity between land-cover classes within the changed categories is illustrated. The darker colors indicate a higher level of similarity.

images, respectively, to get classification maps, which are compared to obtain a "from-to" change map. However, traditional SCD methods cannot satisfy the requirement of finer feature extraction.

With the rapid development of deep learning technology and the availability of a large amount of multitemporal remote sensing images, deep learning-based SCD methods have achieved great progress. Compared with traditional SCD methods, their performance has been significantly improved. To simultaneously solve the problem of semantic segmentation (SS) and CD, most existing deep-learning-based methods focused on designing a proper network structure to effectively encode and integrate semantic context and difference information. As a preliminary attempt, a neural network-based approach for SCD was introduced in [20]. This approach employed a joint convolutional neural network (CNN)–recurrent neural network (RNN) architecture, where the CNN extracted semantic characteristics, and the RNN modeled temporal dependencies for multiclass CD. Daudt et al. [21] explored and compared four common deep learning-based SCD structures (e.g., direct comparison of SS maps, direct SCD (DSCD), separating CD and segmentation, as well as integrating CD and segmentation). SCDNet [22] adopted Siamese UNet to extract semantic context and CD information in the encoder stage and integrated them in the decoder stage and then directly performed SCD. ChangeMask [23] proposed a semantic-aware encoder to model the semantic-change causal relationship and a temporal-symmetric transformer (TST) to learn a robust change representation from semantic representation. Furthermore, current SCD datasets mainly consist of geo-referenced optical remote sensing images. Solutions for mitigating registration errors are proposed in [24] and [25], addressing the registration challenges not only in optical remote sensing images but also in multisource remote sensing images.

However, few works have focused on the problem of class imbalance and class confusion in SCD. To address the aforementioned class imbalance and confusion challenges, this article proposes a strong baseline for SCD tasks, enhanced with semi-supervised contrastive learning of SS. By adopting semi-supervised SS, our proposed method can effectively leverage the inherent semantic context of unchanged regions that lack explicit land-cover labels. Additionally, contrastive learning serves to heighten the distinguishability among different land-cover categories, thus contributing to the overall enhancement of the SCD.

The main contributions of this article are summarized as follows.

1) A simple and effective deep architecture is presented as a strong baseline for SCD tasks. The proposed baseline adopts HRNet40 [26] as an encoder to extract high-resolution semantic context. Then, it simply employs an absolute difference operation on two temporal semantic contexts to capture change information. At the decoding stage, semantic context and change information are integrated by a fusion operation to perform SCD directly. This simple structure outperforms most SCD methods.
2) To address the challenges posed by severe class imbalance and confusing categories in SCD, a semi-supervised contrastive learning module (SSCLM) of SS is proposed. In SSCLM, self-training (ST) is adopted for semi-supervised SS to predict pseudo-labels for the non-changed regions. Simultaneously, contrastive learning with an adaptive sampling strategy for SS is presented to enhance the discriminative capacity of semi-supervised contrastive learning (SSCLNet) in easily confused categories. The adaptive sampling strategy selects challenging negative samples for each category from the other categories that exhibit similar features or attributes. The sampling space is composed of high-quality pseudo-labels and ground-truth labels. By utilizing the high-quality pseudo-labels, the semantic context information of massive unlabeled non-changed image regions can be effectively encoded, which is valuable for scarcely changed categories. Moreover, the high-quality pseudo-labels enlarge the sampling space of contrastive learning. Coupled with the adaptive sampling strategy, SSCLM can effectively learn distinguishable feature representations of easily confused categories in changed regions.
3) To evaluate the effectiveness of SSCLNet, we conduct experiments on two large public datasets, namely SECOND [27] and Landsat-SCD [28]. The experimental results indicate that our baseline achieves the state-of-the-art (SOTA) performance. Furthermore, by integrating SSCLM into the baseline, SSCLNet exhibits a significant improvement in performance. Specifically, on the SECOND dataset, SSCLNet achieves a total score of 40.85%, representing a notable enhancement of

2.07%. Similarly, on the Landsat-SCD dataset, SSCLNet achieves a total score of 73.21%, manifesting a substantial improvement of 4.15%. Code will be available at https://github.com/YizheYoung/SSCLNet.

II. Related Works

A. Deep Learning-Based SCD

Mou et al. [20] proposed a joint CNN–RNN architecture, which was an early attempt to apply neural networks for SCD. The CNN extracted semantic features, while the RNN captured temporal dependencies for multiclass CD. Daudt et al. [21] proposed and compared four common deep learning-based SCD methods. Yang et al. [27] released a well-annotated benchmark dataset (SECOND) for SCD and proposed multiple evaluation metrics. Yang et al. [1] introduced an asymmetric siamese network named ASN for identifying and locating semantic changes. Peng et al. [22] proposed SCDNet, which adopted a Siamese UNet architecture to extract multiscale semantic context and difference information, and then integrated them in the decoder stage to perform SCD directly. Zheng et al. [23] introduced ChangeMask, which designed a semantic-aware encoder to leverage the semantic-change causal relationship sufficiently and proposed a TST to guarantee and leverage temporal symmetry. Zhu et al. [29] proposed a Siam-GL framework composed of an encoder–decoder-based Siamese network and a global hierarchical sampling mechanism to handle imbalanced training samples. It also incorporated a binary change mask to weaken the influence of the no-change regional background on the change regional foreground. Yuan et al. [28] proposed a potential transformer-based SCD model named PyramidSCD-Former, which precisely recognized the small changes and fine edges details of the changes. Ding et al. [30] presented bi-temporal semantic reasoning network (Bi-SRNet), which contained two types of semantic reasoning blocks to reason both single-temporal and cross-temporal semantic correlations. Ding et al. [31] improved SCD accuracy by integrating spatio-temporal dependencies via the development of a semantic change transformer (SCanFormer) and implementing a semantic learning scheme with coherent spatio-temporal constraints. Niu et al. [32] proposed symmetric multi-task network (SMNet), which integrated global and local information for accurate SCD. It used a pre-activated residual blocks and transformation blocks (PRTB) backbone unit to capture semantic features from bitemporal images and a multicontent fusion module (MCFM) to distinguish foreground and background information for enhanced change features.

Previous work primarily concentrated on investigating deep network architectures or designed loss functions to integrate semantic context and change information for SCD. However, limited attention has been given to the issues of class imbalance, category confusion, and the limited proportion of labeled pixels within the dataset, which are critical factors in SCD and are the primary focus of our study.

However, in the fields of CD and SS, some works have discussed and solved such problems, providing us with certain inspiration and inspiration. Hermann et al. [33] introduced the concept of few-shot filtering for CD in remote sensing, which enables the identification of specific types of changes using a relatively large CD dataset and a fine-tuning approach, addressing the challenge of limited labeled data for specialized change categories in remote sensing applications. Cho et al. [34] presented a novel annotation-free unsupervised SS framework, pixel-level feature clustering using invariance and equivariance (PiCIE), utilizing pixel-level clustering extended with geometric consistency without the need for hyperparameter tuning or task-specific pre-processing. Saha et al. [35] proposed a self-supervised SS method for large-scale Earth observation scenes, which sampled small, unlabeled patches from the scene, generates alternate views through simple transformations, processes them with a two-stream network, and refines weights iteratively using deep clustering, spatial consistency, and contrastive learning in the pixel space. Marsocci and Scardapane [36] presented a novel method called "continual Barlow twins," which merges self-supervised and continual learning for remote sensing. This approach combined Barlow twins' simplicity for self-supervision with elastic weight consolidation to mitigate catastrophic forgetting, addressing real-world Earth observation dataset challenges.

B. Semi-Supervised Learning

At present, almost all methods for SCD rely on supervised learning [22], [23], [30], which involves training a model using labeled training data and a supervised loss function such as cross-entropy. However, obtaining a large number of well-annotated training examples is expensive and time-consuming, especially for SCD, which requires both semantic annotations of dual-temporal remote sensing images and change annotations of image pairs. Furthermore, it is common for existing SCD datasets to have incomplete annotations, with only the changed regions receiving semantic annotations [1], [28]. This limitation hinders the attainment of optimal accuracy in SCD tasks, prompting us to consider semi-supervised learning as a solution to exploring available unlabeled data. In this work, we focus on semi-supervised SS.

In recent years, three primary approaches have emerged in semi-supervised learning: generative adversarial network (GAN)-based models [37], [38], [39], [40], consistency regularization [41], [42], [43], [44], [45], and entropy minimization [46], [47], [48].

GANs [40] are utilized as an auxiliary supervision signal for the unlabeled data in a way that the produced predictions are judged by the discriminator to imitate the common structures and semantic information of true segmentation masks [37], [38], [39]. However, GANs are not easy to optimize and may suffer the problem of mode collapse [49], [50].

Consistency regularization enforces the current optimized model to make stable and consistent predictions under various perturbations [42], [45]. Several studies [51], [52], [53], [54], [55], [56], [57], [58] have employed consistency training to enhance SS with promising results. In SCD tasks, Ding et al. [30] utilized dual-temporal consistency as extra supervision to guide semantic information exploitation and reduce dual-temporal result discrepancies.

Entropy minimization is an explicit and bootstrapping approach to leveraging unlabeled data which has been popularized by the ST strategy [46], [47]. In this method,

unlabeled data is assigned pseudo-labels and jointly trained with manually labeled data to minimize entropy [59]. The significance of ST in the SCD tasks is evident as it enables the prediction of pseudo-labels for unlabeled data, thus effectively utilizing all available data. Surprisingly, previous studies have not yet thoroughly explored this approach in SCD, which is precisely the focus of our work.

Although there is a limited exploration of semi-supervised methods in SCD within remote sensing imagery, the field of SS and CD for remote sensing images has witnessed significant research. Sun et al. [60] introduced the boundary-aware semi-supervised semantic segmentation network (BAS4Net), which utilized a discriminator network to infer pseudo-labels from unlabeled images, enhancing semi-supervised learning and improving segmentation network performance. Hong et al. [61] proposed the X-ModalNet, addressing the problem of semi-supervised transfer learning with limited cross-modality data in remote sensing. Protopapadakis et al. [62] presented a stack autoencoder-driven, semi-supervised deep neural network for extracting buildings from cost-effective satellite near-infrared images. These studies provide valuable references and inspiration for our work. Kondmann et al. [63] employed a semi-supervised framework for CD that combined unsupervised and supervised methods. This framework relied on half-sibling regression for optical CD as an unsupervised teacher model to generate pseudo-labels. Our approach used a supervised teacher model to predict pseudo-semantic labels specifically for unchanged areas, aiming to fully exploit the abundant semantic information in unchanged areas.

C. Contrastive Learning

Contrastive learning studies a similarity function to make the same data closer and push different data apart in representation space [64]. In this article, we focus on contrastive learning of SS in SCD, which has not been explored. There are a lot of excellent contrastive learning methods for SS with different design strategies, primarily on the choice of feature extraction and loss design. Alonso et al. [65], Li et al. [66], Wang et al. [67], and Hu et al. [68] focused on constructing high-quality and representative feature vectors for contrastive learning. Alonso et al. [65], Khosla et al. [69], and Zhao et al. [70] designed and used different loss to improve performance. For semi-supervised SS, Zhao et al. [70], [71] and Yang et al. [72] worked to improve the quality of pseudo-labels for ST using contrastive learning. Wang et al. [73] employed unreliable pseudo-labels as negative samples in contrastive learning to make sufficient use of unlabeled data. In addition, Zhao et al. [70] performed contrastive learning via pre-training by using a pixel-wise, label-based contrastive loss.

III. Method

A. SCD Problem Formulation

SCD is the task of predicting pixel-level annotations for image pairs captured at different times. Assume $\mathcal{X} = (X_1, X_2)$ represents dual-temporal input image pairs, where $X_i \in \mathbb{R}^{H\times W\times 3}$ and $i \in \{1, 2\}$. $\mathcal{Y} = (Y_1, Y_2)$ represents the corresponding pixel-level SCD labels with one non-changed class and L land-cover classes in changed regions, where $Y_i \in \mathbb{R}^{H\times W\times(L+1)}$ and $i \in \{1, 2\}$. The task is to design a neural network $\mathcal{F}$ with learnable parameters θ, which can be optimized to learn a mapping function $\mathcal{F}_\theta : \mathcal{X} \mapsto \mathcal{Y}$.

In particular, within the image pairs in the SCD dataset, only changed pixels $\mathcal{D}^c = \{(x_i, y_i)\}_{i=1}^{N^c}$ have semantic labels, where $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, and N^c represents the total number of changed pixels. y_i is the semantic label of x_i, and its value is $l \in L$, signifying the land-cover type in the current phase. Conversely, the non-changed pixels, termed $\mathcal{D}^u = \left\{\left(x_j, y_j\right)\right\}_{j=1}^{N^u}$, lack semantic labels. Here, $x_j \in \mathcal{X}$, $y_j \in \mathcal{Y}$, and N^u signifies the total number of non-changed pixels. y_j is the label of x_j, and its value is 0, indicating the continuity of land-cover type across phases. Notably, N^u significantly exceeds N^c in magnitude.

In this article, we employ ST in semi-supervised SS to predict the land-cover semantic labels for the non-changed pixels. The adopted ST scheme [73] has the teacher model $\mathcal{F}_t$ and student model $\mathcal{F}_s$ with the same architecture, and the teacher model $\mathcal{F}_t$'s parameters are obtained using an exponential moving average (EMA) of the student model $\mathcal{F}_s$. The ST process has three steps: first, train and update the student model $\mathcal{F}_s$ on the semantic labeled changed pixels $\mathcal{D}^c = \{(x_i, y_i)\}_{i=1}^{N^c}$. Second, update the teacher model's weights as an EMA of the student weights, and use the teacher model $\mathcal{F}_t$ to predict pseudo-semantic labels for non-changed pixels $\mathcal{D}^u = \left\{\left(x_j, \hat{y_j}\right)\right\}_{j=1}^{N^u}$. These predictions attribute the value $\hat{y_j}$ as $l \in L$. Last, retrain the student model $\mathcal{F}_s$ using $\mathcal{D}^c = \{(x_i, y_i)\}_{i=1}^{N^c}$ for supervised learning and the high-quality $\mathcal{D}^u = \left\{\left(x_j, \hat{y_j}\right)\right\}_{j=1}^{N^u}$ for unsupervised learning. Iterate the above three steps continuously until convergence. Details of the semi-supervised ST method are presented in Section III-C1.

B. Architecture of the SSCLNet

The architecture of the proposed SSCLNet is depicted in Fig. 2, which comprises a simple and effective SCD baseline incorporated with the SSCLM of SS.

The baseline of SSCLNet conducts SCD directly, and its structure is shown in Fig. 2(a). The baseline of SSCLNet consists of a dual-branch encoder–decoder structure with shared weights and a difference information extraction module in the middle part. Semantic context and difference information are fused at the decoding stage to perform SCD directly. The difference information extraction module only performs an absolute subtraction on semantic context feature maps of dual-temporal input images produced by our encoder. The fusion process consists of a channels-wise concatenation and a 1×1 convolution operation. In detail, the pre-trained HRNet40 [26] is adopted as an encoder expressed by ϕ to extract high-resolution finer semantic context. Pyramid scene parsing network (PSP) [74] decoder head is utilized as our decoder expressed by ψ_c to perform SCD prediction.

The SSCLM module performs semi-supervised contrastive learning for SS within the framework of SCD. We introduced SSCLM following each encoder branch, with both the encoder ϕ and SSCLM in two branches sharing weights. The structure of the encoder ϕ and SSCLM within a single branch is shown in Fig. 2(b). SSCLM incorporates an SS head (SS Head ψ_s)

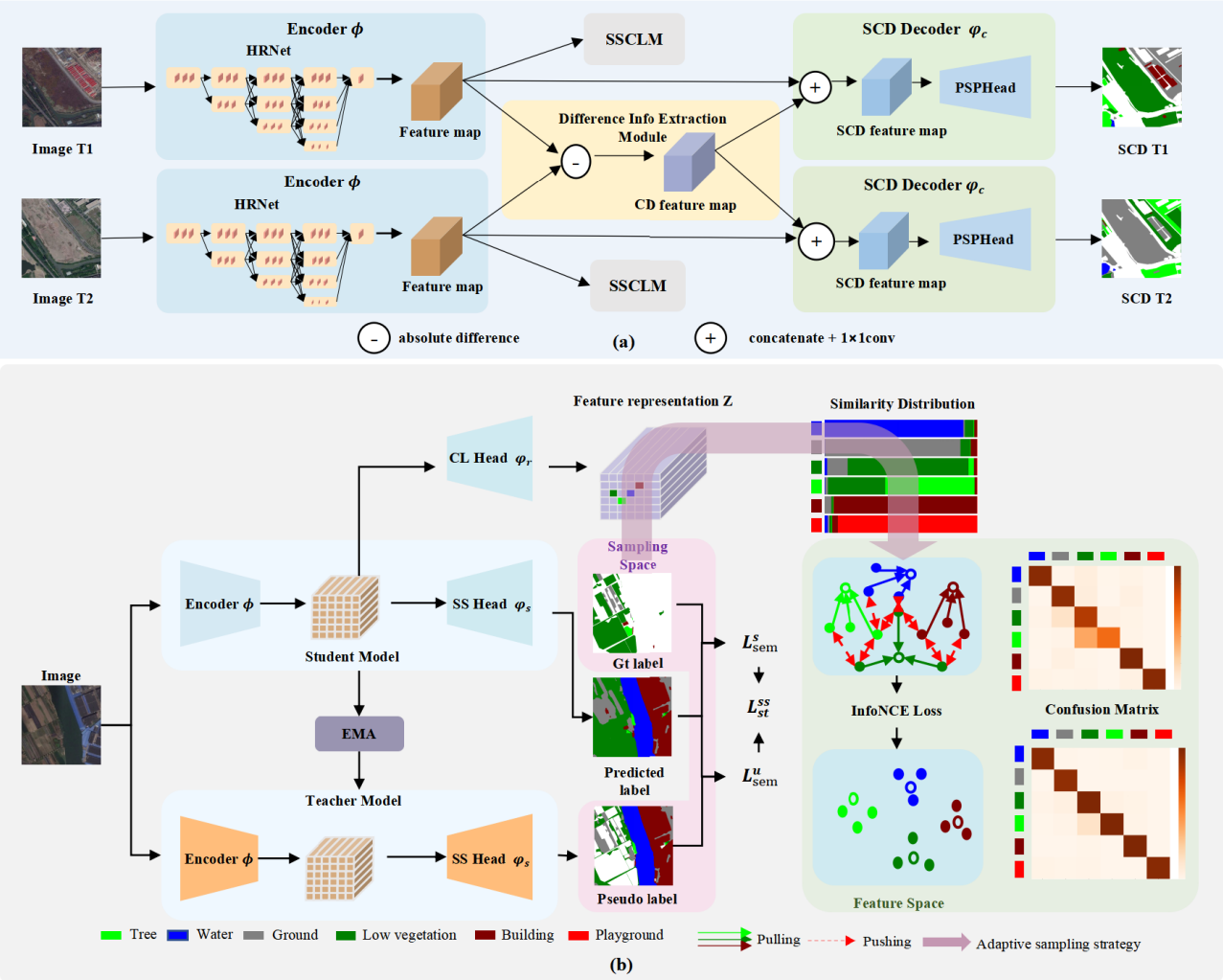

Fig. 2. Overview of SSCLNet. (a) Architecture of SSCLNet is presented, which comprises a simple and effective SCD baseline incorporated with the SSCLM of SS to improve the distinguishability of categories. The baseline consists of two encoder–decoder branches with sharing weights and a difference information extraction module for cross-image constraint. (b) Working principle of SSCLM is illustrated. The ST method is utilized in SS to predict semantic labels for the unlabeled non-changed category. In contrastive learning, the sampling space includes both labeled changed samples and predicted non-changed samples generated by ST. An adaptive sampling strategy is employed to select challenging negative samples for each category from other categories that exhibit similar features or attributes, according to the similarity distribution.

and a contrastive learning feature representation head (CL Head ψ_r) after the semantic context feature maps extracted by the encoder. The SS head ψ_s performs semi-supervised SS to predict land-cover labels for pixels in non-changed regions, and the CL head ψ_r generates a high-dimensional dense representation Z. The feature representations of positive and negative samples in contrastive learning are sampled from Z according to the ground-truth and pseudo-labels. SSCLM enhances the SSCLNet's capability to accurately distinguish land-cover types in changed regions, yielding a comprehensive improvement in SCD performance, as explained in Section III-C.

C. Semi-Supervised Contrastive Learning Module

We employ a semi-supervised contrastive learning module in SS within the framework of SCD to tackle the challenges posed by severe class imbalance and confusing categories in SCD. Specifically, in Section III-C1, we depict ST of SS to predict land-cover semantic labels for the non-changed class, which allows SSCLNet to learn from both semantically labeled and unlabeled data and provide sufficient semantic context information for scarcely land-cover categories in changed class. Then, in Section III-C2, we present contrastive learning of SS, and introduce an adaptive sampling strategy that can enhance the discriminative capacity of SSCLNet in land-cover categories in changed class. Finally, we exhibit the overall loss function in Section III-C3, which combines these techniques to optimize model performance.

1) Self-Training of Semantic Segmentation: There exists a serious class imbalance problem in the SCD dataset. For instance, in the SECOND dataset, the pixels in the non-changed regions make up over 80% of the total dataset, while the pixels in the changed regions with land-cover semantic labels account for less than 20%. Consequently, features of the land-cover categories in changed class with a lower proportion often cannot be learned adequately. Therefore, as illustrated in Fig. 2(b) (left), we conduct ST in semi-supervised SS to make full use of pixels without semantic labels in the non-changed regions.

For each of the two branches, we append the SS head ψ_s after the semantic context feature maps extracted by the encoder ϕ. ϕ and ψ_s form a subnetwork for ST in SS.

In particular, ψ_s of each branch share weights, and ψ_s has the same structure as that of the SCD decoder ψ_c.

To obtain reliable pseudo-labels, we use the entropy of the probability distribution predicted by the teacher model $\mathcal{F}_t$ to evaluate the quality of the pseudo-labels. The entropy is formulated as

$$\mathcal{H}(p_j) = -\sum_{l=0}^{L-1} p_j(l) \log p_j(l) \tag{1}$$

where $p_j \in \mathbb{R}^L$ is the softmax probability generated by the teacher model $\mathcal{F}_t$ for the jth unlabeled pixel and L is the number of land-cover categories. The lower the entropy, the higher the confidence of the pseudo-label. For every category, $l \in L$, rank the entropy of unlabeled pixels from low to high and select the top γ_t^l of them as reliable pseudo-labels at current training iteration t. Note that as training progresses, pseudo-labels become more and more reliable; thus γ_t^l should be dynamically adjusted with iterations. γ_t^l is a percentage defined as follows:

$$\gamma_t^l = \left\{ \alpha_e + (\alpha_0 - \alpha_e)\left(1 - \frac{t}{T}\right) \right\} P_{t-1}^l \tag{2}$$

$$P_{t-1}^l = \frac{\mathrm{TP}_{t-1}^l}{\mathrm{TP}_{t-1}^l + \mathrm{FP}_{t-1}^l} \tag{3}$$

where α_0 is the initial proportion set to 50% and α_e is the ending proportion set to 80%. T is the number of total iterations. TP and FP refer to the number of true positives and false positives in a mini-batch, respectively. P_{t-1}^l is the precision of class l evaluated in the latest iteration $t-1$, meaning the probability of positive samples among all samples classified as l. We start ST in the middle of the training epoch to avoid the accumulation of errors.

The ST loss $\mathcal{L}_{\text{sem}}^{\text{st}}$ for SS can be formulated as

$$\mathcal{L}_{\text{sem}}^{\text{st}} = \mathcal{L}_{\text{sem}}^{s} + \lambda_{\text{sem}}^{u} \mathcal{L}_{\text{sem}}^{u} \tag{4}$$

$$\lambda_{\text{sem}}^{u} = N_t^p / N_t^r \tag{5}$$

where $\mathcal{L}_{\text{sem}}^{s}$ and $\mathcal{L}_{\text{sem}}^{u}$ is a cross-entropy loss for supervised and unsupervised learning in ST. N_t^r is the number of reliable pixels, and N_t^p is the number of pseudo-pixels in a mini-batch at training iteration t. λ_{sem}^{u} is the weight of $\mathcal{L}_{\text{sem}}^{u}$ defined as the reciprocal of the percentage of reliable pixels among pseudo-labels.

2) Contrastive Learning of Semantic Segmentation: Contrastive learning is an effective technique that enables the model to acquire discriminative feature representations of categories presented in the dataset. This is achieved by minimizing the distance between pixels of the same category while simultaneously maximizing the separation between pixels of different categories in the representation space. We apply semi-supervised contrastive learning for SS within SCD.

To enhance the distinguishability of easily confused land-cover categories in changed regions, we propose an adaptive sampling strategy to select challenging negatives for each category from other categories that share similar features or attributes. Additionally, the sampling space of contrastive learning includes both the semantically labeled samples in changed regions and the semantically unlabeled samples predicted by ST in non-changed regions, to ensure that a sufficient number of samples can be obtained for rare categories.

In contrastive learning, the measured feature maps often need to be mapped into a high-dimensional dense representation via a representation head. Therefore, as visualized in Fig. 2(b) (left), for each of the two branches, we append a representation head ψ_r parallel to the SS head ψ_s. ψ_r of each branch share weights, and ψ_r has the same structure as that of the SCD decoder ψ_c. ψ_r maps the semantic context feature maps generated by the encoder ϕ into a high h-dimensional dense representation Z. The pixel-level info noise contrastive estimation (InfoNCE) [75] is utilized as our contrastive loss $\mathcal{L}_c$, which is defined as

$$\mathcal{L}_c = -\frac{1}{L \times M} \sum_{l=0}^{L-1} \sum_{m=1}^{M} \times \log \left[\frac{e^{\langle \mathbf{z}_{lm}, \mathbf{z}_l^+ \rangle / \tau}}{e^{\langle \mathbf{z}_{lm}, \mathbf{z}_l^+ \rangle / \tau} + \sum_{n=1}^{N} e^{\langle \mathbf{z}_{lm}, \mathbf{z}_{lmn}^- \rangle / \tau}} \right] \tag{6}$$

where $\mathbf{z}_{lm}$ is the representation of the mth anchor of class l, $\mathbf{z}_l^+$ is the mean representation averaged across all representations of class l, and $\mathbf{z}_{lmn}^-$ is the representation of the mth anchor's nth negative sample. All of them are sampled from dense representation Z. Each category has M anchors, and each anchor is followed with a positive sample $\mathbf{z}_l^+$, and N negative samples $\mathbf{z}_{lmn}^-$. $\langle \cdot, \cdot \rangle$ is the cosine similarity between feature representations from two different samples, with the range of -1 to 1. τ is the temperature control of the softness of the distribution. For each category l in the current mini-batch, $\mathcal{L}_c$ encourages anchor $\mathbf{z}_{lm}$ to be similar to the positive sample $\mathbf{z}_l^+$ and dissimilar to the negative samples $\mathbf{z}_{lmn}^-$. In particular, we select feature representations of samples in a mini-batch, allowing contrastive learning to learn not only from local context (neighboring pixels) but also from other images in a mini-batch.

We have observed that the SCD dataset often encounters confusion between categories that have similar semantic relationships (e.g., low vegetation and tree in the SECOND dataset). Therefore, it is crucial to employ an adaptive sampling strategy in contrastive learning to enhance the discriminative capability of SSCLNet when dealing with these challenging categories.

Contrastive learning improves the discriminability of categories by creating a clear separation between anchors and their corresponding negatives. Building upon this observation, the adaptive sampling strategy is specifically designed to carefully select challenging negatives for each category from other categories that share similar features or attributes. The proposed adaptive sampling strategy effectively sharpens the classification boundary between confusing categories, ultimately enhancing their distinguishability.

The adaptive sampling strategy's sampling principle is introduced as follows.

In a mini-batch, confusion always happens between classes with high similarity. $\mathcal{D}_i$ is the similarity distribution of class i, which is computed and dynamically updated for each mini-batch. $\mathcal{D}_i$ is defined as follows:

$$\mathcal{D}_i = \{ d_{ij}, \quad \forall j \in L, \text{ and } j \neq i \} \tag{7}$$

where d_{ij} is the similarity between class i and j, which is formulated as

$$d_{ij} = \frac{e^{\left\langle \mathbf{z}_i^+, \mathbf{z}_j^+ \right\rangle}}{\sum_{j=1, j\neq i}^{L} e^{\left\langle \mathbf{z}_i^+, \mathbf{z}_j^+ \right\rangle}} \tag{8}$$

where $\mathbf{z}_i^+$ and $\mathbf{z}_j^+$ are positive samples representing the mean feature for class i and j. $\langle \cdot, \cdot \rangle$ is cosine similarity between the positives of two classes. d_{ij} is similarity normalized after softmax between class i and its negative class j.

Based on the similarity distribution $\mathcal{D}_i$, the adaptive sampling strategy extracts N negatives for class i from its negative classes non-uniformly according to its sampling distribution N_i. N_i is defined as follows:

$$\mathcal{N}_i = N \times \mathcal{D}_i. \tag{9}$$

With the adaptive sampling strategy, contrastive learning can help the SSCLNet to learn a more accurate decision boundary in SS.

The sampling space of contrastive learning includes both the labeled changed samples and the non-changed samples predicted by ST. This sufficient sampling space is particularly significant for addressing the scarcity of samples in rare categories such as water and playground in the SECOND dataset, as well as farmland and buildings in the Landsat-SCD dataset.

3) Overall Loss: The overall training loss of SSCLNet with semi-supervised contrastive learning can be formulated as

$$\mathcal{L} = \lambda_{\text{scd}}\mathcal{L}_{\text{scd}} + \lambda_{\text{st}}^{\text{ss}}\mathcal{L}_{\text{st}}^{\text{ss}} + \lambda_{\text{cl}}^{\text{ss}}\mathcal{L}_{\text{cl}}^{\text{ss}} \tag{10}$$

where $\mathcal{L}_{\text{scd}}$ is cross-entropy loss for SCD task with the weight of λ_{scd}. $\mathcal{L}_{\text{st}}^{\text{ss}}$ is ST loss for SS described in (4) with the weight of $\lambda_{\text{st}}^{\text{ss}}$. $\mathcal{L}_{\text{cl}}^{\text{ss}}$ is contrastive loss described in (6) with the weight of $\lambda_{\text{cl}}^{\text{ss}}$.

IV. Experiments and Analysis

A. Experiments Settings

1) Dataset Description: To evaluate the effectiveness of the proposed method in diverse scenarios, we conduct experiments on two publicly accessible benchmark datasets in SCD, namely the SECOND dataset and the Landsat-SCD dataset. The SECOND dataset is a high-resolution dataset collected in urban areas. The Landsat-SCD dataset is a mid-resolution dataset located at the margin of a desert area.

a) SECOND dataset: SECOND dataset [1] is a publicly available high-resolution SCD dataset including 2968 image pairs, which is obtained in various cities including Hangzhou, Chengdu, and Shanghai in China. Each image has 512 × 512 pixels with the spatial resolution varying from 0.3 to 5 m. SECOND dataset holds seven categories, including one non-changed category and six land-cover categories for changed regions, i.e., water, ground, low vegetation, tree, building, and playground. There exists a serious class imbalance problem between the changed and the non-changed categories. During training, we randomly split the SECOND dataset into the training set and the testing set with a ratio of 9:1.

b) Landsat-SCD dataset: Landsat-SCD dataset [28] is a mid-resolution SCD dataset including 8468 image pairs, which is constructed with Landsat images between the years 1990 and 2020. The observed region is in Tumushuke, Xinjiang, China, which is at the margin of the Taklimakan Desert. Each image has 416 × 416 pixels with a spatial resolution of 30 m. The Landsat-SCD dataset comprises five categories, including one non-changed category and four land-cover categories for changed regions, i.e., farmland, desert, building, and water. Changed pixels account for only 18.89% of the total pixels, making it a challenging dataset to work with. We further split them into training and testing sets with a ratio of 9:1.

2) Evaluation Metrics: Similar with [22] and [27], we adopt mean intersection over union (mIoU) as the evaluation metric for CD, separate Kappa (SeK) [27] coefficient as the metric for SCD, and F1-score as the evaluation metric for each semantic category.

mIoU is the average of IoU between non-changed (IoU_1) and changed classes (IoU_2), which is formulated as

$$\text{mIoU} = 0.5 \times (\text{IoU}_1 + \text{IoU}_2) \tag{11}$$

where IoU_1 and IoU_2 are the IoU of non-changed and changed classes, which can be computed as follows:

$$\text{IoU}_1 = \frac{\text{TN}}{\text{TN} + \text{FP} + \text{FN}} \tag{12}$$

$$\text{IoU}_2 = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{13}$$

where TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives in the CD confusion matrix.

SeK is calculated by the IoU_2 of changed classes and a novel Kappa, which is not affected by the IoU_1 of the non-changed class. SeK is defined as

$$\text{SeK} = e^{\text{IoU}_2 - 1} \cdot \frac{\widehat{\rho} - \widehat{\eta}}{1 - \widehat{\eta}} \tag{14}$$

with

$$\widehat{\rho} = \frac{\sum_{i=2}^{C} Q_{ii}}{\sum_{i=1}^{C}\sum_{i=1}^{C} Q_{ij} - Q_{11}} \tag{15}$$

$$\widehat{\eta} = \frac{\sum_{j=1}^{c} \widehat{Q}_{j+} \cdot \widehat{Q}_{+j}}{\left(\sum_{i=1}^{c}\sum_{i=1}^{c} Q_{ij} - Q_{11}\right)^2} \tag{16}$$

where $Q \in \mathbb{R}^{C\times C}$ is the multiclass confusion matrix, and C is the label category. Q_{11} is the number of true positive pixels in the non-changed class. $\widehat{Q}_{j+}$ and $\widehat{Q}_{+j}$ denote the sum over row and column without Q_{11} on confusion matrix, respectively.

The comprehensive score is calculated as (higher values for better model performance)

$$\text{Score} = 0.3 \times \text{mIoU} + 0.7 \times \text{SeK}. \tag{17}$$

3) Comparative Methods: In comparison experiments, we comprehensively compare six deep learning-based SCD methods.

1) *ResNet-GRU [20]:* The recurrent CNN with the gated recurrent unit (GRU).

TABLE I
COMPARATIVE RESULTS OF DIFFERENT METHODS (RED: THE BEST; BLUE: THE SECOND BEST)

Methods	SECOND			Landsat-SCD		
	mIoU(%)	Sek(%)	Score(%)	mIoU(%)	Sek(%)	Score(%)
ResNet-GRU [20]	60.64	8.99	24.49	74.16	26.51	40.81
ResNet-LSTM [20]	67.27	16.14	31.48	80.88	40.06	52.31
DSCD [21]	62.45	10.20	25.88	74.92	2.89	24.50
SCDS [21]	69.18	14.96	31.22	78.33	31.43	45.50
ICDS [21]	71.95	21.83	36.86	79.10	32.29	46.33
HBSCD	72.40	21.46	36.74	89.82	58.50	67.90
SCDNet [22]	73.06	23.66	38.59	85.23	50.05	60.60
ChangeMask [23]	-	17.89	-	-	-	-
Bi-SRNet [30]	73.41	23.22	36.74	85.53	51.01	61.37
SCanNet [31]	73.42	23.94	38.78	88.96	60.53	69.06
SMNet [32]	71.95	20.29	35.79	85.65	51.14	61.49
Baseline	72.37	24.05	38.55	90.36	62.48	70.84
SSCLNet	74.10	26.59	40.85	91.94	65.41	73.21

2) *ResNet-Long Short Term Memory (LSTM) [20]:* The recurrent CNN with LSTM.
3) *DSCD [21]:* DSCD assigned independent labels to each "from-to" type of semantic change viewing SCD as a simple SS task.
4) *Separate CD and segmentation (SCDS) [21]:* SCDS decoupled SCD into CD and SS. It trained two networks for CD and SS, respectively, and integrated their results finally.
5) *Integrate CD and segmentation (ICDS) [21]:* ICDS used a multitasking network to solve CD and SS simultaneously by passing information from the SS branches to the CD branch and integrating their results finally.
6) *HRNet-Based SCD*[1] *(HBSCD):* HBSCD was the first-place method in the SenseTime CD competition, which used HRNet40 as the backbone, and then attached two SS heads and a CD head to solve CD and SS, respectively.
7) *SCDNet [22]:* SCDNet consisted of two encoders and two decoders with shared weights based on Siamese UNet. SCDNet fully integrated SS and CD information and directly performed SCD in the decoder.
8) *ChangeMask [23]:* ChangeMask decoupled the SCD into a temporal-wise SS and a CD and then integrated these two tasks into a general encoder-transformer–decoder framework.
9) *Bi-SRNet [30]:* Bi-SRNet used two types of semantic reasoning blocks of reasoning both single-temporal and cross-temporal semantic correlations, and used a novel loss function to improve the semantic consistency of CD results.
10) *SCanNet [31]:* SCanNet consisted of a SCanFormer which modeled the semantic transitions between dual-temporal RSIs and a semantic learning scheme that used spatio-temporal constraints to guide the learning of semantic changes.
11) *SMNet [32]:* SMNet was equipped with a backbone named PRTB, which constituted a hybrid unit that combined PRTB, in addition to an MCFM to capture fine-grained changes in complex scenes by distinguishing foreground and background information.

[1]https://github.com/LiheYoung/SenseEarth2020-ChangeDetection

4) Implementation Details: Our model was implemented in PyTorch and trained on two distinct datasets, SECOND and Landsat-SCD. For the SECOND dataset, we trained the model for 30 epochs with a batch size of 8, using two 3090Ti processors. Semi-supervised learning was initiated at epoch 15. For the Landsat-SCD dataset, we trained the model for 100 epochs with the same batch size of 8, and semi-supervised learning was initiated at epoch 80. For contrastive loss, we set $M = 50$, $N = 256$, $\tau = 0.5$, and $h = 256$. For overall loss, we set $\lambda_{\text{scd}} = 2$, $\lambda_{\text{st}}^{\text{ss}} = 1$, $\lambda_{\text{cl}}^{\text{ss}} = 0.2$ in SECOND dataset, and $\lambda_{\text{scd}} = 10$, $\lambda_{\text{st}}^{\text{ss}} = 5$, $\lambda_{\text{cl}}^{\text{ss}} = 0.2$ in Landsat-SCD dataset. During the training process, we employed the AdamW optimizer [76] with a weight decay of 0.0001. We used the poly learning rate scheduling with an initial learning rate of $1.5e^{-4}$ for the backbone and ten times larger for other heads. The learning rate was decayed based on the formula $\text{lr} = \text{baselr} \times (1 - (\text{iter}/\text{totaliter}))^{0.9}$. We augmented the training data by flipping horizontally and vertically, applying random rotation, and random color jitters. During the testing process, we adopted the test time augmentation (TTA) strategy, which involved rotating the input images by 90°, 180°, and 270°.

B. Comparative Experiments

1) SECOND Dataset: The performance comparison with the methods on the SECOND dataset is presented in Table I. The comprehensive experiments demonstrated that our baseline achieved the SOTA performance. Furthermore, by integrating SSCLM into the baseline, the proposed SSCLNet exhibited a significant improvement in performance, achieving 74.10%, 26.59%, 40.85% in mIoU, SeK, and score values, respectively.

For a more overall comparison, the model parameters, computational complexity, and inference time per batch of different methods are reported in Table II. SSCLNet significantly outperformed other methods with similar model parameters and computational complexity. Furthermore, our proposed self-supervised contrastive learning module exclusively contributes to the training process and is removed during inference. Consequently, it does not introduce any supplementary computational burden during inference and does not increase inference time.

We visualize the results of HBSCD, SCDNet, our baseline, and the proposed SSCLNet on the SECOND dataset in Fig. 3. Overall, our proposed method produced more accurate and

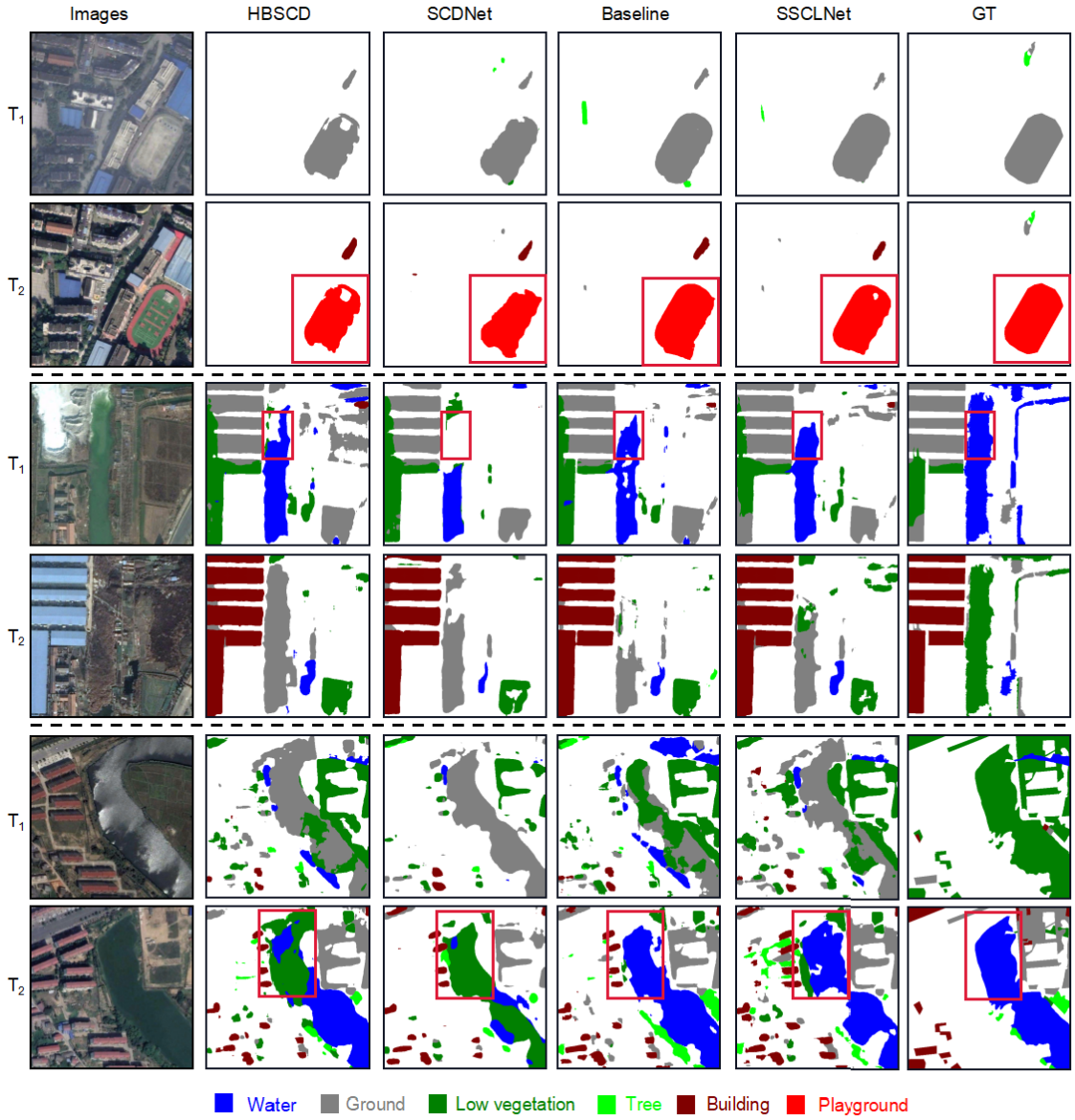

Fig. 3. Visual comparisons of the SCD maps obtained by the different methods and our SSCLNet on the SECOND dataset.

TABLE II

MODEL PARAMETERS, COMPUTATIONAL COMPLEXITY, TRAINING TIME (T-TIME), AND INFERENCE TIME (I-TIME) FOR DIFFERENT METHODS

Methods	Params (M)	Flops (GMac)	T-Time(s)	I-Time(ms)
DSCD	42.03	109.47	-	-
SCDS	79.97	285.67	-	-
ICDS	44.45	122.44	-	-
HBSCD	46.17	128.87	385	164
SCDNet	39.62	116.98	117	72
Baseline	46.80	122.24	365	94
SSCLNet	48.20	122.24	410	94

TABLE III

EFFECT OF SSCLM IN DIFFERENT BASELINES ON THE SECOND DATASET

Methods	mIoU(%)	SeK(%)	Score(%)
ResNet34+FCN	72.03	23.30	37.92
ResNet34+FCN+SSCLM	72.29(+0.26)	23.76(+0.46)	38.32(+0.40)
ResNet34+PSP	71.65	22.67	37.36
ResNet34+PSP+SSCLM	72.02(+0.37)	23.41(+0.74)	37.99(+0.63)
HRNet40+FCN	72.94	24.40	38.97
HRNet40+FCN+SSCLM	73.21(+0.27)	25.32(+0.92)	39.69(+0.72)
Our baseline	72.34	24.01	38.51
SSCLNet	**73.28(+0.94)**	**25.34(+1.33)**	**39.72(+1.21)**

sharper semantic change maps than the other methods, particularly on the shape of water and playground. Additionally, our method made semantic changes more discriminative, leading to more complete predictions.

2) Landsat-SCD Dataset: The performance comparison with the SOTA methods on the Landsat-SCD dataset is shown in Table I. The results of our experiments demonstrated that our baseline achieved the SOTA performance. Furthermore, by integrating SSCLM into the baseline, the proposed SSCLNet showcased a remarkable improvement in performance, with mIoU, SeK, and score values of 91.94%, 65.41%, and 73.21%, respectively.

We visualize the results of HBSCD, SCDNet, our baseline, and proposed SSCLNet on the Landsat-SCD dataset in Fig. 4. It can be seen that our proposed method achieved accurate prediction and reduced errors in semantic boundaries and predicted the complete shape of narrow streams within the water category compared to the other methods.

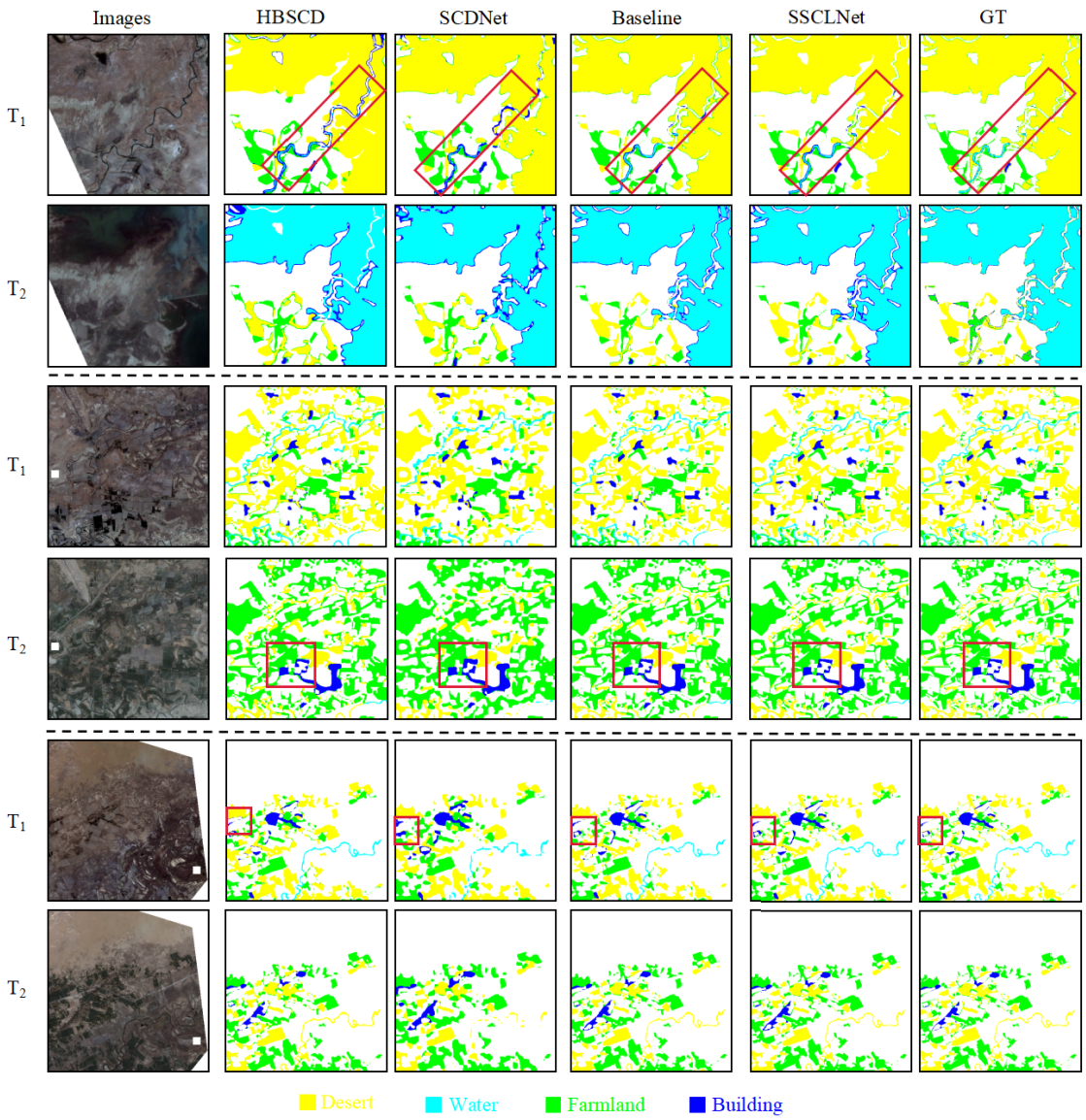

Fig. 4. Visual comparisons of the SCD maps obtained by the different methods and our SSCLNet on the Landsat-SCD dataset.

TABLE IV
EFFECT OF ST AND CONTRASTIVE LEARNING IN SSCLM ON THE SECOND DATASET

Baseline	ST	CL	mIoU(%)	SeK(%)	Score(%)
✓			72.33	24.01	38.51
✓	✓		72.80	24.89	39.26
✓		✓	72.24	24.19	38.60
✓	✓	✓	**73.28**	**25.34**	**39.72**

TABLE V
EFFECT OF ST AND CONTRASTIVE LEARNING IN SSCLM ON THE LANDSAT-SCD DATASET

Baseline	ST	CL	mIoU(%)	SeK(%)	Score(%)
✓			90.40	62.60	70.94
✓	✓		90.72	63.48	71.65
✓		✓	90.66	63.24	71.46
✓	✓	✓	**90.97**	**64.09**	**72.15**

C. Ablation Study

1) Effect of SSCLM in Different Baselines: To demonstrate the effectiveness of self-supervised contrastive learning module, we conducted experiments using multiple baselines on the SECOND dataset. We chose widely adopted ResNet34 [77] and HRNet40 as encoders, and commonly used fully convolutional networks (FCN) and PSP as decoders. These selections enabled us to develop four distinct baseline models.

Table III compares the experimental results of the four baseline models with and without SSCLM. Our results indicated that the integration of the SSCLM significantly improved the performance of the four baseline models. Notably, utilizing HRNet40 as the encoder led to better performance improvement compared to ResNet34. This is due to the fact that HRNet40 could maintain high resolution throughout the encoding process, allowing it to provide a more detailed feature representation that is beneficial for contrastive learning.

2) Effect of Self-Training and Contrastive Learning in SSCLM: We conducted ablation experiments on the SECOND dataset and Landsat-SCD dataset to verify the effectiveness of ST and contrastive learning in SSCLM. Table IV shows the experimental results on the SECOND dataset. Table V shows the experimental results on the Landsat-SCD dataset.

TABLE VI
EFFECT OF SSCLM FOR ENHANCING THE DISTINGUISHABILITY OF DIFFERENT CATEGORIES ON THE SECOND DATASET

Methods	F1-Score(%)							
	Water	Ground	Low vegetation	Tree	Building	Playground	Non-changed	Changed
Baseline	55.80	64.03	55.92	41.26	76.07	55.45	92.50	73.92
SSCLNet	61.54(+5.74)	64.24(+0.21)	57.63(+1.71)	43.19(+1.93)	77.43(+1.36)	67.95(+12.50)	92.54(+0.04)	74.58(+0.66)

TABLE VII
EFFECT OF SSCLM FOR ENHANCING THE DISTINGUISHABILITY OF DIFFERENT CATEGORIES ON THE LANDSAT-SCD DATASET

Methods	F1-Score(%)					
	Farmland	Desert	Building	Water	Non-changed	Changed
Baseline	86.45	89.92	78.96	92.73	97.41	92.39
SSCLNet	86.83(+0.38)	90.14(+0.22)	79.49(+0.53)	92.94(+0.21)	97.51(+0.10)	92.67(+0.28)

As shown in Tables IV and V, only applying ST method to SS and independently applying fully supervised contrastive learning to SS both result in performance improvement; however, it is noteworthy that the application of semi-supervised learning to SS yields a more substantial performance enhancement. This result indicated that applying ST SS to fully utilize the semantic context information of unlabeled data is useful for SCD. However, due to the fact that semantic labels are available for changed categories with a small proportion in most SCD datasets, resulting in an insufficient sample space for contrastive learning. With the cooperation of ST and contrastive learning, in the SECOND dataset, the performance of SSCLNet significantly improved with a gain of 0.95%, 1.33%, and 1.21% for mIoU, Sek, and score, respectively. In the Landsat-SCD dataset, the performance of SSCLNet significantly improved with a gain of 0.37%, 1.49%, and 1.21% for mIoU, Sek, and score, respectively. This improvement is due to that ST can provide a sufficient sample space for contrastive learning, allowing SSCLNet to learn distinguishable feature representations from both labeled and unlabeled data.

3) Effect of SSCLM for Enhancing the Distinguishability of Categories: In order to demonstrate that SSCLM can enhance the distinguishability of different categories, we conducted experiments on the SECOND and Landsat-SCD datasets to observe the segmentation results of the baseline and SSCLNet across all categories.

For the SECOND dataset, we evaluated the F1-score of our baseline and SSCLNet on different land-cover categories in changed regions, which account for less than 20% of the entire dataset. As shown in Table VI, the performance of SSCLNet with SSCLM surpassed that of the baseline by a significant margin. Notably, categories with a small proportion, such as water and playground, exhibited substantial performance gains of 5.74% and 12.50%, respectively. Moreover, categories with similar semantic relationships, such as low vegetation and tree, demonstrated significant improvements of 1.71% and 1.93%. Additionally, categories with a larger proportion, such as ground and building, also experienced performance enhancements of 0.21% and 1.36%, respectively. We further evaluated the F1-score of our baseline and SSCLNet on non-changed and changed categories, where the non-changed category accounts for more than 80% of the entire dataset. The results indicated that SSCLNet outperformed the baseline, exhibiting an enhancement of 0.04% and 0.66% in F1-score for the non-changed and changed categories, respectively. In particular, the changed category showed a greater improvement in performance.

TABLE VIII
COMPARATIVE RESULTS OF DIFFERENT SAMPLING STRATEGIES FOR CONTRASTIVE LEARNING ON THE SECOND DATASET

Methods	mIoU(%)	SeK(%)	Score(%)
Random	72.80	24.89	39.26
EasyNeg	72.62	24.78	39.13
HardNeg	72.97	24.84	39.28
HalfNeg	73.13	24.93	39.39
Ours	**73.28**	**25.34**	**39.72**

For the Landsat-SCD dataset, the evaluation involved comparing the F1-score of our baseline and SSCLNet on various land-cover categories. As indicated in Table VII, the performance of SSCLNet with SSCLM outperformed that of our baseline. Notably, categories with a small proportion, such as building and farmland, exhibited substantial performance gains of 0.53% and 0.38%, respectively. Moreover, categories with close spatial relationships, such as desert and farmland, demonstrated significant improvements of 0.22% and 0.38%. Additionally, categories with a larger proportion, such as desert and water, also experienced performance enhancements of 0.22% and 0.21%, respectively. In addition, we also evaluated the F1-score of our baseline and SSCLNet on both non-changed and changed categories. The SSCLNet outperformed our baseline, demonstrating an improvement of 0.10% and 0.28% in F1-score on the non-changed and changed categories, respectively. The improvement was more significant for changed categories.

The above experiments fully demonstrated the effectiveness of SSCLM for enhancing the distinguishability of different categories.

4) Effect of Adaptive Sampling Strategy for Contrastive Learning: To show the effectiveness of our proposed adaptive sampling strategy for contrastive learning, we presented ablation experiments on the SECOND dataset, as detailed in Table VIII.

Random strategy randomly sampled negatives for each category. EasyNeg strategy only sampled easy negatives for each category. HardNeg strategy only sampled hard negatives for each category. HalfNeg strategy sampled half of the easy negatives and half of the hard negatives for each category.

Here, hard negatives and easy negatives were distinguished by a prediction probability threshold δ, where samples with a prediction probability smaller than δ were considered hard and those with a higher probability were considered easy. In the experiment, we set δ to 0.97.

Experiments demonstrated that our adaptive sampling strategy was the most effective way to sample negative examples in contrastive learning. Specifically, EasyNeg strategy, which selected easy negatives, did not provide significant contributions to contrastive loss during training. On the other hand, HardNeg strategy, which selected hard negatives, can offer valuable discriminative information, but it may hinder network convergence and lead to bad local minima if too many hard samples were chosen. Although HalfNeg strategy performed slightly better than HardNeg strategy, it was still not as effective as our adaptive sampling strategy. Our proposed adaptive sampling strategy involved selecting negative examples for each category from categories that were easily confused with it, which significantly improved the overall SCD performance by enhancing the discriminability of different categories.

V. Conclusion

This article presents SSCLNet, a simple and flexible SCD network boosted by semi-supervised contrastive learning of SS. By incorporating semi-supervised contrastive learning into SS, SSCLNet can effectively use both labeled and unlabeled data and enhance its discriminative capability of confusing categories. Experimental results on the publicly available SECOND dataset and Landsat-SCD dataset demonstrate the effectiveness of our proposed method. In particular, SSCLNet with semi-supervised contrastive learning of SS outperforms the SOTA methods, showing its potential in SCD tasks. In our future research, we plan to extend the application of contrastive learning beyond SS, such as CD in SCD. This expansion will allow us to explore broader potentials of contrastive learning in SCD and provide a more comprehensive understanding of its effectiveness.

References

[1] K. Yang et al., "Asymmetric Siamese networks for semantic change detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2021.

[2] R. S. Lunetta, J. F. Knight, J. Ediriwickrema, J. G. Lyon, and L. D. Worthy, "Land-cover change detection using multi-temporal MODIS NDVI data," *Remote Sens. Environ.*, vol. 105, no. 2, pp. 142–154, Nov. 2006.

[3] S. Jin, L. Yang, Z. Zhu, and C. Homer, "A land cover change detection and classification protocol for updating Alaska NLCD 2001 to 2011," *Remote Sens. Environ.*, vol. 195, pp. 44–55, Jun. 2017.

[4] C. Zhang et al., "Joint deep learning for land cover and land use classification," *Remote Sens. Environ.*, vol. 221, pp. 173–187, Feb. 2019.

[5] T. Suzuki, S. Shirakabe, Y. Miyashita, A. Nakamura, Y. Satoh, and H. Kataoka, "Semantic change detection with hypermaps," 2016, *arXiv:1604.07513*.

[6] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[7] S. Jabari, M. Rezaee, F. Fathollahi, and Y. Zhang, "Multispectral change detection using multivariate Kullback–Leibler distance," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 163–177, Jan. 2019.

[8] J. Yan, L. Wang, W. Song, Y. Chen, X. Chen, and Z. Deng, "A time-series classification approach based on change detection for rapid land cover mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 249–262, Dec. 2019.

[9] G. Chen, G. J. Hay, L. M. T. Carvalho, and M. A. Wulder, "Object-based change detection," *Int. J. Remote Sens.*, vol. 33, no. 14, pp. 4434–4457, 2012.

[10] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS J. Photogramm. Remote Sens.*, vol. 80, pp. 91–106, Jun. 2013.

[11] X. Chen, J. Chen, Y. Shi, and Y. Yamaguchi, "An automated approach for updating land cover maps based on integrated change detection and classification methods," *ISPRS J. Photogramm. Remote Sens.*, vol. 71, pp. 86–95, Jul. 2012.

[12] S. Jin, L. Yang, P. Danielson, C. Homer, J. Fry, and G. Xian, "A comprehensive change detection method for updating the national land cover database to circa 2011," *Remote Sens. Environ.*, vol. 132, pp. 159–175, May 2013.

[13] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, Jun. 2019.

[14] W. Yu, W. Zhou, Y. Qian, and J. Yan, "A new approach for land cover classification and change analysis: Integrating backdating and an object-based method," *Remote Sens. Environ.*, vol. 177, pp. 37–47, May 2016.

[15] M. Hao, W. Shi, K. Deng, H. Zhang, and P. He, "An object-based change detection approach using uncertainty analysis for VHR images," *J. Sensors*, vol. 2016, pp. 1–17, Nov. 2016.

[16] Z. Huang, X. Jia, and L. Ge, "Sampling approaches for one-pass land-use/land-cover change mapping," *Int. J. Remote Sens.*, vol. 31, no. 6, pp. 1543–1554, Mar. 2010.

[17] J. Hu and Y. Zhang, "Seasonal change of land-use/land-cover (LULC) detection using MODIS data in rapid urbanization regions: A case study of the pearl river delta region (China)," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 4, pp. 1913–1920, Aug. 2013.

[18] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion," *Remote Sens. Environ.*, vol. 199, pp. 241–255, Sep. 2017.

[19] A. Smith, "Digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 17, no. 11, pp. 2043–2057, 1996.

[20] L. Mou, L. Bruzzone, and X. X. Zhu, "Learning spectral–spatial–temporal features via a recurrent convolutional neural network for change detection in multispectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 924–935, Feb. 2019.

[21] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Understand.*, vol. 187, Oct. 2019, Art. no. 102783.

[22] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, and P. He, "SCDNET: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, Dec. 2021, Art. no. 102465.

[23] Z. Zheng, Y. Zhong, S. Tian, A. Ma, and L. Zhang, "ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 228–239, Jan. 2022.

[24] D. Xiang, Y. Xie, J. Cheng, Y. Xu, H. Zhang, and Y. Zheng, "Optical and SAR image registration based on feature decoupling network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5235913.

[25] D. Xiang, Y. Xu, J. Cheng, Y. Xie, and D. Guan, "Progressive keypoint detection with dense Siamese network for SAR image registration," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 5, pp. 5847–5858, Oct. 2023, doi: 10.1109/TAES.2023.3266415.

[26] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2020.

[27] K. Yang et al., "Semantic change detection with asymmetric Siamese networks," 2020, *arXiv:2010.05687*.

[28] P. Yuan, Q. Zhao, X. Zhao, X. Wang, X. Long, and Y. Zheng, "A transformer-based Siamese network and an open optical dataset for semantic change detection of remote sensing images," *Int. J. Digit. Earth*, vol. 15, no. 1, pp. 1506–1525, Dec. 2022.

[29] Q. Zhu et al., "Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 184, pp. 63–78, Feb. 2022.

[30] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620014.

[31] L. Ding, J. Zhang, K. Zhang, H. Guo, B. Liu, and L. Bruzzone, "Joint spatio-temporal modeling for the semantic change detection in remote sensing images," 2022, *arXiv:2212.05245*.

[32] Y. Niu, H. Guo, J. Lu, L. Ding, and D. Yu, "SMNet: Symmetric multi-task network for semantic change detection in remote sensing images based on CNN and transformer," *Remote Sens.*, vol. 15, no. 4, p. 949, Feb. 2023.

[33] M. Hermann, S. Saha, and X. X. Zhu, "Filtering specialized change in a few-shot setting," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1185–1196, 2023.

[34] J. Hyun Cho, U. Mall, K. Bala, and B. Hariharan, "PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16789–16799.

[35] S. Saha, M. Shahzad, L. Mou, Q. Song, and X. X. Zhu, "Unsupervised single-scene semantic segmentation for Earth observation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022, Art. no. 5228011, doi: 10.1109/TGRS.2022.3174651.

[36] V. Marsocci and S. Scardapane, "Continual barlow twins: Continual self-supervised learning for remote sensing semantic segmentation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5049–5060, 2023.

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

[38] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," 2018, *arXiv:1802.07934*.

[39] 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. 2019.

[40] I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[41] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with Ladder networks," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 28, 2015.

[42] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[43] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2016, *arXiv:1610.02242*.

[44] 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.

[45] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6256–6268.

[46] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 17, 2004.

[47] D.-H. Lee et al., "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, vol. 3, no. 2, p. 896.

[48] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy student improves ImageNet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 10687–10698.

[49] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[50] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," 2017, *arXiv:1701.04862*.

[51] Y. Guo, F. Wang, Y. Xiang, and H. You, "Semisupervised semantic segmentation with certainty-aware consistency training for remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2900–2914, 2023.

[52] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, varied perturbations," 2019, *arXiv:1906.01916*.

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

[54] Z. Ke, D. Qiu, K. Li, Q. Yan, and R. W. Lau, "Guided collaborative training for pixel-wise semi-supervised learning," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2613–2622.

[56] J. Li, B. Sun, S. Li, and X. Kang, "Semisupervised semantic segmentation of remote sensing images with consistency self-training," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022, Art. no. 5615811, doi: 10.1109/TGRS.2021.3134277.

[57] J.-X. Wang, S.-B. Chen, C. H. Q. Ding, J. Tang, and B. Luo, "RanPaste: Paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, Art. no. 2002916, doi: 10.1109/TGRS.2021.3102026.

[58] J. Chen et al., "Semi-supervised semantic segmentation framework with pseudo supervisions for land-use/land-cover mapping in coastal areas," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102881.

[59] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "ST++: Make self-training work better for semi-supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4268–4277.

[60] X. Sun, A. Shi, H. Huang, and H. Mayer, "BAS^{4}Net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observat. Remote Sens.*, vol. 13, pp. 5398–5413, 2020.

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

[62] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Stacked autoencoders driven by semi-supervised learning for building extraction from near infrared remote sensing imagery," *Remote Sens.*, vol. 13, no. 3, p. 371, Jan. 2021.

[63] L. Kondmann, S. Saha, and X. X. Zhu, "SemiSiROC: Semisupervised change detection with optical imagery and an unsupervised teacher model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3879–3891, 2023.

[64] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," 2021, *arXiv:2104.04465*.

[65] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8219–8228.

[66] J. Li, C. Xiong, and S. C. H. Hoi, "CoMatch: Semi-supervised learning with contrastive graph regularization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9475–9484.

[67] W. Wang, T. Zhou, F. Yu, J. Dai, E. Konukoglu, and L. V. Gool, "Exploring cross-image pixel contrast for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7303–7313.

[68] H. Hu, J. Cui, and L. Wang, "Region-aware contrastive learning for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16271–16281.

[69] P. Khosla et al., "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.

[70] X. Zhao et al., "Contrastive learning for label efficient semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10623–10633.

[71] Z. Zhao, L. Zhou, L. Wang, Y. Shi, and Y. Gao, "LaSSL: Label-guided self-training for semi-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 8, pp. 9208–9216, doi: 10.1609/aaai.v36i8.20907.

[72] F. Yang et al., "Class-aware contrastive semi-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14421–14430.

[73] Y. Wang et al., "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 4248–4257.

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

[75] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[76] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

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