

# GraSS: Contrastive Learning with Gradient Guided Sampling Strategy for Remote Sensing Image Semantic Segmentation

Zhaoyang Zhang, Zhen Ren, Chao Tao, Yunsheng Zhang, Chengli Peng, and Haifeng Li\*

**Abstract**—Self-supervised contrastive learning (SSCL) has achieved significant milestones in remote sensing image (RSI) understanding. Its essence lies in designing an unsupervised instance discrimination pretext task to extract image features from a large number of unlabeled images that are beneficial for downstream tasks. However, existing instance discrimination based SSCL suffers from two limitations when applied to the RSI semantic segmentation task: 1) Positive sample confounding issue, SSCL treats different augmentations of the same RSI as positive samples, but the richness, complexity, and imbalance of RSI ground objects lead to the model actually pulling a variety of different ground objects closer while pulling positive samples closer, which confuse the feature of different ground objects. 2) Feature adaption bias, SSCL treats RSI patches containing various ground objects as individual instances for discrimination and obtains instance-level features, which are not fully adapted to pixel-level or object-level semantic segmentation tasks. To address the above limitations, we consider constructing samples containing single ground objects to alleviate positive sample confounding issue, and make the model obtain object-level features from the contrastive between single ground objects. Meanwhile, we observed that the discrimination information can be mapped to specific regions in RSI through the gradient of unsupervised contrastive loss, these specific regions tend to contain single ground objects. Based on this, we propose contrastive learning with Gradient Guided Sampling Strategy (GraSS) for RSI semantic segmentation. GraSS consists of two stages: 1) the instance discrimination warm-up stage to provide initial discrimination information to the contrastive loss gradients, 2) the gradient guided sampling contrastive training stage to adaptively construct samples containing more singular ground objects using the discrimination information. Experimental results on three open datasets demonstrate that GraSS effectively enhances the performance of SSCL in high-resolution RSI semantic segmentation. Compared to eight baseline methods from six different types of SSCL, GraSS achieves an average improvement of 1.57% and a maximum improvement of 3.58% in terms of mean intersection over the union. Additionally, we discovered that the unsupervised contrastive loss gradients contain rich feature information, which inspires us to utilize gradient information more extensively during model training to attain additional model capacity. The source code is available at <https://github.com/GeoX-Lab/GraSS>.

**Index Terms**—Self-supervised learning, contrastive loss, gradient guided, semantic segmentation, remote sensing image (RSI).

The work is financial supported by the Major Program/Open Project of Xiangjiang Laboratory (No. 22XJ01010), The National Natural Science Foundation of China (No. 42171376, 41771458), The Natural Science Foundation of Hunan for Distinguished Young Scholars (No. 2022JJ10072); and the High-Performance Computing Center of Central South University. (Corresponding author: H.F. Li, [lihaifeng@csu.edu.cn](mailto:lihaifeng@csu.edu.cn))

Z. Zhang, Z. Ren, C. Tao, Y. Zhang, C. Peng, and H. Li are with the School of Geosciences and Info-Physics, Central South University, Changsha 410083, China. Z. Zhang and H. Li are also with the Xiangjiang Laboratory, Changsha 410205, China.

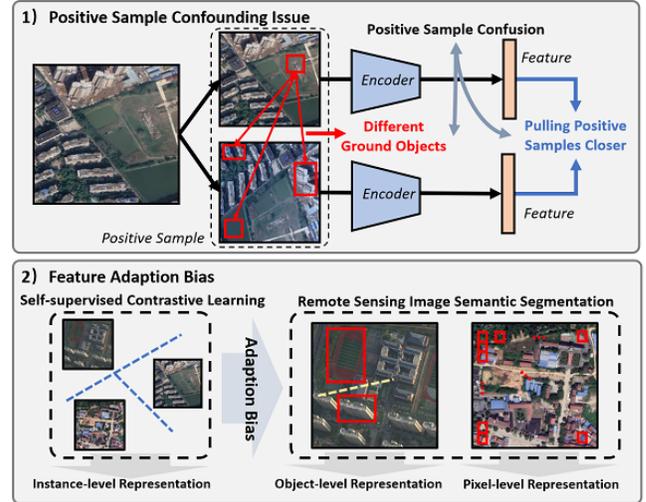


Fig. 1. Example of the positive sample confounding issue and feature adaption bias in self-supervised contrastive learning for RSI semantic segmentation. 1) The richness, and complexity of RSI ground objects can result in positive samples containing different ground objects, self-supervised contrastive learning (SSCL) lead to positive sample confounding issue when pulling these positive samples closer in the feature space. 2) SSCL treats RSI patches containing various ground objects as individual instances for discrimination, resulting in feature representations at the instance level, it introduces a feature adaption bias when applied to semantic segmentation tasks that require pixel-level or object-level features.

## I. INTRODUCTION

SELF-supervised contrastive learning has achieved significant success in various downstream tasks such as remote sensing image (RSI) scene classification [1]–[5], hyperspectral image classification [6]–[11], object detection [12], [13], change detection [14], [15], and semantic segmentation [16]–[18]. Its core idea is to learn effective image representations by designing an unsupervised instance discrimination pretext task [19]–[27].

However, there are two limitations of the unsupervised instance discrimination pretext task when applied to the task of RSI semantic segmentation that requires capturing features of ground objects [16]–[18], [28], [29]. First, the positive sample confounding issue. Positive sample confounding issue is one aspect of the sample confounding issue (SCI) [18], [30]. Due to the richness, complexity, and imbalance of ground objects contained in RSIs [18], [31], [32], the sample confounding issue in self-supervised contrastive learning models manifests in two aspects: The first aspect is that for negative samples,

the self-supervised contrastive learning model treats augmentations of all different images as negative samples, which leads the model to inevitably push away negative samples that contain the same ground objects as the anchor sample, we call this the negative sample confounding issue, which is also often referred to as the false negative sample issue [18], [33], [34]. The second aspect is that for positive samples, since RSIs contain a variety of ground objects, the model actually pulls a variety of different ground objects in positive samples closer while pulling positive samples closer, which makes the model confuse the features of different ground objects, we call this the positive sample confounding issue (as shown in Fig. 1). The positive sample confounding issue undermines the identity assumption of self-supervised contrastive learning [35]–[38], which is the focus of this study. Second, the feature adaptation bias. Self-supervised contrastive learning treats RSI patches containing various ground objects as individual instances, focusing more on the relationship between instances and ignoring the relationship between ground objects in RSIs [16], [29]. It obtains features at the instance level, with feature adaptation bias for RSI semantic segmentation tasks that require pixel-level or object-level features [12], [13], [29], [39]–[42].

To address the positive sample confounding issue, ContrastiveCrop [43] and Leopart [44] use the activation map forwarded from the image to the feature layer to guide sampling, and construct higher quality positive samples. However, these methods ignore the gradient of the contrastive loss backpropagated to the feature layer, and do not make full use of the discriminative information contained in the contrastive loss. Recent works such as LCR [45] add an additional branch to align the feature activation map and the contrastive loss gradient activation map, which effectively improves the performance of self-supervised contrastive learning on fine-grained visual recognition. However, this method only changes the feature of the samples, and does not use the activation map to reconstruct the positive samples. For the RSI semantic segmentation task, the positive samples still contain a variety of ground objects, which cannot effectively alleviate the positive sample confounding issue.

To address the feature adaptation bias, DenseCL [28], VADeR [29], and IndexNet [17] use dense contrastive approach to optimize the pixel-level contrastive loss between the two views of the input image. However, for the semantic segmentation task of high-resolution remote sensing images, these approaches inevitably lead to higher contrastive learning overhead. In addition, GLCNet [16] considers adding a local contrastive module for decoder feature maps to the original instance-level contrastive, but this requires the decoder structure for semantic segmentation to be specified in the self-supervised pretraining stage, although the main target of the pretraining stage is to obtain the encoder network.

Unlike the above methods, we observed that the discrimination information contained in the contrastive loss can be mapped to specific regions in RSI through the gradient of unsupervised contrastive loss, these specific regions tend to contain single ground objects. Therefore, we utilize the gradient of the contrastive loss backpropagation to the feature layer to

guide sampling and iteratively construct positive and negative samples that contain more singular ground objects during the training process. The major difference between the proposed GraSS and previous work is that the GraSS fully utilizes the contrastive loss gradient to resample the RSI as input to the model, without adding a dense contrastive module or local contrastive module in the pretraining stage. The experimental results indicate that can effectively alleviate the positive sample confounding issue caused by positive samples containing various ground objects, and because the positive and negative samples constructed contain more singular ground objects, our approach will also make the instance-level contrastive closer to the object-level contrastive, effectively mitigating the feature adaptation bias of the instance discrimination pretext task to the downstream semantic segmentation task.

The main contributions of this paper are:

- 1) We propose self-supervised contrastive learning with Gradient guided Sampling Strategy (GraSS) for remote sensing image semantic segmentation, which uses the positive and negative sample discrimination information from the contrastive loss gradient to guide the positive and negative sample construction. It effectively alleviates the positive sample confounding issue and feature adaptation bias of the self-supervised contrastive learning for RSI semantic segmentation, without adding the additional dense contrastive module or local contrastive module.
- 2) We find that the positive and negative sample discrimination information contained in the contrastive loss gradient can be mapped to specific regions on the RSI, which often contain more singular ground objects. This indicates that the gradient of contrastive loss contains rich feature information, which inspires us to make more use of gradient information to obtain additional model capability in the process of model training.
- 3) The experimental results on three open datasets, Potsdam, LoveDA Urban, and LoveDA Rural, show that GraSS achieves the best performance compared with eight self-supervised contrastive learning baseline methods from six different types of positive and negative sample construction, and its improved by 1.57% on average and 3.58% on maximum of mean intersection over the union (mIoU).

## II. RELATED WORK

### A. Construction of Positive and Negative Samples

The construction of positive and negative samples is the basis of self-supervised contrastive learning [2], [20], [35], [38], [46], [47], which usually regards different data augmentations of the same image as positive samples and data augmentations of different images as negative samples [18], [20], [23], [46]. The data augmentation method can be divided into two categories according to the different image attributes changed: one is spectral transformation, such as random color distortion [16], Gaussian blur [48], and the other is spatial change, such as random resize crop [20], [48], random flip [48], [49]. Different data augmentation methods have different

impacts on the self-supervised contrastive learning model [3], [16], [20], [21], [50]. Among them, the augmentation combination of random resize crop and color distortion has been proven to bring greater performance improvement to the self-supervised contrastive learning model [20]. In addition, considering the spatio-temporal heterogeneity of RSIs, STICL [3] and SeCo [50] make full use of the temporal-shifting characteristics of RSIs to propose a positive and negative sample construction method that is more adaptable to remote sensing image processing.

However, due to the richness, complexity, and imbalance of remote sensing images, the self-supervised contrastive learning model for RSI semantic segmentation suffers from severe sample confounding issue [18], [19], [27]: First, negative sample confounding issue, which is often called false negative sample issue [18], [33], [34]. In order to solve the false negative sample issue, FALSE [18] considers the self-correcting signal based on positive samples and true negative samples giving feedback to the model to guide the model to improve the construction of negative samples and alleviate the false negative sample issue, while IFND [34] and FNC [33] considers the semantic structure of feature space to dynamically detect false negative samples.

The second is the positive sample confounding issue, which undermines the identity assumption of self-supervised contrastive learning [35]–[38] and is the focus of this paper. In order to solve the positive sample confounding issue, some recent research [43], [44] used the feature activation maps of images to select specific regions from the original images to generate positive samples. These approach aims to obtain positive samples with semantic consistency guarantees using the activation information forward propagated from the image to the model feature layer. In addition, recent works such as LCR [45] introduce the information of contrastive loss gradient, and consider adding a GradCAM fitting branch (GFB) [45] to the original contrastive learning model to align the feature activation map and the contrastive loss gradient activation map, which effectively improves the performance of self-supervised contrastive learning on fine-grained visual recognition. However, different from our proposed GraSS, the LCR adds an additional branch and loss function, and only changes the features of the sample, the activation map is not used to guide the resampling. For the RSI semantic segmentation, the positive samples still contain a variety of ground objects, which makes it difficult to effectively alleviate the positive sample confounding issue.

### B. Dense Contrastive Learning

The instance discrimination pretext task of self-supervised contrastive learning acquires image instance-level features, which is naturally adapted to image-level downstream tasks such as RSI scene classification [2], [12], [13], [16], [17], [27], [39], but suffers from feature adaptation bias for RSI semantic segmentation that requires object-level or pixel-level [12], [13], [29], [39]–[41]. A natural idea for mitigating feature adaptation bias is to optimize the pixel-level contrastive loss between the two views of the input image by dense contrastive

approach, giving the model the ability to capture features at the pixel-level or object-level of the image. Methods such as IndexNet [17], DenseCL [28], and VADeR [29] add a dense contrastive module to the original instance-level contrastive and obtain stable performance gains in RSI semantic segmentation and object detection, but this inevitably leads to higher computational overhead.

In addition, GLCNet [16] considers adding a local contrastive module for semantic segmentation decoder feature maps to the original instance-level contrastive, but this requires the decoder structure for semantic segmentation to be specified in the self-supervised pretraining stage.

## III. METHOD

### A. Overview

The core idea of the method is derived from the basic characteristics of self-supervised contrastive learning models: self-supervised contrastive learning constrains the model to obtain image features by designing unsupervised instance discrimination pretext task, which can be seen as an image classifier that treats each RSI sample as an independent category. Inspired by the fact that deep network image classifiers tend to rely on a major region of an instance and ignore information about other regions when discriminating between different image instances [51]–[53]: We expect to obtain the regions of semantic consistency that the self-supervised contrastive learning model focuses on during instance discrimination and use the obtained semantic consistency regions to construct positive and negative samples. We observe that the positive and negative sample discrimination information contained in the contrastive loss gradients can be mapped to specific regions in RSI through the backpropagation of contrastive loss. These specific areas tend to contain single ground objects. Extracting these specific areas as positive and negative samples can effectively solve the positive sample confounding issue and feature adaptation bias of self-supervised contrastive learning for RSI semantic segmentation.

Therefore, we designed two training stages: 1) instance discrimination warm-up and 2) gradient guided sampling contrastive training. The overall framework of the GraSS is shown in Fig. 2. The instance discrimination warm-up stage aims to give the initial positive and negative sample discrimination information to the contrastive loss gradient, which is used to constrain the model to perform instance-level discrimination. The gradient guided sampling contrastive training stage aims to use the gradients of contrastive loss to obtain regions in RSI patches that contain more singular ground objects, in order to construct new positive and negative samples. In this stage, we calculated the contrastive loss twice: the first calculation is to obtain the gradient of the contrastive loss backpropagation to the image feature layer and obtain the activation map. The second calculation is to update the model parameters.

### B. Instance Discrimination Warm Up

The purpose of the instance discrimination warm-up stage is to train the model to acquire initial instance discrimination capabilities, with contrastive loss at this stage used to constrain

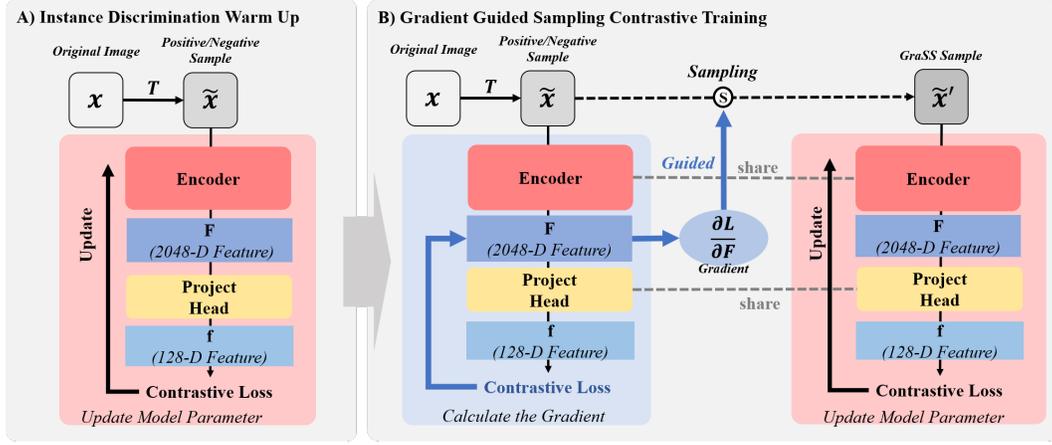


Fig. 2. Overview of contrastive learning with Gradient guided sampling strategy (GraSS) for remote sensing image semantic segmentation. A) is the Instance Discrimination Warm Up, the original image is processed by the augmentation function  $T$  to obtain positive and negative samples, and the positive and negative samples are input into the model to calculate the contrastive loss and update the model parameters; B) is Gradient Guided Sampling Contrastive Training, the original image is processed by the augmentation function  $T$  to obtain positive and negative samples, and the positive and negative samples are input into the model to calculate the gradient, and then the new samples obtained by gradient-guided sampling are input into the model to calculate the contrastive loss and update the model parameters.

the model to perform instance discrimination. This stage mainly includes the construction of positive and negative sample, model feature extraction, calculation of contrastive loss, and updating of model parameters.

1) *Construction of Positive and Negative Samples:* For the RSI data set  $x = \{x_i\}_{i=1}^N$ , it is augmented by function  $T$  to get  $N \cdot K$  sample instances  $\tilde{x} = \{\tilde{x}_i\}_{i=1}^N$ , where  $\tilde{x}_i = \{\tilde{x}_{ij}\}_{j=1}^K$ . The function  $T$  consists of three operations: image copy  $c(\cdot)$ , random spectral augmentation  $rc(\cdot)$ , and random spatial augmentation  $rs(\cdot)$ . Therefore, this process can be described as:

$$\tilde{x} = T(x) = rs(rc(c(x))) \quad (1)$$

$$\tilde{x}_i^c = c(x_i) = [\tilde{x}_{i1}^c, \tilde{x}_{i2}^c, \dots, \tilde{x}_{iK}^c] \quad (2)$$

$$\tilde{x}_i^{Rc} = rc(\tilde{x}_i^c) = [\tilde{x}_{i1}^{Rc}, \tilde{x}_{i2}^{Rc}, \dots, \tilde{x}_{iK}^{Rc}] \quad (3)$$

$$\tilde{x}_i = rs(\tilde{x}_i^{Rc}) = [\tilde{x}_{i1}, \tilde{x}_{i2}, \dots, \tilde{x}_{iK}] \quad (4)$$

where  $\tilde{x}_{i1}^c = \tilde{x}_{i2}^c = \dots = \tilde{x}_{iK}^c = x_i$ . All image instances  $\tilde{x}_i$  within any  $\tilde{x}_{ij}$  are obtained from the same original image  $x_i$ , and they are positive samples of each other. Any two image instances  $\tilde{x}_p$  and  $\tilde{x}_q$  ( $p \neq q$ ) are negative samples of each other.

2) *Model Feature Extraction and Contrastive Loss Calculation:* For image sample instances  $\tilde{x}$ , we input them into the encoder feature extraction network  $E(\cdot)$  to obtain high-dimensional features  $F$ , and further input high-dimensional features  $F$  into the feature projection head  $P(\cdot)$  to obtain low-dimensional features  $f$  to calculate the contrastive loss  $L$ , and iteratively update the model parameters. Specifically, for the image instance  $\tilde{x}_{ij}$ , model feature extraction and contrastive loss calculation can be described as:

$$F_{ij} = E(\tilde{x}_{ij}) \quad (5)$$

$$f_{ij} = P(F_{ij}) \quad (6)$$

$$l_{ij} = -\log\left(\frac{\sum_{n=1, n \neq j}^K \exp(\text{sim}(f_{ij}, f_{in})/\tau)}{\sum_{m=1, m \neq i}^N \sum_{n=1}^K \exp(\text{sim}(f_{ij}, f_{mn})/\tau)}\right) \quad (7)$$

where  $K$  is typically 2,  $\tau$  is the temperature parameter, and  $\text{sim}(\cdot, \cdot)$  usually uses cosine similarity. For each iterative parameter update process, the contrastive loss is finally defined as:

$$L = \frac{1}{N \cdot K} \sum_{i=1}^N \sum_{j=1}^K l_{ij}, \quad (8)$$

### C. Gradient Guided Sampling Contrastive Training

The gradient guided sampling contrastive training stage aims to use the gradient of the contrastive loss to obtain regions in RSI patches that contain more singular ground objects, in order to reconstruct positive and negative samples. This stage involves the construction of positive and negative sample instances, the acquisition of the Discrimination Attention Region (DAR), the reconstruction samples, and the calculation of contrastive loss and model parameter updates.

The settings for the construction of positive and negative sample instances and calculation of contrastive loss are kept consistent with the instance discrimination warm-up. More details of the acquisition of the Discrimination Attention Region (DAR), the reconstruction samples are shown in Fig. 3.

1) *Acquisition of Discrimination Attention Region (DAR):* After the construction of positive and negative sample instances using Eq. (1)-(4), we project the image instances to the low-dimensional features  $f$  according to Eq. (5)-(6), and calculate the contrastive loss using Eq. (7)-(8) to characterize the distribution of positive and negative sample instances in the feature space. Then, we use backpropagation to calculate the gradient of the contrastive loss to the high-dimensional feature  $\frac{\partial L}{\partial F}$ . For the image instance  $\tilde{x}_{ij}$ , this process can be described as:

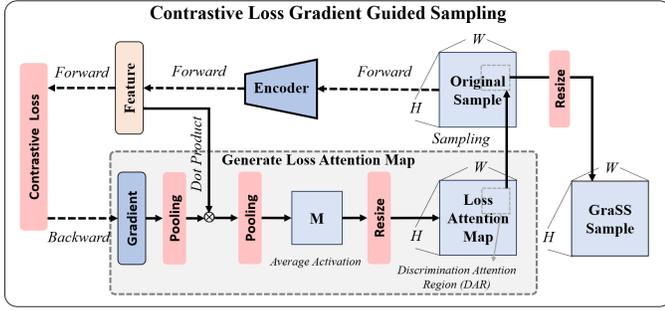


Fig. 3. Details of contrastive loss gradient guided sampling. The original samples are input into the model to obtain features and calculate the contrastive loss. Then, the contrastive loss is backpropagated to the feature layer to obtain the gradient to generate the loss attention map. Finally, the original samples were sampled according to the loss attention map to obtain the gradient-guided sampling samples.

$$\frac{\partial L}{\partial \mathbf{F}_{ij}} = \frac{\partial L}{\partial l_{ij}} \frac{\partial l_{ij}}{\partial \mathbf{f}_{ij}} \frac{\partial \mathbf{f}_{ij}}{\partial \mathbf{F}_{ij}} \quad (9)$$

where  $\frac{\partial L}{\partial l} = \frac{1}{N \cdot K}$  only related to the amount of original image data  $N$  and the number of copies  $K$  of the image copy function  $c(\cdot)$ . Then, we calculate the average activation  $M$  for the dot product of contrastive loss gradient and the feature  $\mathbf{F}$  and resize  $M$  to the same size as the image sample to obtain the contrastive Loss Attention Map (LAM). For the image instance  $\tilde{x}_{ij}$ , this process can be described as:

$$M = \frac{1}{D} \sum_{d=1}^D \text{pooling} \left( \frac{\partial L}{\partial \mathbf{F}_{ij}^d} \right) \mathbf{F}_{ij}^d, \quad (10)$$

$$\text{LAM} = \underset{H=h, W=w}{\text{Resize}} (M). \quad (11)$$

In the Eq. (10),  $\mathbf{F}_{ij}^d$  denotes the  $d$ -th dimensional component of the  $D$ -dimensional feature  $\mathbf{F}_{ij}$ .  $\text{pooling}(\cdot)$  denotes the global pooling operation applied to the gradients.  $\underset{H=h, W=w}{\text{Resize}}(\cdot)$  represents the resizing of the activation  $M$  into a two-dimensional activation map of height  $h$  and width  $w$ , where  $h$  and  $w$  correspond to the height and width of the input image instances.

Finally, we obtain the discriminative attention region (DAR) based on the contrastive Loss Attention Map (LAM). Specifically, we define the Discrimination Attention Region acquisition function  $G(\text{LAM}; T_A)$ , where  $T_A$  is the activation map threshold for selecting DAR. We regard the regions in LAM with a value higher than  $T_A$  as the candidate discrimination attention region  $R$ , calculate the maximum activation value of all candidate discrimination attention regions  $R$ , and select the candidate region with the highest maximum activation value as the Discrimination Attention Region (DAR). The above process can be described as:

$$\text{DAR} = G(\text{LAM}; T_A = t), \quad (12)$$

$$R = \{R_i\} = (\text{LAM} > t), \quad (13)$$

$$\max(\text{DAR}) = \max(\max(R_i)), \quad (14)$$

where  $R_i$  refers to the  $i$ -th 4-connected closed region in  $R$ . The rules for the operation of the function  $G$ , as described in Eq. (12), are defined by Eq. (13), and Eq. (14).

2) *Reconstruction of Positive and Negative Samples*: We reconstructed positive and negative samples based on the DAR. Specifically, we first obtain the coordinates of the centroid  $(x, y)$ , width  $w$ , and height  $h$  of the smallest outer rectangle of the DAR corresponding to the original image sample. Afterward, we crop the corresponding RSI region based on the coordinates and resize it to the original image size to obtain a new sample. We refer to the operation of cropping an image based on the DAR as DACrop. The above process can be described as:

$$x, y, h, w = \text{Box}(\text{DAR}_{ij}), \quad (15)$$

$$\tilde{x}'_{ij} = \underset{X=x, Y=y, H=h, W=w}{\text{DACrop}} (\tilde{x}_{ij}). \quad (16)$$

Finally, we input the updated image instance  $\tilde{x}_{ij}$  into the model to extract features, calculate the contrastive loss and update the model parameters.

The proposed GraSS can be described in Algorithm 1.

---

#### Algorithm 1 Pseudocode for GraSS

---

**Require:** RSI dataset  $X$ ; encoder  $E(\cdot)$ , project head  $P(\cdot)$ ; augmentation function  $T(\cdot)$ ; batch size  $N$ ; warm-up epoch  $e$ , threshold  $t$ , current model training epoch  $e_c$

- 1: **for** batch  $x$  from  $X$  **do**
- 2:   Data Augmentation:  $\tilde{x} = T(x)$
- 3:   Get features:  $\mathbf{f}_{ij} = P(E(\tilde{x}))$
- 4:   Calculate contrastive loss  $L$  using Equation (7) and (8)
- 5:   **if**  $e_c \leq e$  **then**
- 6:     Update  $E(\cdot)$  and  $P(\cdot)$
- 7:   **else**
- 8:     Get LAM using Equation (9), (10), and (11)
- 9:     Get DAR:  $\text{DAR} = G(\text{LAM}; T_A = t)$
- 10:    Construct sample  $\tilde{x}'_{ij}$  with DACrop using Equation (15) and (16)
- 11:    Get features:  $\mathbf{f}'_{ij} = P(E(\tilde{x}'_{ij}))$
- 12:    Calculate contrastive loss  $L$  using Equation (7) and (8)
- 13:    Update  $E(\cdot)$  and  $P(\cdot)$
- 14:   **end if**
- 15: **end for**

---

## IV. EXPERIMENT

### A. Experimental Setup

1) *Dataset*: We selected three high-resolution RSI semantic segmentation datasets Potsdam [54], LoveDA Urban, and LoveDA Rural [55] to evaluate the semantic segmentation performance of the self-supervised contrastive learning model on high-resolution RSIs. TABLE I shows more detailed information about the Potsdam, LoveDA Urban, and LoveDA Rural datasets.

TABLE I  
 DETAIL INFORMATION OF POTSDAM, LOVE DA URBAN, AND LOVE DA RURAL DATASET.

Dataset	Potsdam	LoveDA Urban	LoveDA Rural
Resolution (m)	0.05	0.3	0.3
Crop Size	256× 256	256× 256	256× 256
The amount of data for self-supervised pretraining	13824	18496	21856
The amount of data for semantic segmentation fine-tuning	138	184	218
The amount of data for semantic segmentation test dataset	8064	10832	15872

2) *Baselines*: We selected eight state-of-the-art methods from six different types of positive and negative sample construction as baselines to evaluate the performance of GraSS.

a) *Original Contrastive Learning Method*: We selected the representative SimCLR [20] and MoCo v2 [24] as the classical self-supervised contrastive learning baselines. Both SimCLR [20] and MoCo v2 [24] use the typical positive and negative sample construction methods, where different augmentations of the same image are treated as positive samples and augmentations of different images are treated as negative samples. The difference is that SimCLR constructs negative samples from other images in the same training batch, which limits the number of negative samples by the batch size [20], while MoCo v2 updates the negative samples by maintaining a queue momentum, and the number of negative samples is not limited by the number of samples in the training batch [23], [24].

b) *Contrastive Learning Method with Clustering*: We selected PCL [56] as a self-supervised contrastive learning baseline that introduces clustering to build positive and negative samples. Unlike SimCLR and MoCo v2, PCL introduces a clustering strategy to construct positive and negative samples based on data augmentation, which treats cluster centers of the same class of image clustering as positive samples and cluster centers of different classes of clusters as negative samples [56].

c) *Contrastive Learning Method without Negative Samples*: We selected Barlow Twins [49] and BYOL [57] as self-supervised contrastive learning baselines that do not construct negative samples and only construct positive samples. To avoid the model collapse caused by only bringing positive samples closer, BYOL uses an asymmetric network structure to project different augmentations of the same image into different feature spaces for comparison, while Barlow Twins does not directly pull positive samples closer, it only constrains the dimensions of sample features to be relatively independent.

d) *Negative Aware Contrastive Learning Method*: We selected FALSE [18] as a self-supervised negative aware contrastive learning baseline. FALSE adds a determination module to correct false negative samples into positive samples based on the classical positive and negative sample construction method, the positive samples are not only derived from different augmentations of the same image but also from images that contain the same ground objects as the positive samples [18].

e) *Positive Aware Contrastive Learning Method*: We selected ContrastiveCrop [43] as a self-supervised positive aware contrastive learning baseline. ContrastiveCrop uses the

activation information propagated forward from the image to the feature layer of the model to construct positive samples [43].

f) *Dense Contrastive Learning Method*: We selected DenseCL [28] as a self-supervised contrastive learning baseline with the dense contrastive module. DenseCL adds dense feature contrastive constraints to the instance contrastive. It gives the model the ability to capture certain object-level or pixel-level features of the image [28].

3) *Metrics*: We selected three metrics, the mean Intersection-over-Union (mIoU), the Overall Accuracy (OA), and the mean class Accuracy (mAcc) to quantitatively evaluate the performance of the self-supervised contrastive model on the test dataset for the downstream semantic segmentation task. The mIoU is a common metric for the semantic segmentation task, and for a single ground object, the intersection-over-Union (IoU) is defined by the following equation:

$$IoU = \frac{prediction \cap target}{prediction \cup target} \quad (17)$$

Where prediction refers to the predicted result of the model for the ground object, and target refers to the ground truth of the ground object. The mIoU is equal to the average of the IoU of all objects.

The OA represents the overall accuracy of the predicted result on the test dataset, which is defined by the following equation:

$$OA = \frac{TP}{N} \quad (18)$$

where the TP means the total number of pixels that are correctly predicted, and the N means the total number of pixels.

Slightly different from OA, mAcc is used to indicate the average level of accuracy of the predicted result for each ground object class. Specifically, the prediction Accuracy (Acc) for a single ground object class can be defined by the following equation:

$$Acc = \frac{TP_i}{N_i} \quad (19)$$

where the  $TP_i$  means the number of correctly predicted pixels for a specific ground object class and the  $N_i$  means the number of pixels for a specific ground object class in ground truth. The mAcc is the average of the Acc of all ground object classes.

TABLE II  
QUANTITATIVE COMPARISON RESULTS WITH EIGHT STATE-OF-THE-ART SELF-SUPERVISED CONTRASTIVE LEARNING BASELINE METHODS AND GLCNET.

Method	Pretraining Module	Potsdam			LoveDA Urban			LoveDA Rural			Time* (Training/Test)
		OA	mIoU	mAcc	OA	mIoU	mAcc	OA	mIoU	mAcc	
SimCLR	Encoder	61.18	43.02	54.90	41.91	33.09	45.31	<u>62.96</u>	<u>41.30</u>	<u>52.97</u>	3.0h/12min
MoCo v2	Encoder	60.21	42.81	54.53	40.61	32.92	45.77	58.01	36.28	47.84	2.5h/12min
PCL	Encoder	61.45	43.13	55.15	40.07	33.28	45.99	59.40	37.42	49.41	8.5h/12min
Barlow Twins	Encoder	61.42	43.17	54.95	<u>43.05</u>	<u>34.32</u>	<u>46.09</u>	56.29	33.93	51.55	7.5h/12min
BYOL	Encoder	<u>61.54</u>	43.93	55.73	35.18	28.49	39.94	60.21	37.39	48.34	5.0h/12min
FALSE	Encoder	60.65	43.12	55.45	42.44	33.69	46.08	62.44	40.91	51.82	3.5h/12min
ContrastiveCrop	Encoder	60.89	42.86	54.64	40.24	33.35	45.33	61.98	39.28	51.85	7.5h/12min
DenseCL	Encoder	61.44	<u>44.34</u>	<u>56.40</u>	37.01	30.85	42.07	62.85	38.69	49.62	5.5h/12min
<b>GraSS(Ours)</b>	<b>Encoder</b>	<b>62.28</b>	<b>44.39</b>	<b>56.50</b>	<b>43.68</b>	<b>34.77</b>	<b>46.77</b>	<b>65.25</b>	<b>42.58</b>	<b>53.79</b>	4.5h/12min
GLCNet	Encoder+Decoder	77.91	60.57	75.26	50.78	41.26	56.55	63.36	40.41	53.36	8.5h/12min
<b>GraSS(Ours)</b>	<b>Encoder+Decoder</b>	<b>78.79</b>	<b>61.41</b>	<b>75.62</b>	<b>52.60</b>	<b>42.05</b>	<b>57.82</b>	<b>65.65</b>	<b>40.96</b>	<b>57.62</b>	9.5h/12min

\* The training time is measured on the Potsdam training dataset using an A800 GPU for 350 epochs, and the test time is measured on the Potsdam test dataset using an RTX 3090 GPU for 150 epochs.

4) *Implementation Details*: For both the eight self-supervised contrastive learning baselines and the proposed GraSS, we used ResNet50 [58] as the backbone network.

In the self-supervised pretraining stage, we train 350 epochs using the entire training dataset without labels, and the batch size is set to 256. For each baseline method, we use the data augmentation and optimization settings recommended in the original paper, and all self-supervised pretraining methods were used to train the feature extractor only. For the proposed GraSS, the number of instance discrimination warm-up training epochs is also included in the total number of pretraining epochs for a fair comparison with baselines.

In the RSI semantic segmentation fine-tuning stage, in order to accurately evaluate the performance of the features extracted from different self-supervised pretraining methods, we freeze the weights of the entire backbone network and update only the parameters of the feature decoder used to obtain the RSI semantic segmentation results. We randomly select 1% of the entire training dataset for fine-tuning training. For the eight baseline methods and proposed GraSS, the randomly selected fine-tuning data were kept consistent. We uniformly use the Stochastic Gradient Descent (SGD) optimizer [59] for fine-tuning training for 150 epochs with the batch size set to 16.

## B. Experimental Result

In this section, we present five aspects of evaluation for the proposed GraSS.

First is performance analysis, we compare the proposed GraSS with six types, a total of eight self-supervised contrastive learning baseline methods on the Potsdam, LoveDA Urban, and LoveDA Rural datasets, and provide both quantitative and qualitative analysis results. In order to explore the applicability of the gradient guided sampling strategy, unlike the eight baseline methods mentioned above that train only the feature extractor in the pre-training stage, we also compare with GLCNet [16], which requires a specified RSI

TABLE III  
THE SEMANTIC SEGMENTATION RESULTS OF GLCNET WITH GRADIENT GUIDED SAMPLING STRATEGY.

Dataset and Metric		GLCNet	
		w/o Gradient Guided	w/ Gradient Guided
Potsdam	Kappa	71.77	<b>72.79</b>
	OA	77.91	<b>78.79</b>
	mIoU	60.57	<b>61.41</b>
	mAcc	75.26	<b>75.62</b>
LoveDA Urban	Kappa	42.08	<b>43.23</b>
	OA	50.78	<b>52.60</b>
	mIoU	41.26	<b>42.05</b>
	mAcc	56.55	<b>57.82</b>
LoveDA Rural	Kappa	49.22	<b>50.45</b>
	OA	63.36	<b>65.65</b>
	mIoU	40.41	<b>40.96</b>
	mAcc	53.36	<b>57.62</b>

TABLE IV  
COMPARISON OF SEMANTIC SEGMENTATION RESULTS BETWEEN GRASS AND TWO STATE-OF-THE-ART SAMPLING METHOD.

Dataset and Metric		Sampling Method			
		Original	ContrastiveCrop	LCR	<b>Ours</b>
Potsdam	OA	60.21	60.89	61.26	<b>62.28</b>
	mIoU	42.81	42.86	43.88	<b>44.39</b>
	mAcc	54.53	54.64	55.52	<b>56.50</b>
LoveDA Urban	OA	40.61	40.24	42.50	<b>43.68</b>
	mIoU	32.92	33.35	33.53	<b>34.77</b>
	mAcc	45.77	45.33	45.97	<b>46.77</b>
LoveDA Rural	OA	58.01	61.98	64.85	<b>65.25</b>
	mIoU	36.28	39.28	40.39	<b>42.58</b>
	mAcc	47.84	51.85	50.26	<b>53.79</b>

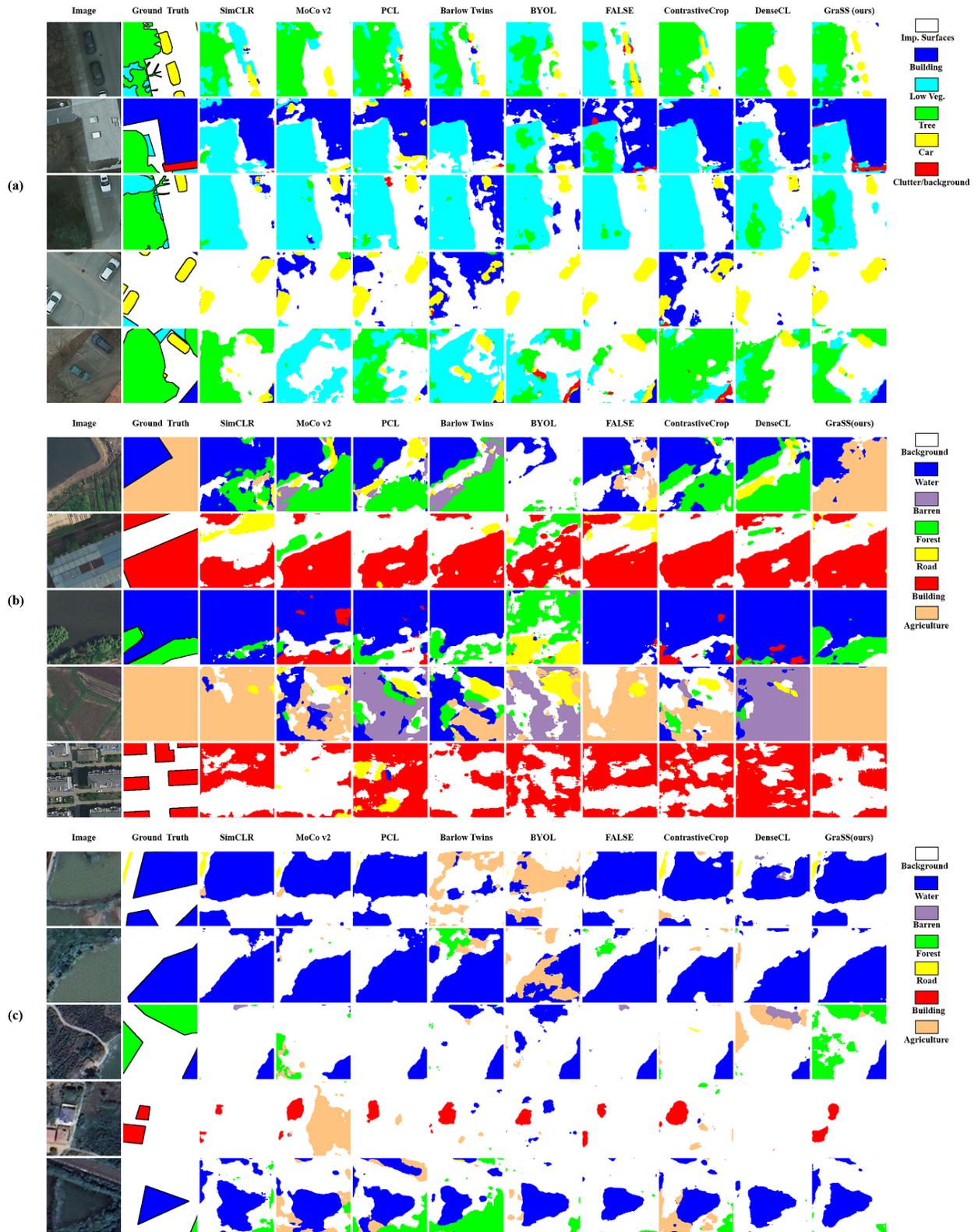


Fig. 4. Qualitative comparison results with eight self-supervised contrastive learning baseline methods. Where the five RSIs in a) are from the test dataset of Potsdam, the five RSIs in b) are from the test dataset of LoveDA Urban, and the five RSIs in c) are from the test dataset of LoveDA Rural.

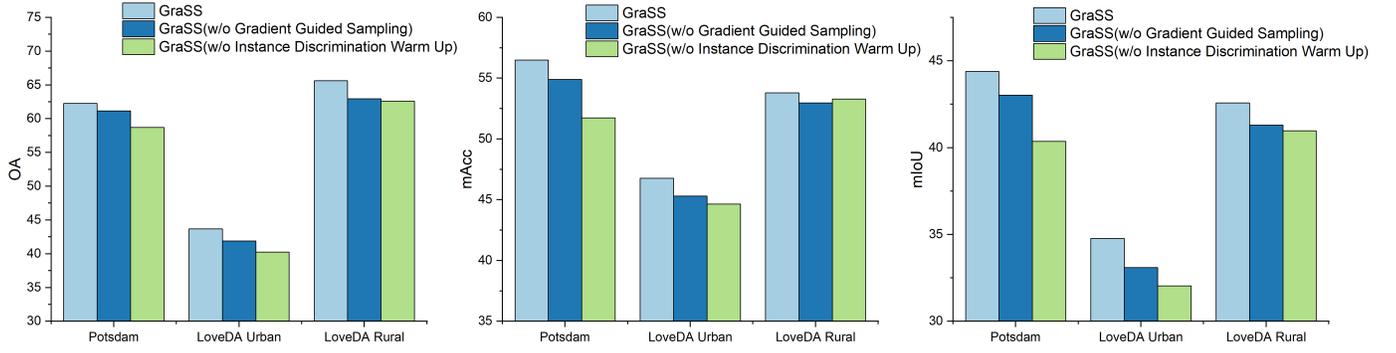


Fig. 5. Results of the ablation study exploring the effectiveness of each module of the proposed GraSS.

semantic segmentation decoder in the pre-training stage. In addition, to further validate the performance of the proposed GraSS, we also compare GraSS with two sampling methods, ContrastiveCrop [43] and LCR [45].

Second is the ablation study, we perform ablation experiments to verify the effectiveness of each module of the proposed GraSS.

Third is performance sensitivity analysis, we examine the effects of two hyperparameters corresponding to the two training stages of GraSS: the instance discrimination warm-up epoch and the threshold  $T_A$  of the activation map for selecting DAR on the RSI semantic segmentation performance of the model. The instance discrimination warm-up epoch hyperparameter corresponds to the instance discrimination warm-up stage, and the threshold of activation map hyperparameter corresponds to the gradient guided sampling contrastive training stage.

Fourth is the analysis of the number of ground objects contained in the sample. We quantitatively evaluated the number of ground objects contained in the samples and found that positive and negative samples obtained by GraSS contain more singular ground object types compared to the samples obtained by original self-supervised contrastive methods.

Finally is the visual analysis of the contrastive Loss Attention Map (LAM). We examined the effect of the warm-up epoch and the ground objects contained in RSI on the LAM.

#### 1) Performance analysis:

a) *Quantitative Analysis:* To evaluate the performance of the proposed GraSS, we first compared it with six types, a total of eight self-supervised contrastive learning baseline methods and GLCNet [16], which requires a specified RSI semantic segmentation decoder in the pre-training stage, and the experiment results are shown in TABLE II. In addition, in order to explore the applicability of the gradient guided sampling strategy, we align the experimental conditions and metrics of GraSS with GLCNet to show the performance improvement of GLCNet with the gradient guided strategy, and the experimental results are shown in TABLE III. Finally, in order to further verify the performance of the proposed GraSS, we also compared GraSS with the two sampling methods: ContrastiveCrop [43] and Learning Common Rationale (LCR) [45], and the experimental results are shown in TABLE IV.

TABLE II shows that GraSS achieves the best results on all three metrics for the three datasets compared to the eight self-

supervised contrastive learning baseline methods of SimCLR, MoCo v2, PCL, Barlow Twins, BYOL, FALSE, Contrastive-Crop, and DenseCL. On the Potsdam dataset, GraSS only slightly outperformed DenseCL, but on the LoveDA Urban dataset, GraSS is 3.85% higher than DenseCL in terms of mIoU, and on the LoveDA Rural dataset, GraSS is 3.89% higher than DenseCL in terms of mIoU, this demonstrates the stable semantic segmentation performance improvement of GraSS.

In addition, compared to the original self-supervised contrastive learning method SimCLR, we observed that MoCo v2, PCL, Barlow Twins, BYOL, and DenseCL show significant performance degradation on the LoveDA Rural dataset, with DenseCL and BYOL also showing significant performance degradation on the LoveDA Urban dataset, exhibiting unstable semantic segmentation performance.

TABLE III shows the experimental results of using the gradient guided sampling strategy for the GLCNet that requires a specified semantic segmentation decoder in the self-supervised pretraining stage. The experimental results indicate that the gradient guided sampling strategy further improves the semantic segmentation performance of the GLCNet. Meanwhile, It indicates that the gradient guided sampling strategy is also applicable to the self-supervised contrastive learning method that trains both the feature extractor and the semantic segmentation decoder in the pretraining stage.

TABLE IV shows the comparison results of semantic segmentation performance of the proposed GraSS with Original, ContrastiveCrop, and LCR, where Original represents the original self-supervised contrastive learning method. The experimental results indicate that on Potsdam, LoveDA Urban, and LoveDA Rural datasets, compared with Original, Contrastive-Crop, and LCR, the GraSS achieves the best performance on three indicators. Specifically, on the LoveDA Rural dataset, the proposed GraSS improves the mIoU by 2.19% and the mAcc by 3.53% compared with the best baseline.

b) *Qualitative Analysis:* We qualitatively analyzed the visualization results of semantic segmentation of RSIs, and the experimental results are shown in Fig. 4. The experimental results show that the semantic segmentation results of GraSS present richer details than the eight self-supervised contrastive learning baselines, especially for small-scale ground objects in the RSIs, which are difficult to be captured by the instance-level self-supervised contrastive learning methods. In addition, the GraSS also presents a more stable semantic segmentation

effect for RSIs containing a large range of homogeneous ground objects.

Quantitative and qualitative experimental results indicate that the proposed GraSS effectively improves the performance on RSI semantic segmentation tasks, and performs better in both quantitative and qualitative aspects. This is because the proposed GraSS fully utilizes the discriminative information in the contrastive loss gradient to construct samples containing more singular ground objects, which effectively alleviates the positive sample confounding issue in the process of contrastive learning. Meanwhile, the model can benefit from the contrastive of samples containing a single ground object, and obtain more accurate features of ground objects.

Although the proposed GraSS effectively alleviates the positive sample confounding issue of self-supervised contrastive learning in RSI semantic segmentation tasks compared with other methods, since the process of constructing positive and negative samples is unsupervised, the proposed GraSS cannot absolutely guarantee that the obtained samples only contain a single type of ground object, and cannot completely eliminate the positive sample confounding issue.

2) *Ablation study*: In order to explore the effectiveness of each module of the proposed GraSS, we conducted an ablation study on three datasets: Potsdam, LoveDA Urban, and LoveDA Rural. Three sets of experiments were conducted, the first set of experiments used the complete GraSS, the second set of experiments removed the gradient guided sampling module on the basis of the proposed GraSS, and the third set of experiments removed the instance discriminative warm-up training module on the basis of the GraSS. The experimental results are shown in Fig. 5.

The experimental results in Fig. 5 show that each module of GraSS improves the model performance. Compared with complete GraSS and the GraSS without gradient guided sampling, GraSS without instance discrimination warm-up demonstrates the worst performance on the above three datasets. This indicates that it is necessary to conduct instance discrimination warm up in the initial stage of the model training and train the model to obtain the initial instance discrimination ability. The instance discrimination warm up gives discriminative information to the contrastive loss gradient, which can help the gradient-guided sampling to further improve the performance of the model.

In addition, compared with the complete GraSS, the performance of the GraSS without gradient guided sampling decreases in all indicators of the three datasets, which indicates that the gradient guided sampling contrastive training can effectively improve the semantic segmentation performance of the model after the instance discrimination warm up.

3) *Performance sensitivity analysis*: In this section, we examine the effects of two hyperparameters of GraSS: the instance discrimination warm-up epochs and the threshold of the activation map for selecting DAR on the performance of semantic segmentation. The instance discrimination warm-up epochs hyperparameter corresponds to the first training stage of GraSS and the threshold of the activation map for selecting the DAR hyperparameter corresponds to the second training stage of GraSS.

a) *Analysis of the Instance Discrimination Warm-Up Epochs*: The self-supervised contrastive learning methods with gradient guided sampling rely on the instance discrimination ability of the model. Therefore, we analyze the impact of the warm-up training epochs on the semantic segmentation performance of the GraSS. We fixed the gradient guided sampling training epochs to 150, and selected six instance discrimination warm-up epochs of 0, 50, 100, 150, and 200 on the Potsdam, LoveDA Urban, and LoveDA Rural datasets for comparison.

Fig. 6 shows the semantic segmentation results for the GraSS with six different warm-up epochs. The experimental results show that for the Potsdam, LoveDA Rural, and LoveDA Urban datasets, when the instance discrimination warm-up epoch is increased to 200, the semantic segmentation performance of the GraSS still obtains an improvement, although it fluctuates slightly. Among them, the semantic segmentation performance of the LoveDA Urban dataset fluctuates significantly, and for the Potsdam and LoveDA Rural datasets, the best semantic segmentation OA was obtained when the warm-up epoch is 200.

b) *Analysis of the Threshold of the Activation Map for Selecting DAR*: To analyze the effect of the threshold  $T_A$  of the activation map for selecting DAR on the semantic segmentation performance of the model, we selected five thresholds of 0, 0.3, 0.5, 0.7 and 0.9 for comparison experiments on the Potsdam, LoveDA Urban, and LoveDA Rural datasets.

The threshold of 0 indicates that no gradient guided sampling is performed, and a larger threshold indicates that a smaller sampling region is obtained. The results in Fig. 7 show that gradient guided sampling effectively improves the semantic segmentation performance of the contrastive learning model and achieves the best semantic segmentation performance at a threshold value of 0.5. However, too large threshold  $T_A$  will cause the model to select too small RSI regions, resulting in the constructed samples containing too little or missing ground object information, which causes a degradation of the semantic segmentation performance.

4) *Analysis of the Number of Ground Objects Contained in the Sample*: To analyze the number of ground objects contained in the samples obtained by GraSS, we use the label information to count the number of ground objects contained in the sample by gradient guided sampling strategy (GraSS) on Potsdam, LoveDA Urban, and LoveDA Rural datasets and compared it with the original RSI sample and random resize crop. In this analysis experiment, the batch size was set to 32, and the gradient guided sampling training was started after 150 epochs of instance discrimination warm-up and continued to be observed for 50 epochs. The experimental results are shown in Fig. 8.

We first analyzed the number of ground objects contained in the average single sample in the first batch of each training epoch, and the experimental results are shown in the first column of Fig. 8. The experimental results show that the samples obtained by GraSS contain the lowest number of ground objects compared to the original RSI and random resize crop in the 50 observed epochs, which indicates that GraSS effectively reduces the number of ground objects contained in

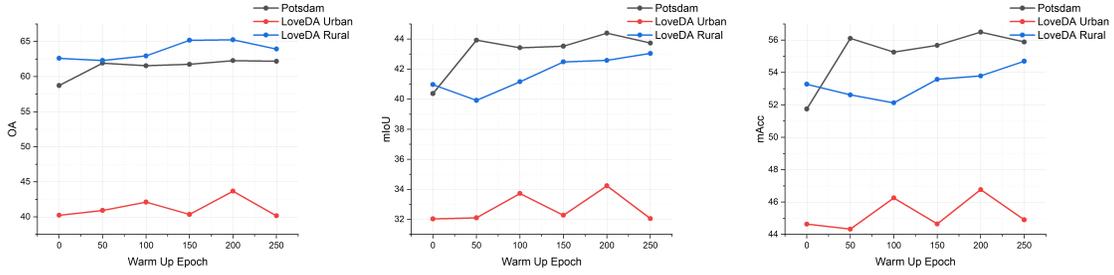


Fig. 6. Semantic segmentation results with 6 different instance discrimination warm-up epochs.

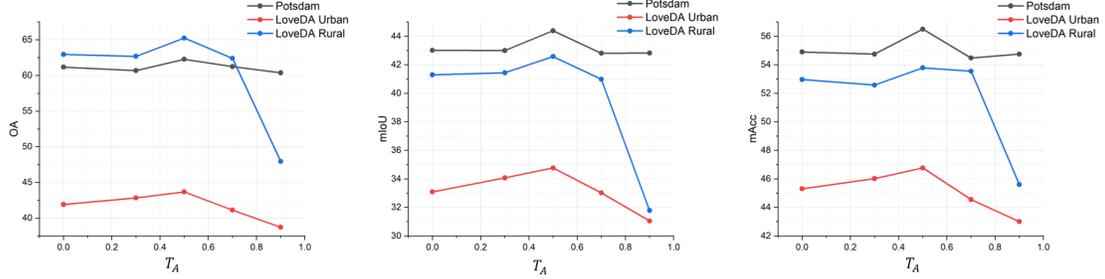


Fig. 7. Semantic segmentation results with 5 different threshold  $T_A$  of the activation map for selecting DAR.

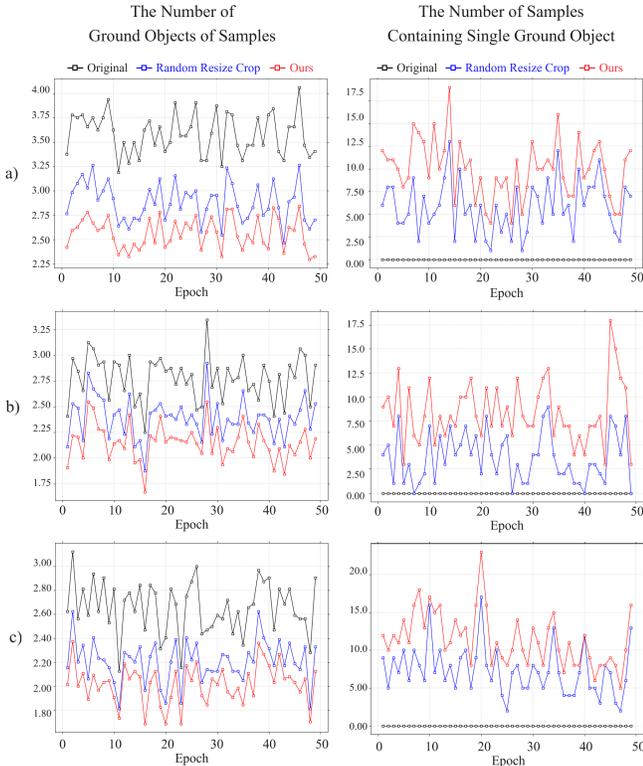


Fig. 8. Result of analysis of the Number of Ground Objects Contained in the Sample. The first-row a) shows the results on the Potsdam dataset, the second-row b) shows the results on the LoveDA Urban dataset, and the third-row c) shows the results on the LoveDA Rural dataset.

the original RSI samples.

In addition, we also analyzed the number of samples containing single ground objects in the first batch of each

training epoch, and the experimental results are shown in the second column of Fig. 8. The experimental results show that GraSS obtains the highest number of samples containing single ground objects compared to the original RSI and the random resize crop in the 50 observed epochs. For the Potsdam dataset, GraSS obtained a maximum of 19 samples containing single ground objects compared to the original RSI, for the LoveDA Urban dataset, GraSS obtained a maximum of 18 samples containing single ground objects compared to the original RSI, and for the LoveDA Rural dataset, GraSS obtained a maximum of 23 samples containing single ground objects compared to the original RSI. This indicates that GraSS can obtain more samples containing single ground objects, effectively mitigating positive sample confounding issue and feature adaptation bias.

5) Visual analysis of contrastive Loss Attention Map (LAM):

a) Effect of Instance Discrimination Warm-Up Epoch on the LAM: To explore the effect of instance discrimination warm-up epoch on the contrastive Loss Attention Map (LAM), we visualized the LAM obtained from different instance discrimination warm-up epochs, and the experimental results are shown in Fig. 9.

We selected 5 instance discrimination warm-up epochs of 0, 50, 100, 150, and 200 for visualization and analysis of the LAM. The experimental results show that as the instance discrimination warm-up proceeds, the obtained LAMs gradually focus on a certain region of the RSI, which tends to contain more singular ground objects in RSI.

b) Effect of Ground Objects Contained in RSI on the LAM: To explore the effect of ground objects contained in RSI on LAM, we artificially constructed different batches of image data and acquired LAM for visualization, and the batch size is set to 8. Specifically, we first specify the anchor sample images

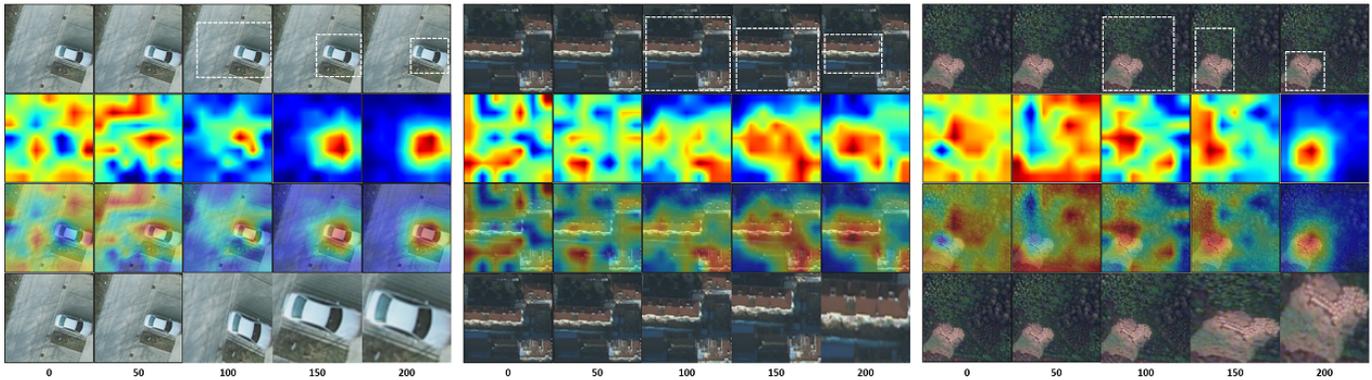


Fig. 9. The result of the effect of instance discrimination warm-up Epoch on LAM. The first row indicates the original RSI, the white dashed box is the cropped area obtained when the threshold  $T_A$  is set to 0.5, the second row is the LAM corresponding to the RSI, the third row shows the superimposed results of the RSI and the corresponding LAM, the fourth row is the samples reconstructed from GraSS after instance discrimination warm-up, and the number in the last row indicates the corresponding instance discrimination warm-up epoch.

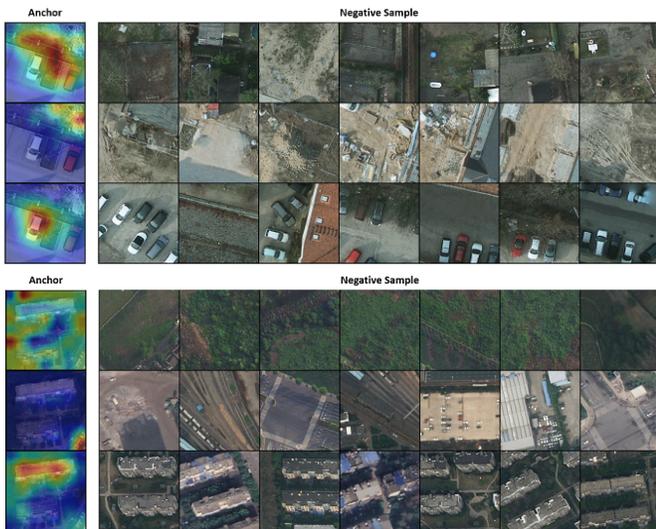


Fig. 10. The experiment result of the effect of ground objects contained in RSI on LAM. Each row represents a batch of remote sensing images that are input to the model. Among them, the anchor samples in the first three rows are the same, and the negative samples are different. The anchor samples in the last three rows are the same, and the negative samples are different.

and observe the changes in the anchor sample’s LAM by replacing other images in the same batch. Where other images in the same batch are considered as negative samples by self-supervised contrastive learning, the experimental results are shown in Fig. 10.

The experimental results show that the regions with higher activation values in LAM tend to be concentrated on a relatively large number of ground objects in the same batch. As shown in the first to third rows of Fig. 10, for the same anchor sample image, when other images in the same batch contain a high number of grass or low vegetation, the regions with higher activation values in the anchor sample’s LAM are concentrated in the low vegetation area (as shown in the first row of Fig. 10). When other images in the same batch contain a high number of clutter, the regions with higher activation values in the anchor sample’s LAM are concentrated in the clutter area (as shown in the second row of Fig. 10). When other images in the same batch contain a high number of cars, the regions

with higher activation values in the anchor sample’s LAM are concentrated in the car area (as shown in the third row of Fig. 10). And the fourth to sixth rows of Figure 9 show similar results.

A possible explanation for such a result is that when ground objects in the negative sample are close to those in the anchor sample, a larger contrastive loss gradient will be generated, which leads to a higher activation value of the image area corresponding to the LAM obtained from the contrastive loss gradients.

## V. CONCLUSION

In this paper, we propose contrastive learning with Gradient guided Sampling Strategy (GraSS) for RSI semantic segmentation. It uses the positive and negative sample discrimination information contained in the self-supervised contrastive loss gradients to construct samples containing more singular ground objects, alleviate the sample confounding issue of semantic segmentation of RSIs for self-supervised contrastive learning, and mitigate the feature adaptation bias between instance-level pretext task and pixel-level RSI semantic segmentation tasks. The experiments show that the GraSS effectively improves the performance of the self-supervised contrastive learning model for the RSI semantic segmentation task and outperforms a total of eight self-supervised contrastive learning methods of six types at present. In addition, we have conducted extensive experiments and probes on the proposed GraSS and preliminarily discussed the effects of two factors, the instance discrimination warm-up epoch and the ground objects contained in RSI, on the LAM obtained by contrastive loss gradients, which is expected to deepen our understanding of self-supervised contrastive learning models.

Although the current gradient guided sampling strategy effectively mitigates the positive sample confounding issue of self-supervised contrastive learning for the RSI semantic segmentation task, since self-supervised contrastive learning is essentially unsupervised, our proposed GraSS cannot absolutely guarantee that the obtained sample contains only a single type of ground objects.

In addition, we found that the contrastive loss gradient contains rich feature information, which inspires us to make

more use of the gradient information in the process of model training to obtain additional model capabilities. However, we currently lack a clear understanding of several factors that affect the contrastive Loss Attention Map (LAM) obtained by the contrastive loss gradients. In the future, we will further explore the relationship between the contrastive loss gradient and the spatio-temporal characteristics of RSI, which may provide guidance for designing a self-supervised contrastive learning model that can capture the features of RSIs more effectively.

#### ACKNOWLEDGMENT

The authors would like to thank the valuable comments from anonymous reviewers.

#### REFERENCES

- [1] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1182–1191.
- [2] C. Tao, J. Qi, W. Lu, H. Wang, and H. Li, "Remote sensing image scene classification with self-supervised paradigm under limited labeled samples," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.
- [3] H. Huang, Z. Mou, Y. Li, Q. Li, J. Chen, and H. Li, "Spatial-temporal invariant contrastive learning for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
- [4] X. Wang, J. Zhu, Z. Yan, Z. Zhang, Y. Zhang, Y. Chen, and H. Li, "Last: Label-free self-distillation contrastive learning with transformer architecture for remote sensing image scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
- [5] D. Ye, J. Peng, H. Li, and L. Bruzzone, "Better memorization, better recall: A lifelong learning framework for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
- [6] L. Zhao, W. Luo, Q. Liao, S. Chen, and J. Wu, "Hyperspectral image classification with contrastive self-supervised learning under limited labeled samples," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
- [7] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, "Hyperspectral and lidar data classification based on structural optimization transmission," *IEEE Transactions on Cybernetics*, 2022.
- [8] Y. Zhang, W. Li, W. Sun, R. Tao, and Q. Du, "Single-source domain expansion network for cross-scene hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 32, pp. 1498–1512, 2023.
- [9] Y. Zhang, W. Li, M. Zhang, Y. Qu, R. Tao, and H. Qi, "Topological structure and semantic information transfer network for cross-scene hyperspectral image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [10] Y. Zhang, M. Zhang, W. Li, S. Wang, and R. Tao, "Language-aware domain generalization network for cross-scene hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.
- [11] Y. Zhang, W. Li, M. Zhang, S. Wang, R. Tao, and Q. Du, "Graph information aggregation cross-domain few-shot learning for hyperspectral image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [12] E. Xie, J. Ding, W. Wang, X. Zhan, H. Xu, P. Sun, Z. Li, and P. Luo, "Detco: Unsupervised contrastive learning for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8392–8401.
- [13] C. Yang, Z. Wu, B. Zhou, and S. Lin, "Instance localization for self-supervised detection pretraining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3987–3996.
- [14] S. Saha, P. Ebel, and X. X. Zhu, "Self-supervised multisensor change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2022.
- [15] Y. Chen and L. Bruzzone, "A self-supervised approach to pixel-level change detection in bi-temporal rs images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.
- [16] H. Li, Y. Li, G. Zhang, R. Liu, H. Huang, Q. Zhu, and C. Tao, "Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.
- [17] D. Muhtar, X. Zhang, and P. Xiao, "Index your position: A novel self-supervised learning method for remote sensing images semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.
- [18] Z. Zhang, X. Wang, X. Mei, C. Tao, and H. Li, "False: False negative samples aware contrastive learning for semantic segmentation of high-resolution remote sensing image," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
- [19] C. Tao, J. Qi, G. Zhang, Q. Zhu, W. Lu, and H. Li, "Tov: The original vision model for optical remote sensing image understanding via self-supervised learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.
- [20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [21] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.
- [22] H. Li, Z. Cui, Z. Zhu, L. Chen, J. Zhu, H. Huang, and C. Tao, "Rsmetanet: Deep meta metric learning for few-shot remote sensing scene classification," *arXiv preprint arXiv:2009.13364*, 2020.
- [23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
- [24] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.
- [25] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9640–9649.
- [26] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6707–6717.
- [27] C. Tao, J. Qi, M. Guo, Q. Zhu, and H. Li, "Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
- [28] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li, "Dense contrastive learning for self-supervised visual pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3024–3033.
- [29] P. O. O. Pinheiro, A. Almahairi, R. Benmalek, F. Golemo, and A. C. Courville, "Unsupervised learning of dense visual representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4489–4500, 2020.
- [30] M. Kim, J. Tack, and S. J. Hwang, "Adversarial self-supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2983–2994, 2020.
- [31] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "Map-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 6169–6181, 2020.
- [32] H. Li, K. Qiu, L. Chen, X. Mei, L. Hong, and C. Tao, "Scattnet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 5, pp. 905–909, 2020.
- [33] T. Huynh, S. Kornblith, M. R. Walter, M. Maire, and M. Khademi, "Boosting contrastive self-supervised learning with false negative cancellation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2785–2795.
- [34] T.-S. Chen, W.-C. Hung, H.-Y. Tseng, S.-Y. Chien, and M.-H. Yang, "Incremental false negative detection for contrastive learning," in *International Conference on Learning Representations*, 2022.
- [35] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 42–62, 2022.
- [36] K. Yang, T. Zhou, X. Tian, and D. Tao, "Identity-disentangled adversarial augmentation for self-supervised learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 364–25 381.

- [37] J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo, “Predicting what you already know helps: Provable self-supervised learning,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 309–323, 2021.
- [38] T. Wang and P. Isola, “Understanding contrastive representation learning through alignment and uniformity on the hypersphere,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 9929–9939.
- [39] J. Ding, E. Xie, H. Xu, C. Jiang, Z. Li, P. Luo, and G.-S. Xia, “Deeply unsupervised patch re-identification for pre-training object detectors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [40] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, “Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 684–16 693.
- [41] F. Wei, Y. Gao, Z. Wu, H. Hu, and S. Lin, “Aligning pretraining for detection via object-level contrastive learning,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 682–22 694, 2021.
- [42] J. Chen, H. Huang, J. Peng, J. Zhu, L. Chen, C. Tao, and H. Li, “Contextual information-preserved architecture learning for remote-sensing scene classification,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.
- [43] X. Peng, K. Wang, Z. Zhu, M. Wang, and Y. You, “Crafting better contrastive views for siamese representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 031–16 040.
- [44] A. Ziegler and Y. M. Asano, “Self-supervised learning of object parts for semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 502–14 511.
- [45] Y. Shu, A. van den Hengel, and L. Liu, “Learning common rationale to improve self-supervised representation for fine-grained visual recognition problems,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 392–11 401.
- [46] J. D. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, “Contrastive learning with hard negative samples,” in *International Conference on Learning Representations*, 2021.
- [47] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, “What makes for good views for contrastive learning?” *Advances in neural information processing systems*, vol. 33, pp. 6827–6839, 2020.
- [48] O. Henaff, “Data-efficient image recognition with contrastive predictive coding,” *International Conference on Machine Learning*, pp. 4182–4192, 2020.
- [49] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, “Barlow twins: Self-supervised learning via redundancy reduction,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 310–12 320.
- [50] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, “Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9414–9423.
- [51] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
- [52] X. Wang, S. You, X. Li, and H. Ma, “Weakly-supervised semantic segmentation by iteratively mining common object features,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1354–1362.
- [53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
- [54] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, and U. Breitkopf, “The isprs benchmark on urban object classification and 3d building reconstruction,” *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1*, vol. 1, no. 1, pp. 293–298, 2012.
- [55] J. Wang, Z. Zheng, X. Lu, and Y. Zhong, “Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021, pp. 1–16.
- [56] J. Li, P. Zhou, C. Xiong, and S. Hoi, “Prototypical contrastive learning of unsupervised representations,” in *International Conference on Learning Representations*, 2020.
- [57] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, “Bootstrap your own latent—a new approach to self-supervised learning,” *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.
- [58] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [59] M. Hardt, B. Recht, and Y. Singer, “Train faster, generalize better: Stability of stochastic gradient descent,” in *International conference on machine learning*. PMLR, 2016, pp. 1225–1234.