Remote Sensing Images Semantic Segmentation with General Remote Sensing Vision Model via a Self-Supervised Contrastive Learning Method

Haifeng Li, Yi Li, Guo Zhang, Ruoyun Liu, Haozhe Huang, Qing Zhu, Chao Tao

Abstract—Recently, supervised deep learning has achieved great success in remote sensing images (RSIs) semantic segmentation. However, supervised learning for semantic segmentation requires a large number of labeled samples which is difficult to obtain in the field of remote sensing. A new learning paradigm, self-supervised learning (SSL), can be used to solve such problems by pre-training a general model with large unlabeled images and then fine-tuning on a downstream task with very few labeled samples. Contrastive learning is a typical method of SSL, which can learn general invariant features. However, most of the existing contrastive learning is designed for classification tasks to obtain a image-level representation, which may be sub-optimal for semantic segmentation tasks requiring pixel-level discrimination. Therefore, we propose Global style and Local matching Contrastive Learning Network (GLCNet) for remote sensing semantic segmentation. Specifically, 1) The global style contrastive module is used to learn a image-level representation better, as we consider the style features can better represent the overall image features. 2) The local features matching contrastive module is designed to learn representations of local regions which is beneficial for semantic segmentation. We evaluate on four remote sensing semantic segmentation datasets, and the experimental results show that our method mostly outperforms state-of-the-art self-supervised methods and ImageNet pre-training. Specifically, with 1% annotation from the original dataset, our approach improves Kappa by 6% on the ISPRS Potsdam dataset and 3% on Deep Globe Land Cover Classification dataset relative to the existing baseline. Moreover, our method outperforms supervised learning when there are some differences between the datasets of upstream tasks and downstream tasks. Our study promotes the development of self-supervised learning in the field of remote sensing semantic segmentation. Since SSL could directly learn the essential characteristics of data from unlabeled data which is easy to obtain in remote sensing filed, this may be of great significance for tasks such as global mapping.

Index Terms—self-supervised learning, contrastive learning, remote sensing images semantic segmentation.

I. INTRODUCTION

WITh the development of remote sensing techniques, high-resolution satellite images are getting easily to be obtained. Remote sensing images are widely used in Urban Planning, Disaster Monitoring, Environment Protection, Agricultural Management, etc. [1]–[3]. The extraction and recognition of information from remote sensing images is the basis of these applications. Semantic segmentation as a pixel-level image analysis technology is one of the most important and challenging research directions in remote sensing image interpretation filed [4].

The traditional remote sensing semantic segmentation algorithms are mostly machine learning approaches based on hand-crafted features, such as support vector machine (SVM, [5, 6]), random forest (RF, [7]) and artificial neural networks (ANN, [8]). Since AlexNet [9] won the ILSVR champion in 2012, deep learning especially deep convolution neural network (DCNN) has attracted more and more attention [10]–[12]. Compared with traditional methods, deep learning is completely data-driven, and it can extract more abstract high-level features, achieving remarkable results on image classification tasks [13]. Subsequently, full convolution networks based approaches, such as FCN [13], U-Net [15], Deeplab series [16]–[18], almost dominate the field of computer vision image semantic segmentation. In remote sensing, researchers have improved the general semantic segmentation network, taking into account the specific characteristics of remote sensing, which further improves the accuracy of remote sensing semantic segmentation tasks [19], [20]. For example, Hammadi et al. [21] design a new Full Convolutional Network (FCN) architecture specifically for the classification of wetland complexes using polarimetric SAR (PolSAR). Ding et al. [22] focus on the problem of remote sensing images (RSIs) semantic segmentation with large-size. To better exploit global context information in RSIs, they propose a two-stage semantic segmentation network, by scaling the image at different sizes to obtain global contextual information and local detail information respectively, and then fusing the features to improve the accuracy.

However, the deep learning based supervised remote sensing semantic segmentation methods relies heavily on a large amount of high-quality labeled samples. As remote sensing semantic segmentation technology plays an increasingly important role in global sustainable development, its desire for a large number of global-wise, high-quality labeled samples is growing [23], [24]. The semantic segmentation task requires pixel-level labeling which is very costly, and remote sensing images have huge heterogeneity in time and space, so the existing labeled data is actually only a interception of the images, and it is difficult to obtain a large number of annotated
samples with extremely high richness which cover global areas, multi-resolution, multi-season and multi-spectral. To solve the problem of insufficient labeled samples, one strategy is to generate more samples through data augmentation [25], GAN [26], etc.; a second strategy is to use other annotated data, such as pre-training [27] or transfer learning [28], [29], which aims to transfer the knowledge learned from a larger or more related domain to improve performance on the target domain or reduce the dependence on labeled samples; another strategy is to learn how to have better performance on only a few labeled samples, such as meta-learning [30]. However, all the above methods are based on the paradigm of supervised learning, which is highly related to specific tasks and datasets, and it is impossible to obtain a general model. For example, transfer learning may have negative transfer when the difference between the source domain and the target domain is large [31]. In addition, these methods do not make use of the abundant unlabeled data.

Self-supervised learning provides a new paradigm, as shown in fig. 1 which first learns knowledge from unlabeled image data by designing self-supervised signals, and then transfer to downstream tasks, achieving comparable performance to supervised learning on downstream tasks with just limited labeled samples [32]. While large amounts of labeled data are not available, unlabeled image data covering the whole world with great diversity and richness are readily accessible, and the information contained in the image data is much richer than sparse labels, so we can expect to learn potentially more general knowledge through self-supervised learning.

In this work, we focus on contrastive learning which is a typical and successful self-supervised method [33]–[35]. We introduce the self-supervised contrastive learning paradigm into the remote sensing semantic segmentation task. In the pre-training stage, we use contrastive learning to enhance the consistency of the sample on the label-free data to learn a general remote sensing vision model (G-RSvM). G-RSvM enhances invariance, such as illumination invariance, rotation invariance, scale invariance, etc. In addition, the previous contrastive learning is mainly designed for image classification tasks, which only focus on the learning of image-level representations. However, there is a balance between global feature learning and local feature learning for remote sensing semantic segmentation tasks: from the perspective of global representations, remote sensing images have overall differences due to the disparity in time (spring, summer, autumn, and winter), weather, sensors, etc.; from the perspective of local features, pixel-level semantic segmentation requires more local information. In view of this, we propose the Global style and Local matching Contrastive learning Network (GLCNet) framework, in which the global style contrastive learning module focuses on the global representation, and the local matching contrastive learning module is used to learn pixel (local) level features.

The main contributions of this paper are summarized as follows:

1) As far as we know, we are the first to apply self-supervised contrastive learning to remote sensing semantic segmentation tasks, and verify it on multiple datasets that it can directly learn features from unlabeled images to guide the downstream semantic segmentation tasks with limited annotations;

2) We propose a new self-supervised contrastive learning framework - Global style and Local matching Contrastive learning Network (GLCNet) - which focuses on the problem of the balance between global and local feature learning in remote sensing semantic segmentation tasks;

3) We evaluate our proposed method on two public datasets and two realistic datasets. Experiment results show that with only 1% of the original annotated data, our method improves Kappa by 6% on the Potsdam Dataset and 3% on Deep Globe Land Cover Dataset compared to the existing benchmark. It also outperforms supervised learning in the situation where the upstream dataset and the downstream dataset are not highly similar.

The remainder of this paper is organized as follows. In Section II, we introduce the semantic segmentation approach based on the self-supervised contrastive learning and our further improved self-supervised approach for the semantic segmentation task – GLCNet. The experiments and results are provided in Section III. Section IV states further discussion, and the conclusion is provided in Section V.

II. Method

A. Overview

The success of supervised deep learning relies on a large number of labeled samples, which is difficult to meet in remote sensing semantic segmentation. As shown in fig. 1, self-supervised learning provides a new paradigm for learning potential useful knowledge directly from a large amount of readily available unlabeled data, and then transfer it to downstream tasks to achieve better performance especially with limited labeled samples. In our work, the downstream task is the semantic segmentation of remote sensing images, so we concentrate on designing a self-supervised model for the semantic segmentation task of remote sensing. In this paper, we introduce contrastive learning to learn the general invariant representation, and at the same time, we propose the GLCNet self-supervised method taking into account the characteristics of the semantic segmentation task, as shown in fig. 2, which mainly contains two modules:

![Fig. 1. The self-supervised learning paradigm](image-url)
1) The global style contrastive learning module mainly considers that the global average pooling features used in the existing contrastive learning to measure sample features are not a good substitute for the overall features of an image, so the style features that are more representative of the overall features of an image are introduced to help the model learn global representations better;

2) The local matching contrastive learning module is mainly considered in view of the richness of feature classes on a single image of the semantic segmentation dataset, which may lose a lot of detailed information by extracting only global features, while the image-level representation may be sub-optimal for semantic segmentation tasks which require pixel-level discrimination.

B. Contrastive learning

Contrastive learning is to learn by forcing positive sample pairs to be similar and negative sample pairs to be dissimilar [34], [36]. The key to contrastive learning methods is to construct positive and negative samples. The most recent breakthrough methods [33], [34] classify instances as their own labels, which means that different enhanced versions of a sample are treated as positive samples, and other samples are treated as negative samples. Contrastive learning can encourage the model to learn the invariance of the transformation and the ability of distinguishing different samples. In this work, we use contrastive learning to learn general spatiotemporal invariance features for remote sensing. Specifically, we perform random rotation, cropping, scaling and other operations on samples in order to make the model learn spatial invariance such as rotation invariance and scale invariance. In addition, the temporal difference of remote sensing images mainly lies in the overall texture and color differences caused by seasonal factors and imaging conditions. Due to the lack of multi-temporal image data, we simulate the time transformation by applying random color distortion, random noise, etc. on the samples in order to make the model learn temporal-invariant features.

Inspired by SimCLR [34], we apply contrastive learning to train the encoder part of the semantic segmentation network, as shown in fig. 3, which consists of the following four main components:

1) Data augmentation: In order to encourage the model to learn general spatiotemporal invariance features, as shown in fig. 3, we perform spatial transformations such as random cropping, resizing, flipping and rotation for the learning of spatial invariance features, and simulate temporal transformations with color distortion, Gaussian blur, random noise, etc. for the learning of temporal invariance features. Specifically,
two augmented views $\hat{x}$ and $\tilde{x}$ are generated from a given sample $x$ by data augmentation $t_1$ and $t_2$, i.e., $\hat{x} = t_1(x)$, $\tilde{x} = t_2(x)$. In this work, $t_1$ means random cropping followed by resize to a fixed resolution (e.g., $224 \times 224$), $t_2$ means sequentially applying several augmentations: random cropping followed by resizing to a fixed resolution, random flipping, random rotating, random color distortion, random Gaussian blur, random noise and random gray.

2) Feature extraction: The global feature is extracted from the augmented sample instances using encoder network $e(\cdot)$:

$$\tilde{f}_i = \mu(e(\hat{x}_i)), \hat{f}_i = \mu(e(\tilde{x}_i))$$

Where $\mu$ represents the calculation of the mean value of each channel in the feature map, i.e. the global average pooling. And in this work, the $e(\cdot)$ is the encoder part of the semantic segmentation network DeepLabV3+ [18].

3) Projection head: As shown in equation 2, the projection head $g(\cdot)$ is a MLP with one hidden layer (ReLU). The presence of $g(\cdot)$ in SimCLR [34] has been proved to be very beneficial, may due to it can allow $e(\cdot)$ form and retain more potential useful information for downstream tasks.

$$z_i = g(\tilde{f}_i) = W^{(2)}\sigma(W^{(1)}\tilde{f}_i), \hat{z}_i = g(\hat{f}_i)$$

Where $\sigma$ is a ReLU non-linearity.

4) Contrastive loss: Contrastive loss expects positive sample pairs to be similar and negative sample pairs dissimilar. Specifically, $N$ samples from a mini-batch are augmented to be $2N$ samples. A pair of samples augmented from the same sample form a positive pair, other $2(N-1)$ samples are negative samples, so the contrastive loss $L_C$ is defined as

$$L_C = \frac{1}{2N} \sum_{k=1}^{N} (\ell(\hat{x}_i, \hat{x}_i) + \ell(\hat{x}_i, \tilde{x}_i))$$

with:

$$\ell(\hat{x}_i, \hat{x}_i) = -\log \frac{\exp(sim(z_i, \hat{z}_i)/\tau)}{\sum_{s \in \Lambda^{-}} \exp(sim(z_i, g(\hat{f}(x))/\tau)}$$

Where $sim$ denotes 2 (N-1) negative samples in addition to the positive sample pair, $\tau$ denotes a temperature parameter.

Although powerful image-level representations can be learned by existing contrastive learning paradigm, there are still some problems: First, the existing contrastive learning uses global average pooling features for feature extraction of a sample, which may not be a good representation of the overall characteristics of a sample; second, more critically, the image-level representation learned by the instance-wise contrastive learning may be suboptimal for semantic segmentation tasks that require pixel-level discrimination. Therefore, we proposed the GLCNet.

C. Global style and Local matching Contrastive learning Network (GLCNet)

Our proposed GLCNet method is shown in fig. 2 which mainly contains two modules: the global style contrastive learning module is mainly focusing on the problem that the global average pooling features used in existing contrastive learning are not a good substitute for the overall features of a complex remote sensing image; the local matching contrastive learning module is mainly considered that most existing contrastive learning methods are designed for image classification tasks to obtain image-level features, which may be suboptimal for semantic segmentation requiring pixel-level discrimination. The details are as follows.

1) Global style contrastive learning module: The global style contrastive learning learns by forcing different augmented views of a sample to be similar and dissimilar with other samples, analogous to the existing instance-wise contrastive learning. The difference is that we use style features instead of the simple average pooling features used in instance-wise contrastive learning, as we believe it is more representative of overall features of an image. Huang and Belongie [37] indicate that channel-wise mean and variance of the feature map extracted by CNN can represent the style features of the image, so we calculate the channel-wise mean and variance of the features extracted by the encoder $e(\cdot)$ to extract the global style feature vector, which is defined as

$$f_s(x_i) = \text{concat}(\mu(e(x_i)), \sigma(e(x_i)))$$

Where $\mu$ denotes the channel-wise mean of the feature map, $\sigma$ denotes the channel-wise variance.

Therefore, for $N$ samples from a mini-batch, similar to equation 3, the global style contrastive learning loss is defined as follows

$$L_G = \frac{1}{2N} \sum_{k=1}^{N} (\ell_g(\hat{x}_i, \hat{x}_i) + \ell_g(\hat{x}_i, \tilde{x}_i))$$

with:

$$\ell_g(\hat{x}_i, \hat{x}_i) = -\log \frac{\exp(sim(\hat{z}_i^s, \hat{z}_i^s)/\tau)}{\sum_{s \in \Lambda^{-}} \exp(sim(\hat{z}_i^s, g(\hat{f}(x))/\tau)}$$

Where $\hat{z}_i^s = g(f_s(\hat{x}_i)), \hat{z}_i^s = g(f_s(\hat{x}_i))$
2) Local matching contrastive learning: The local matching contrastive learning module is proposed mainly because of the following two reasons: firstly, the land cover categories in a single remote sensing semantic segmentation image are extremely rich. Only extracting the global features of the whole image to measure and distinguish will lose a lot of information; secondly, instance-wise contrastive learning methods are used to obtain image-level features which may be suboptimal for semantic segmentation as it requires pixel-level discrimination. Therefore, the local matching contrastive learning module is designed to learn the representation of local regions which is beneficial for pixel-level semantic. It consists of the following main components:

\[ \hat{f}_L^j = f_L(\hat{p}_j) = \mu(\hat{p}_j), \hat{f}_L^j = \mu(p_j) \]  

where \( \mu \) represents the calculation of the mean value of each channel in the feature map.

\[ \ell_L(\hat{p}_j, \hat{p}_j) = -\log \frac{\exp \left( \frac{\text{sim}(\hat{p}_j, \hat{p}_j)}{\tau} \right)}{\sum_{p \in \Lambda_L} \exp \left( \frac{\text{sim}(p, g_L(f_L(p)))}{\tau} \right)} \]  

\[ \mu_j = g_L(\hat{f}_L^j) = g_L(f_L(\hat{p}_j)), \hat{\mu}_j = g_L(\hat{f}_L^j) \]  

where, \( N_L \) denotes the number of all local regions selected from a mini-batch of \( N \) samples, \( n_p \) is the number of matched local regions obtained from a sample, \( \Lambda_L \) is a set of feature maps corresponding to all local regions except the two matched local regions, \( g_L(\cdot) \) is a projection head which is similar to \( g(\cdot) \).

3) Total loss: The local matching contrastive loss updates a complete semantic segmentation encoder-decoder network by forcing the feature representations of matching local regions to be similar and the feature representations of different local regions not to be similar. For \( N \) samples from a mini-batch, the local matching contrastive loss is defined as follows:

\[ L = \frac{1}{2N_L} \sum_{j=1}^{N_L} (\ell_L(\hat{p}_j, \hat{p}_j) + \ell_L(p_j, \hat{p}_j)) \]

where, \( \lambda \) is the constant 0.5 in this work. \( \ell_L \) represents the global style contrastive learning which is only used to update the encoder network. And \( L \) represents the local matching contrastive loss in equation \[7\], which can update both the encoder part and decoder part at the same time.

Furthermore, we provide algorithm \[1\] to describe our proposed GLCNet in detail.

III. Experiments and Results

A. Data Description

We evaluate the proposed GLCNet and other supervised methods on four datasets for remote sensing semantic segmentation. The ISPRS Potsdam Dataset and the Deep Globe Land Cover Classification Dataset are publicly available datasets, and the other Hubei and Xiangtan datasets come from real world which have the same spatial resolution and similar classification system that is convenient for studying the impact of domain differences. The details of the four datasets are explained below.
We select 730 images for training and 2448 DeepGlobe Land Cover Classification dataset [38] provides 14 patches that are cropped into 256 test set contains 1500 samples randomly selected from the remaining 14 patches which are cropped into 13824 images with a size of 6000 pixels in Potsdam dataset. 24 patches which are cropped into 13824 images with a size of 256×256 pixels are used for self-supervised training. To evaluate the performance of self-supervised learning, 1% of the self-supervised training set is selected by default as the training set for the downstream semantic segmentation task, and the test set contains 1500 samples randomly selected from the remaining 14 patches that are cropped into 256×256 pixels.

2) Deep Globe Land Cover Classification Dataset (DGLC): DeepGlobe Land Cover Classification dataset [38] provides high-resolution sub-meter satellite images with a size of 2448×2448 pixels. The labels are far from perfect covering seven classes: urban, agriculture, rangeland, forest, water, barren, and unknown. We select 730 images for training and 73 images for downstream testing. Moreover, each image is cropped to a size of 512×512 pixels, and the final number of samples used for each stage is shown in table I.

3) Hubei Dataset: The images of the Hubei dataset are acquired from the Gaofen-2 satellite, covering the Hubei Province of China. The images have a spatial resolution of 2 m with three spectral bands(RGB). The labels are with poor quality, covering 10 classes: background, farmland, urban, rural areas, water, woodland, grassland, other artificial facilities, road and other. We first divided the entire Hubei province into several patches with a size of 13889×9259 pixels. We randomly select 34 of them for training and 5 for testing. Moreover, each image is cropped to a size of 256×256 pixels, and the final number of samples used for each stage is shown in table I.

4) Xiangtan Dataset: The images of the Xiangtan dataset also come from the Gaofen-2 satellite, covering the Xiangtan city of China. The labels are with higher quality than the Hubei dataset, covering 9 classes: background, farmland, urban, rural areas, water, woodland, grassland, road and other. The entire Xiangtan City is divided into patches with a size of 4096×4096 pixels, and we randomly select 85 of them for training and 21 for testing. Moreover, each image is cropped to a size of 256×256 pixels, and the final number of samples used for each stage is shown in table I.

B. Experiment Setup

1) Baseline: In order to evaluate the performance of self-supervised learning, we randomly initialize the network on specific downstream semantic segmentation tasks as the basic baseline. In addition, the common ImageNet pre-training strategy is also used as the baseline. Moreover, we compare 3 typical self-supervised tasks (predicting context [39], [40], image inpainting [41], and instance-wise contrastive learning [34], [36], [42]), the specific methods we used for comparison are summarized as follows:

1. Random baseline: Train from scratch on a specific semantic segmentation task without pre-training;
2. ImageNet Pre-training: Use the supervised training model on ImageNet to initialize the backbone of the semantic segmentation model;
3. Jigsaw [40]: This self-supervised method constructs self-supervised tasks by solving puzzles. Specifically, as shown in fig. 6 1), the given image is divided into multiple patches, and then the patch is shuffled before being sent to the CNN network. It is expected that the

| Datasets         | Potsdam | DGLC | Hubei | Xiangtan |
|------------------|---------|------|-------|----------|
| Ground resolution| 0.05m   | 0.5m | 2m    | 2m       |
| Spectral bands   | NIR,RGB| RGB  | RGB   | RGB      |
| Crop size        | 256 × 256 | 512 × 512 | 256 × 256 | 256 × 256 |
| Amount of self-supervised data | 13824 | 18248 | 66471 | 16051 |
| Default amount of training data | 138 | 182 | 664 | 160 |
| Amount of testing data | 1500 | 1825 | 9211 | 3815 |

Algorithm 1 Algorithm of GLCNet Pre-training

Require: A set of images X; structure of e(·), d(·), g(·), gL(·); augmentation t1,t2; batch size N; parameters τ, λ, s_p, n_p

1. for sampled mini-batch x = {x_k}^{N}_{k=1} from X do
2. Build index label y = {y_k}^{N}_{k=1} using the size of x
3. for all do
4. Draw augmentations: \( ẋ_k = t_1(x_k), ẏ_k = t_1(y_k), \)
5. end for
6. Get globe style features:
   \( f_s(\hat{x}) = \text{concat}(\mu(e(\hat{x})), \sigma(e(\hat{x}))) \)
   \( f_s(x) = \text{concat}(\mu(e(x)), \sigma(e(x))) \)
7. Get local matching feature maps p̃ and ̃p from d(e(\hat{x})) and d(e(\hat{x})) according to s_p, n_p, ẏ and ẏ
8. Compute globe style contrastive loss \( \mathcal{L}_G \) using eq. 5
9. Compute local matching contrastive loss \( \mathcal{L}_L \) using eq. 7
10. Compute total loss: \( \mathcal{L} = \lambda \cdot \mathcal{L}_G + (1 - \lambda)\mathcal{L}_L \)
11. Update network e(·), d(·), g(·) and gL(·) to minimize \( \mathcal{L} \)
12. end for
13. return encoder e(·) and decoder d(·)
network will learn contextual relationships between the shuffled patches;

4. **Inpainting** [41]: A typical method of designing self-supervised signals with the idea of image restoration. To be specific, as shown in fig. 6 2), a random region of the image is first discarded, and then the CNN model is trained to predict the original image from the corrupted image, thus enabling the CNN model to learn contextual information;

5. **SimCLR** [34]: The SimCLR method is based on the idea of instance-wise contrastive learning, which is to learn by forcing positive samples enhanced from a sample to be similar and negative samples enhanced from different samples in a mini-batch to be dissimilar;

6. **MoCo v2** [42]: The MoCo v2 is also based on the idea of instance-wise contrastive learning, which focuses on obtaining negative samples far beyond the batch size. Therefore, a dynamic queue is proposed to save the features of negative samples, and a momentum update encoder is proposed to avoid the consistency problem of the representations of negative samples from the rapid change of the encoder.

2) **Evaluation Metrics**: The performance of self-supervised methods needs to be evaluated on specific downstream semantic segmentation tasks. Therefore, we use OA and Kappa to measure the overall accuracy on the test set of downstream tasks, which are defined as follows

\[
OA = \frac{TP}{N} \quad (9)
\]

\[
\text{Kappa} = \frac{OA - p_e}{1 - p_e} \quad (10)
\]

Where \(TP\) denotes the total number of pixels that are correctly predicted, i.e. true positive. \(N\) represents the total number of pixels, \(p_e = \frac{\sum_{c} a_c \times b_c}{N \times N}\), \(a_c\) denotes the actual number of pixels of class \(c\), \(b_c\) denotes the number of predicted pixels of class \(c\).

In addition, we use the F1-score to measure the accuracy of a single category, which is defined as follows

\[
F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (11)
\]

Where, precision = \(\frac{TP}{TP+FP}\), recall = \(\frac{TP}{TP+FN}\). \(TP\) is true positive, \(FP\) is false positive, \(FN\) is false negative.

3) **Implementation Details**: In the self-supervised pre-training phase, Jigsaw, SimCLR and MoCo v2 are only designed to train the encoder part of DeepLabV3+ with ResNet50 backbone, while inpainting and the proposed GLCNet can train the complete encoder-decoder part of DeepLabV3+. We use Adam optimizer for 400 epochs, with a batch size of 64. The initial learning rate is set as 0.01 with the cosine decay schedule. Moreover, for the proposed GLCNet method, we choose 4 local regions with a size of 48×48 from a sample, i.e. \(s_p = 48, n_p = 4\). The model with the lowest loss in the self-supervised pre-training process is saved for the downstream task.

Although inpainting and GLCNet method can train both encoder part and decoder part of the network during the self-supervised training, since methods such as SimCLR used for comparison are designed to train only the encoder part, we only load the encoder part from self-supervised pre-training stage by default at the fine-tune stage. In the fine-tuning phase, we use only a limited amount of annotated data for semantic segmentation training, such as 1% of the amount of self-supervised data. We use Adam optimizer for 150 epochs, with a batch size of 16. The initial learning rate is set as 0.001, and decreases to 0.98 every epoch.

C. Experimental Results

In this section, we first compare the proposed GLCNet with other self-supervised methods and ImageNet pre-training method on several RSIs semantic segmentation datasets. And then we explore two factors that may affect the self-supervised pre-training performance on the target RSIs semantic segmentation task, which are the amount of self-supervised pre-training data and the domain difference between pre-training dataset and fine-tuning dataset.

1) **Comparison with other methods**: In this section, we evaluate the performance of the proposed GLCNet on the remote sensing semantic segmentation task with limited annotations on multiple datasets, comparing with other self-supervised learning method, ImageNet pre-training and random initialization method. The amount of data used for self-supervised pre-training on each dataset is shown in table 1 and 1% of the amount of self-supervised data is used for
TABLE II
COMPARISON WITH OTHER METHODS USING LIMITED LABELED DATA ON FOUR RSIS SEMANTIC SEGMENTATION TASKS

| Pretext task       | Potsdam | DGLC   | Hubei  | Xiangtan |
|--------------------|---------|--------|--------|----------|
|                    | Kappa   | OA     | Kappa  | OA       |
| Random Baseline    | 58.27   | 67.39  | 51.47  | 71.70    |
| ImageNet Baseline  | 67.65   | 74.83  | 65.64  | 79.17    |
| Jigsaw             | 60.99   | 69.68  | 44.05  | 68.88    |
| Inpainting         | 63.62   | 71.70  | 43.82  | 67.78    |
| Moco v2            | 59.81   | 68.73  | 54.02  | 72.93    |
| SimCLR             | 65.62   | 73.21  | 64.20  | 78.79    |
| Ours(GLCNet)       | 71.80   | 78.05  | 67.19  | 79.40    |

Fig. 7. F1-score for each class on four RSIs semantic segmentation tasks

fine-tuning. From the results in table II, we can find that our proposed GLCNet method brings great improvements on all datasets compared to not using any pre-training strategy. At the same time, we also find that different self-supervised methods have a significant impact on the results, and inappropriate self-supervised methods even having a negative impact, while our method achieves state-of-the-art results. As illustrated in fig. 8, we also show some visualization results, where our method is relatively better overall. Meanwhile, in order to measure whether our method has advantages in each category, we calculate the single-class accuracy, and the results are shown in fig. 7, where it can be found that our method achieves superior results in most classes compared to other self-supervised methods on multiple datasets.

In addition, we are surprised to find that our method outperformed the ImageNet pre-training method on most datasets, where the ImageNet pre-training method is obtained by supervised training on ImageNet with millions of data, which is much more than the amount of self-supervised data used in our experiments. This shows that although the ImageNet pre-training method can provide a significant improvement, it is not the optimal approach due to the large differences between natural images and remote sensing images. For example, as shown in table II, our method on the Potsdam dataset has the most obvious improvement compared to ImageNet pre-training, may due to the fact that the dataset has four
bands, which is the most different from RGB natural images. Therefore, it is more reasonable if we can train a general model directly from unlabeled remote sensing images. Furthermore, it is worth noting that in our experiments, the images used for the self-supervised pre-training are similar to those for the downstream task, both originating from the same dataset. Such a situation is available, as we can easily obtain a large number of images from the same source through satellite technology.

2) Effect of the amount of self-supervised data: Since the self-supervised pre-training does not need annotated data, and a large amount of image data is easy to obtain, this section mainly explores whether more self-supervised pre-training data can improve performance. We conduct experiments on Potsdam and Xiangtan datasets, where 20%, 50% to 100% of self-supervised data is randomly selected. The results are shown in fig. 9, where None means that no self-supervised training is performed. From the results, we can find that in both datasets, there is an overall increasing trend as the amount of self-supervised data increases, and the improvement of our method is relatively more obvious compared with the SimCLR method. Therefore, it is foreseeable that it may be more beneficial when using larger datasets for self-supervised training.
Fig. 9. Classification accuracies with different amounts of self-supervised data. 1) results of Potsdam dataset, 2) results of Xiangtan dataset

### TABLE III
RESULTS ON THE DOMAIN DIFFERENCE BETWEEN SELF-SUPERVISED PRE-TRAINING DATASET AND FINE-TUNING DATASET

| Pre-training dataset | Fine-tuning dataset          | Domain differences |
|----------------------|------------------------------|--------------------|
|                      | Xiangtan Kappa | OA | Hubei Kappa | OA |                |
|                      | Kappa | OA | Kappa | OA |                |
| /                    | 64.85 | 78.17 | 46.69 | 58.02 | /                |
| Postdam              | Supervised Baseline | | SimCLR | 66.80 | 79.24 | 47.69 | 58.91 | Very huge |
|                      | GLCNet | | 68.37 | 80.18 | 48.94 | 59.95 |                |
|                      | DGLC |                | 69.35 | 80.72 | 50.81 | 60.93 |                |
|                      | GLCNet |                | 69.84 | 80.88 | 50.09 | 60.81 | huge |
|                      | Supervised Baseline | | / | / | 70.62 | 81.58 | 51.01 | 61.43 |                |
| Xiangtan             | SimCLR | | 70.76 | 81.55 | 49.20 | 59.99 | small |
|                      | GLCNet | | 72.16 | 82.57 | 51.16 | 61.94 |                |
|                      | Supervised Baseline | | 71.16 | 81.89 | / | / |                |
|                      | GLCNet | | 70.67 | 81.53 | 51.85 | 62.12 | small |
| Hubei                | SimCLR | | 71.55 | 82.13 | 53.80 | 63.62 |                |
|                      | GLCNet | | 75.92 | 81.10 | 50.44 | 60.77 | Mixed |
| Xiangtan+Hubei       | SimCLR | | 70.81 | 81.81 | 53.11 | 63.00 | domains |
|                      | GLCNet | | 70.17 | 81.23 | 49.32 | 59.86 | Mixed |
|                      | +DGLC | | 71.19 | 81.76 | 51.62 | 62.05 | domains |

3) Effect of the domain difference: In this section we evaluate the impact of domain differences on the performance of the self-supervised pre-training model, and the results are shown in table III. We find that training with the self-supervised dataset which is more similar to the downstream task dataset led to better model performance on the downstream task. In addition, our method mostly outperforms supervised learning, except in cases where the domain differences are extremely small (e.g., Hubei→Xiangtan, Xiangtan→Hubei), mainly because the two domains not only have the same image resolution and are close in physical location, but also, crucially, the classification system is almost completely consistent, so it is difficult to exceed the accuracy of supervised learning. Although we find in section III-C2 that model performance further improves as the amount of self-supervised training data increases, in this experiment we find that the model performance may not improve or may even be damaged if the self-supervised pre-training dataset is mixed with a large number of images that are not very similar to the downstream task dataset. Fortunately, since the self-supervised pre-training does not require labels, it is feasible to obtain a large amount of image data that is similar to the target dataset.

### D. Ablation Study
In this section, we perform ablation experiments to investigate the effectiveness of the modules of our proposed GLCNet method and the effectiveness of the decoder parameters of the models trained with our method.

1) The effectiveness of the modules of the proposed GLCNet: In this section, we explore the effectiveness of each module in our proposed method on the Potsdam dataset. The experimental results are shown in fig. [10] where:
- **ours_nostyle** indicates the global part without using style features, i.e. directly using the traditional global average pooling features;
- **ours noglobe** indicates that the global style contrastive learning module is completely removed;
- **ours_nolocal** indicates that the local matching contrastive module is completely removed;
- **ours_nostyle_and_nolocal** indicates that the local matching contrastive module is removed and the global part does not use style features.
insufficient annotated samples. And we have some further discussions about our experimental results in this section.

We find that the design of the self-supervised task has a great impact on the final performance, and our proposed method achieves optimal results, even outperforming the ImageNet pre-training method on several datasets. Furthermore, through the experiments in Section III-C2, it is found that when the amount of self-supervised data increases, the accuracy of fine-tuning is further improved. Therefore, it can be foreseen that the model performance is expected to be further improved when more images are used for self-supervision, and a large number of remote sensing images are extremely easy to obtain, so this will have great practical application value.

The self-supervised trained model show potential capability for remote sensing image understand since it only depends on intrinsic supervisor signal instead of task depended labels. As the experimental results shown in section III-C3, our proposed self-supervised approach outperforms supervised learning when there is some difference between the self-supervised and fine-tuned datasets, so it illustrates that the model trained by self-supervised learning is more robust. In practice, we will face many situations where labels are lacking in some local areas. It will be extremely meaningful to learn a general model from images of the global area through self-supervised learning and then migrate to the local area. However, from the experimental results, we found that if multiple datasets with large differences are mixed for self-supervised training, the performance will be impaired when migrating to a local dataset. This reason might be that it is more difficult to perform self-supervised training when mixing multiple domains, as it tends to distinguish between images in different domains with large differences first, while the ability to distinguish between images in a local domain may be reduced. It would be of great practical value if we could subsequently find a way to mix images from multiple domains for self-supervised training without degrading its effectiveness in a single domain, so that we could perform self-supervised training to obtain a more general model by constructing a large image dataset mixing of different resolutions, different regions, different times, etc.

We find that each of the modules we designed have some benefit, while the local matching contrastive learning module brought more benefit on the whole, which illustrates the significance for local scale discrimination on the semantic segmentation dataset. However, the $s_p$ and $n_p$ are set the same for all datasets without much exploration, which may not be the optimal setting. Furthermore, as the distribution of surface features in the actual images can be extremely heterogeneous, the random selection of local areas will be biased towards the more dominant feature classes, and further improvement might be achieved if the selection of local areas could be made more homogeneous.

IV. DISCUSSIONS

In this work, we apply self-supervised mechanisms to remote sensing semantic segmentation datasets, bringing significant improvements to semantic segmentation tasks with insufficient annotated samples. And we have some further discussions about our experimental results in this section.

2) The effectiveness of the decoder part trained with GLCNet: Our method is originally designed to train the full semantic segmentation network, but since most of the methods used for comparison are designed to train only the encoder, we only loaded the self-supervised pre-trained encoder part in the previous experiments during the fine-tuning phase for fairness of comparison. In this section, to investigate whether the decoder trained by our method is valid, we conduct experiments on the Potsdam and Xiangtan datasets. The experimental results are shown in table IV, where $d(1,2)$ denotes loading the decoder parameters of the first two layers and $d(1,2,3)$ denotes loading the complete decoder parameters except for the final classification layer. From the results, it can be found that the decoder parameters trained with our method do not bring out significant improvement, which may be due to the fact that the encoder part of the semantic segmentation network is mainly used for detail recovery, while our current local matching contrastive learning module still performs an average pooling operation on local regions to extract features, losing detail information such as edge localization.

| Load pre-training parameters | Potsdam Kappa | Xiangtan Kappa | Potsdam OA | Xiangtan OA |
|-----------------------------|---------------|---------------|------------|------------|
| encoder                     | 71.80         | 78.05         | 72.16      | 82.57      |
| encoder+d(1,2)              | 72.02         | 78.13         | 72.25      | 82.52      |
| encoder+d(1,2,3)            | **72.17**     | **78.29**     | 71.54      | 82.19      |

V. CONCLUSION

In this work, we introduce self-supervised contrastive learning to remote sensing semantic segmentation tasks for learning general spatio-temporal invariant features from a large number
of unlabeled images in order to reduce the dependence on labelled samples. Furthermore, considering that the existing contrastive learning methods are mainly designed for image classification tasks to obtain image-level representations, which may not be optimal for semantic segmentation tasks that require pixel-level discrimination, so we propose the GLCNet. Experiments portray that our method mostly outperforms traditional ImageNet pre-training method and other self-supervised methods in semantic segmentation tasks with limited labeled data. We also find that more self-supervised pre-training samples can bring performance improvements, and in actual situations we can easily obtain a large amount of remote sensing data, so our method may have great practical application significance.

There are still some shortcomings in our method, for example, we would like to use the GLCNet to better learn general temporal invariance features. However, at present we only simulate temporal transformations by randomly enhancing the images in terms of colour and texture due to the lack of multi-temporal image data. This cannot really imitate the complex transformations caused by seasons, imaging conditions, etc. So, the true temporal features might not be learnt sufficiently, which can subsequently be complemented by using real multi-temporal images. In the future, the method of this paper will be further improved and then applied to large-scale image data to alleviate the critical lack of labeling in tasks such as global land cover.

REFERENCES

[1] G. J. Schumann, G. R. Brakenridge, A. J. Kettner, R. Kashif, and E. Niebuhr, “Assisting flood disaster response with earth observation data and products: A critical assessment,” Remote Sensing, vol. 10, no. 8, p. 1230, 2018.
[2] M. Weiss, F. Jacob, and G. Duveiller, “Remote sensing for agricultural applications: A meta-review,” Remote Sensing of Environment, vol. 236, p. 111402, 2020.
[3] S. Shi, Y. Zhong, J. Zhao, P. Lv, Y. Liu, and L. Zhang, “Land-use/land-cover change detection based on class-prior object-oriented conditional random field framework for high spatial resolution remote sensing imagery,” IEEE Transactions on Geoscience and Remote Sensing, 2020.
[4] J. Xing, R. Sieber, and M. Kalacska, “The challenges of image segmentation in big remotely sensed imagery data,” Annals of GIS, vol. 20, no. 4, pp. 233–244, 2014.
[5] C. Cortes and V. Vapnik, “Support-vector networks,” Machine learning, vol. 20, no. 3, pp. 273–297, 1995.
[6] M. Pal and P. M. Mather, “Support vector machines for classification in remote sensing,” International journal of remote sensing, vol. 26, no. 5, p. 1007–1011, 2005.
[7] B. Breiman, “Random forests,” Machine learning, vol. 45, no. 1, pp. 5–32, 2001.
[8] R. M. Basse, H. Omrani, O. Charif, P. Gerber, and K. Bódis, “Land use changes modelling using advanced methods: Cellular automata and artificial neural networks. the spatial and explicit representation of land cover dynamics at the cross-border region scale,” Applied Geography, vol. 53, pp. 160–171, 2014.
[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in neural information processing systems, vol. 25, pp. 1097–1105, 2012.
[10] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.
[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE conference on computer vision and pattern recognition, ser. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.
[12] 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, ser. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[13] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proceedings of the IEEE conference on computer vision and pattern recognition, ser. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
[14] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, ser. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
[15] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention, ser. International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 4, pp. 834–848, 2017.
[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” arXiv preprint arXiv:1706.05587, 2017.
[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in Proceedings of the European conference on computer vision (ECCV), ser. Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.
[19] F. Waldner and F. I. Diakogiannis, “Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network,” Remote Sensing of Environment, vol. 245, 2020.
[20] R. Liu, L. Mi, and Z. Chen, “Attnet: Adaptive fusion network for remote sensing image semantic segmentation,” IEEE Transactions on Geoscience and Remote Sensing, 2020.
[21] F. Mohammadi-maneshe, B. Salehi, M. Mahdianpari, E. Gill, and M. Molinier, “A new fully convolutional neural network for semantic segmentation of polarimetric sar imagery in complex land cover ecosystem,” ISPRS journal of photogrammetry and remote sensing, vol. 151, pp. 223–236, 2019.
[22] L. Ding, J. Zhang, and L. Bruzzone, “Semantic segmentation of large-size vhr remote sensing images using a two-stage multiscale training architecture,” IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 8, pp. 5367–5376, 2020.
[23] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, and J. Wang, “Deep learning in environmental remote sensing: Achievements and challenges,” Remote Sensing of Environment, vol. 241, p. 111716, 2020.
[24] A. Vali, S. Comai, and M. Matteucci, “Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review,” Remote Sensing, vol. 12, no. 15, p. 2495, 2020.
[25] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” Journal of Big Data, vol. 6, no. 1, pp. 1–48, 2019.
[26] C. Bowles, L. Chen, R. Guerero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, “Gan augmentation: Augmenting training data using generative adversarial networks,” arXiv preprint arXiv:1810.10862, 2018.
[27] M. Frid-Adar, A. Ben-Cohen, R. Amer, and H. Greenspan, “Improving the segmentation of anatomical structures in chest radiographs using u-net with an imagenet pre-trained encoder,” in Image Analysis for Moving Organ, Breast, and Thoracic Images. Springer, 2018, pp. 159–168.
[28] B. C. X. Chen, and Y. Lu, “Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection,” IEEE Access, vol. 8, pp. 116744–116755, 2020.
[29] J. Geng, X. Deng, X. Ma, and W. Jiang, “Transfer learning for sar image classification via deep joint distribution adaptation networks,” IEEE Transactions on Geoscience and Remote Sensing, vol. 58, no. 8, pp. 5377–5392, 2020.
[30] H. Li, Z. Cui, Z. Zhu, L. Chen, J. Zhu, H. Huang, and C. Tao, “R-Snet: Deep meta metric learning for few-shot remote sensing scene classification,” arXiv preprint arXiv:2009.13364, 2020.
[31] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, “Dada: Depth-aware domain adaptation in semantic segmentation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, ser. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7364–7373.

[32] L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[33] 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, ser. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[34] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in International conference on machine learning, ser. International conference on machine learning. PMLR, 2020, pp. 1597–1607.

[35] J. Li, P. Zhou, C. Xiong, R. Socher, and S. C. Hoi, “Prototypical contrastive learning of unsupervised representations,” arXiv preprint arXiv:2005.04966, 2020.

[36] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, ser. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742.

[37] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in Proceedings of the IEEE International Conference on Computer Vision, ser. Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.

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

[39] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in Proceedings of the IEEE international conference on computer vision, ser. Proceedings of the IEEE international conference on computer vision, 2015, pp. 1422–1430.

[40] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in European conference on computer vision, ser. European conference on computer vision. Springer, 2016, pp. 69–84.

[41] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in Proceedings of the IEEE conference on computer vision and pattern recognition, ser. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544.

[42] X. Chen, H. Fan, R. Girshick, and K. He, “Improved baselines with momentum contrastive learning,” arXiv preprint arXiv:2003.04297, 2020.