



Article

RiSSNet: Contrastive Learning Network with a Relaxed Identity Sampling Strategy for Remote Sensing Image Semantic Segmentation

Haifeng Li ^{1,2} , Wenxuan Jing ¹, Guo Wei ³, Kai Wu ³, Mingming Su ³, Lu Liu ³, Hao Wu ³, Penglong Li ⁴ and Ji Qi ^{1,*}

¹ School of Geosciences and Info-Physics, Central South University, Changsha 410083, China; lihaifeng@csu.edu.cn (H.L.); 215012162@csu.edu.cn (W.J.)

² Xiangjiang Laboratory, Changsha 410205, China

³ Inner Mongolia Civil-Military Integration Development Research Center, Hohhot 010070, China; guow0505@gmail.com (G.W.); kaiwu5353@gmail.com (K.W.); mingmsu05@gmail.com (M.S.); lovinlechengtest@gmail.com (L.L.); wuhao960201@gmail.com (H.W.)

⁴ Chongqing Geomatic and Remote Sensing Center, Chongqing 401147, China; penglongli@whu.edu.cn

* Correspondence: erenturing@csu.edu.cn

Abstract: Contrastive learning techniques make it possible to pretrain a general model in a self-supervised paradigm using a large number of unlabeled remote sensing images. The core idea is to pull positive samples defined by data augmentation techniques closer together while pushing apart randomly sampled negative samples to serve as supervised learning signals. This strategy is based on the strict identity hypothesis, i.e., positive samples are strictly defined by each (anchor) sample's own augmentation transformation. However, this leads to the over-instanting of the features learned by the model and the loss of the ability to fully identify ground objects. Therefore, we proposed a relaxed identity hypothesis governing the feature distribution of different instances within the same class of features. The implementation of the relaxed identity hypothesis requires the sampling and discrimination of the relaxed identical samples. In this study, to realize the sampling of relaxed identical samples under the unsupervised learning paradigm, the remote sensing image was used to show that nearby objects often present a large correlation; neighborhood sampling was carried out around the anchor sample; and the similarity between the sampled samples and the anchor samples was defined as the semantic similarity. To achieve sample discrimination under the relaxed identity hypothesis, the feature loss was calculated and reordered for the samples in the relaxed identical sample queue and the anchor samples, and the feature loss between the anchor samples and the sample queue was defined as the feature similarity. Through the sampling and discrimination of the relaxed identical samples, the leap from instance-level features to class-level features was achieved to a certain extent while enhancing the network's invariant learning of features. We validated the effectiveness of the proposed method on three datasets, and our method achieved the best experimental results on all three datasets compared to six self-supervised methods.

Keywords: semantic segmentation; remote sensing (RS); self-supervised learning; contrastive learning



Citation: Li, H.; Jing, W.; Wei, G.; Wu, K.; Su, M.; Liu, L.; Wu, H.; Li, P.; Qi, J. RiSSNet: Contrastive Learning Network with a Relaxed Identity Sampling Strategy for Remote Sensing Image Semantic Segmentation. *Remote Sens.* **2023**, *15*, 3427. <https://doi.org/10.3390/rs15133427>

Academic Editor: Biswajeet Pradhan

Received: 16 May 2023

Revised: 20 June 2023

Accepted: 3 July 2023

Published: 6 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the maturation of high-definition satellite technology, the acquired RSIs have greatly improved in both quantity and quality, and the extracted feature content and details are becoming increasingly rich [1–4], allowing satellites to become a true “eye on the earth”. However, the massive scale of RSIs presents new difficulties for manual annotation. In the face of a large number of high-quality RSIs, starting from the detailed features of the images themselves becomes the key to quickly decoding the information content of these images. Traditional unsupervised representation learning methods lack an effective feedback mechanism after abandoning manual annotations, resulting in the learned features not being

discriminative and invariant enough for RSI understanding tasks. Self-supervised learning uses pretext tasks [5–7] to generate pseudo-labels for massive unlabeled data, learning good representations that benefit various downstream tasks.

As a representative method of self-supervised learning, contrastive learning is both efficient and effective, and is therefore widely used in the remote sensing community. It treats each sample (called an anchor sample) as a separate class and augments the anchor sample to obtain corresponding positive samples while treating the rest of the samples in the dataset as negative samples. Based on the construction of such similar and different samples, the distances between the similar samples (anchor samples and positive samples) are reduced, and the distances between different samples (anchor samples and negative samples) are increased, thereby constructing comparative supervision signals to enable a model to learn the features of similar and different instances. This supervision signal construction approach based on sample contrast is widely used, and several typical methods such as SimCLR [8,9] and MoCo [10–12] follow the approach of treating each sample as a separate class and generating positive samples through data augmentation. Different data augmentation methods simulate anchor samples with different time and state representations and aid in the learning of the intraclass invariance [13] and interclass differentiability [14] of features by correspondingly reducing and increasing the distances between features. Interclass invariance means that the features extracted from similar samples remain the same or essentially the same regardless of changes in the external conditions. Interclass differentiability means that the features extracted from samples of different classes are differentiable or significantly different, regardless of how similar the features are to each other.

Traditional contrastive learning methods use data augmentation [8,15,16] as a means of learning invariance, and various studies have also shown that data augmentation is an important part of bridging intraclass distances. The hypothesis underlying the method of obtaining positive samples through the data augmentation of anchor samples only is called the strict identity hypothesis; in this method, positive samples are strictly defined as the samples obtained through anchor sample augmentation. This definition allows one to reduce the distances between the anchor sample and the positive samples under the strict identity hypothesis without establishing other connections between samples in the same class while ignoring other samples of the same class as the anchor sample in the dataset. This limited data augmentation approach cannot model the features of different samples within a class, let alone constrain the distribution of different instances of the same class in the feature space. Especially in the field of remote sensing, since the features of samples in the same class often exhibit considerable richness, complexity, and imbalance in space [17,18], the data augmentation of anchor samples alone is insufficient to support the learning of the feature invariance of samples within the same class. Therefore, we proposed a relaxed identity hypothesis under which the definition of positive samples is extended from the augmentation of anchor samples to the augmentation of anchor samples and similar samples. Under this hypothesis, the selection of similar samples is particularly important. Some works have attempted to increase the number of similar samples by prior knowledge [19,20]. However, this type of approach is too certain in regard to the samples obtained from prior knowledge, leading false-positive samples to have a large impact on the network. In addition, some works have used clustering [21,22] to select similar samples, that is, selecting samples close to the anchor sample cluster center as similar samples. In the field of remote sensing, the selection of similar samples relies on hyperspectral RSIs for spectral clustering [23,24]. The results of such work will be affected by the clustering results, and poor clustering results will erroneously pull in the distance of different classes. Even works that attempt to reduce or eliminate the impact of false-positive samples on the network through correction methods still cannot escape from this over-reliance on prior knowledge. In summary, the available methods for contrastive learning from similar positive samples still suffer from the limitations of using only data augmentation methods, employing simple clustering methods, and relying excessively on prior knowledge.

To address the above problem, we proposed a relaxed positive sample selection and discrimination method for RSI contrastive learning. Since self-supervised learning cannot rely on labels to help find similar features and distinguish different features, we defined a double constraint of semantic similarity under the relaxed identity sampling strategy and feature similarity under relaxed sample discrimination to sample and discriminate relaxed positive samples. First, to perform relaxed identity sampling, we sampled the features within the neighborhood of the anchor samples according to the fact that nearby objects in a remote sensing dataset tend to have a strong correlation [25]; the samples obtained in this way were grouped into the relaxed sample queue, and a semantic similarity measure was defined. Second, to discriminate the samples in the relaxed sample queue, we defined the feature similarity, ranked the samples in the relaxed sample queue by means of the self-discriminating ability of the network, and fused the samples at the front end of the ranked relaxed sample queue as the relaxed positive samples; these fused samples were used as positive samples to close the distance with the anchor sample. We used a two-branch network to implement the sampling and discrimination of samples under the above relaxed identity sampling strategy. One of the network branches performed instance-level feature learning to provide the initial network weights for the sampling and discrimination of relaxed samples, and the other performed the discrimination of relaxed samples based on this instance-level learning. Our approach moved away from sole reliance on either data augmentation, clustering, or prior knowledge while elevating feature learning from instance-level to class-level learning. We present experiments conducted on three datasets to demonstrate the effectiveness of our method.

The main contributions of this paper are as follows:

- (1) We proposed a relaxed identity hypothesis as an extension of the strict identity hypothesis currently used to define positive samples for contrastive learning, i.e., to consider similar samples as positive samples instead of considering only augmented views of the same sample. Our proposed relaxed identity hypothesis could fundamentally alleviate the problem of incomplete object recognition due to the over-instantiation of features obtained by contrastive learning.
- (2) Following the relaxed identity hypothesis, we proposed a novel contrastive learning method, RiSSNet, which used spatial proximity sampling and visual similarity discrimination strategies to filter out similar samples for the construction of positive sample pairs. With the dual constraints of spatial proximity and visual similarity used to construct positive samples, RiSSNet achieved a more compact feature space through contrastive learning to obtain category-level features instead of instance-level features.
- (3) We experimentally verified the effectiveness of the proposed RiSSNet on three representative semantic segmentation datasets, which contributed to a deeper understanding of the relaxation hypothesis.

2. Materials and Methods

2.1. Related Works

2.1.1. Semantic Segmentation of Remote Sensing Images

The semantic segmentation task for RSIs is regarded as an important way to obtain information from RSIs, and these segmentation results are widely used in many fields, such as disaster acquisition and land resource exploration. Long et al. [26] proposed a fully convolutional neural network (FCN) semantic segmentation model, which achieved more robust results by combining the results of different layers for image segmentation. Ronneberger et al. [27] designed a U-Net model with a symmetric network structure to perform the stitching operation, using the symmetric structure of the network to fuse shallow and deep image features. Chen et al. [28] proposed DeeplabV3+, which more effectively preserved the shallow information of the network through feature fusion compared to DeepLabV3 [29] in order to obtain stable and highly accurate segmentation results. Because RSIs have the distinctive characteristics of large scale differences, a large number

of spectra, and a high resolution, the semantic segmentation methods based on the above-mentioned models must often be optimized specifically for RSIs. Wu et al. [30] solved the problem of small-feature loss in ASPP by constructing feature pyramids and improved the segmentation results for roads by using dense links. Ding et al. [31] proposed an attention-based mechanism for LAnet, which added contextual information by means of patch-level attention modules to complement the spatial structure in high-level features.

The abovementioned works focused on the importance of dense contextual relationships for the semantic segmentation of RSIs, but most of them simply supplemented semantic segmentation with spatial correlations and failed to establish the semantic correlations between different dense blocks.

2.1.2. Semantic Segmentation for Remote Sensing Images Based on Self-Supervised Contrastive Learning

Since self-supervised contrastive learning methods were proposed, they have seen wide application in the field of semantic segmentation for RSIs. Contrastive learning has been proven to enable models to learn good instance-level features under the strict identity hypothesis and has been widely used in various fields, becoming a mainstream method for unsupervised learning. Wu et al. [32] first proposed the instance discrimination task, which treats an anchor sample and its augmented counterpart as a positive sample pair to be pulled closer together and the rest of the samples as negative samples to be pushed away, as a means to learn the features of different instances. Since the introduction of instance-discriminative contrastive learning, data augmentation has been regarded as an important part of contrastive learning, as data augmentation can simulate the variations in sample instances at different times and locations in order to learn instance features under different conditions. Chen et al. [8] investigated the effect of data augmentation in the self-supervised domain and found that among three augmentation methods (random cropping, color distortion, and Gaussian blur), random cropping and color distortion were more beneficial for the learning of instance sample representations. Subsequently, further studies [33–37] were conducted on data augmentation. Tian et al. [15] proposed an idea for how positive sample pairs should be designed from the perspective of mutual information. Peng et al. [38] started from cropping as a data augmentation tool; cropping the foreground part of an image can allow a network to distinguish foreground from background and thus learn the foreground better. In addition, Li et al. [39] proposed augmentation methods for the spatiotemporal learning of RSIs and supplemented the learned feature details using global and local contrastive learning. Mao et al. [40] developed a multilevel self-supervised contrastive learning framework for constructing local semantic-level positive and negative sample pairs.

Although the above methods optimized the features learned under the strict identity condition from various perspectives, the overly strict identity constraint nevertheless restricts feature learning to the instance level and prevents the construction of feature associations at the class level. Especially for RSIs, the complexity and variability of the features make the gap between the instance level and the class level highly prominent. In [39–41], we attempted to use patch-level feature learning to overcome the above problem but still failed to extend the content considered in patch-level learning beyond instance-level features.

2.1.3. Positive Sampling Strategies

With the continuous improvement of contrastive learning techniques, work related to positive sample sampling is also emerging, and research extends from self-supervised contrastive learning to supervised contrastive learning. The current sampling methods for contrastive learning can be broadly divided into three main categories: work related to label-assisted sampling under supervised learning conditions, work related to clustering sampling under self-supervised learning and other for work related to sampling of dataset characteristics under self-supervised conditions.

Positive sampling under supervised learning was first proposed by Khosla et al. [42], who integrated supervised learning and contrastive learning to sample positive samples with the help of labels and suggested that harder positive samples would bring greater gain to a network. In this way, the ability of a network to learn the invariant features of similar samples can be improved by associating similar samples with the help of labels. The work related to positive sampling in the context of self-supervised learning can be broadly divided into two categories: positive sampling based on clustering and positive sampling based on the characteristics of the dataset itself. Li et al. [43] performed positive sampling via clustering from a self-supervised perspective by searching neighboring samples in the clusters to which the anchor samples belonged and retaining samples close to the center of a cluster in the corresponding anchor sample category as positive samples; however, the positive samples selected by this method were highly similar to the anchor samples, and the differences were too small to achieve a large beneficial impact. Ayush et al. [44] achieved the spatial alignment of RSIs using the spatiotemporal structure of the images and considered regions of both temporal and geographic proximity for positive sampling in order to sample positive samples for self-supervised contrastive learning. This method was able to expand the capabilities of data augmentation to some extent. Jean et al. [45] extended Word2Vec to remote sensing datasets based on the fact that words in similar contexts in natural language often have similar meanings and accordingly proposed the Tile2Vec method to sample positive samples from augmentations of only the corresponding anchor sample itself to augmentations of similar samples, making it possible to select more diverse positive samples without relying on labels. Jung et al. [46] proposed using multiple positive samples for learning based on Tile2Vec, i.e., the fusion of features, as a way to avoid the impact of false-positive samples on the network.

The above methods suffer from limitations such as relying solely on clustering results or a priori knowledge of random sampling results; consequently, the constraints on the sampling methods for positive samples are too restrictive, meaning that the selected positive samples are not optimal. To overcome these limitations, we established a doubly constrained sampling and discrimination method for positive samples under relaxed conditions to ensure that the obtained samples had greater similarity to the anchor samples while controlling the distribution of different sample instances of the same class in the feature space, thereby newly realizing the leap from instance-level feature learning to class-level feature learning.

2.2. Method

This section introduces our proposed RiSSNet architecture. The overall framework of the network is summarized and the RiSSNet schematic is presented in Section 2.2.1, the relaxed identity sampling principle is introduced in Section 2.2.2, and the relaxed identical sample discrimination strategy is discussed in Section 2.2.3.

2.2.1. Overview of RiSSNet

RiSSNet is a two-branch contrastive learning network built under relaxed identity constraints. The two branches of the network run in parallel. One branch learns the features of each sample itself by enforcing strict identity constraints to enhance the discriminative power of the base network. The other branch performs sampling and discrimination at the patch level with relaxed identity constraints, enabling feature augmentation learned from the instance level to some extent. The specific network structure is shown in Figure 1. The two-branch network consists of four main parts: data augmentation, an encoder, a projection head, and a loss function.

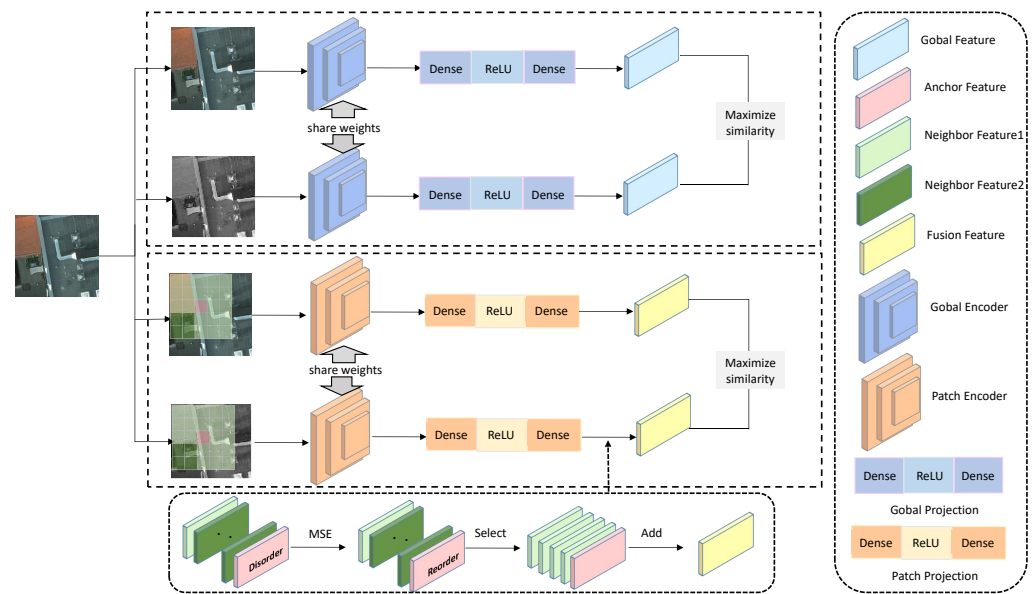


Figure 1. Overall structure of RiSSNet. The network consists of two branches for comparison, where one branch still performs instance-level contrastive learning, and the other branch performs sampling and discrimination with relaxed identity constraints.

RiSSNet first performs data augmentation on the input image. The data augmentation methods used in this paper were random cropping, color distortion, and Gaussian blur. Specifically, for a given anchor sample x_i and the remaining samples $y_1, y_2, \dots, y_{N-1}$ in the batch, the positive sample pair $t(x_i)t'(x_i)$ and the negative sample pair $t(y_1)t'(y_1) \dots t(y_{N-1})t'(y_{N-1})$ were obtained after data augmentation. Based on the data augmentation results, relaxed identity sampling was performed in the local neighborhood of the anchor sample in the data-augmented image, and all samples in the neighborhood were grouped into the relaxed identical sample cohort. The data-augmented image and the relaxed identical sample cohort were passed to the encoder of the strict identity sampling branch and the encoder of the relaxed identity sampling branch, respectively, and each encoder performed feature extraction. The specific formula of the encoder part is as follows:

$$h_{x_i} = f_{x_i} = e(t(x_i)) \quad (1)$$

where h_{x_i} represents the output after the pooling layer, and $e()$ represents the feature vector after the encoder.

Feature mapping was then performed by the projection head. The projection head was composed of an MLP with a hidden layer. The output after the projection head was

$$z_{x_i} = g(h_{x_i}) = W^{(2)}\sigma(W^{(1)}h_{x_i}) \quad (2)$$

where z_{x_i} is the representation output of the projection head, $g()$ is the projection head, and σ is the ReLU layer. After the projection head obtained the strictly identical sample features and the relaxed identical sample features, the relaxed identical sample features were used to discriminate the relaxed identical samples, the MSE loss was used to reorder the relaxed identical sample queue, and the samples at the front end of the queue were selected as the final positive samples under the relaxed identity hypothesis. Finally, the InfoNCE loss between the positive sample features and the anchor sample features under the strict identity hypothesis and the relaxed identity hypothesis was calculated separately to complete the learning of the image features by RiSSNet.

2.2.2. Relaxed Identity Sampling

The relaxed identity sampling strategy is a self-supervised sampling strategy without labels. In order to obtain similar samples for the anchor samples, we performed the neighborhood sampling of the anchor samples by virtue of the fact that the ground objects that are closer to the anchor samples in remote sensing have a stronger correlation than those that are farther away from the anchor samples. We defined the semantic similarity index to measure whether the anchor samples had the same semantics as the samples in the sampling queue (Figure 2). In this paper, semantic similarity is defined as the label similarity between anchor samples and sampled samples. The larger the semantic similarity, the more the sample matches the expectation.



Figure 2. Schematic diagram of the sampling of relaxed identity samples. Based on the fact that the closest neighbors of remote sensing features tend to have greater similarity, we sampled the anchor sample neighborhood, and the extracted samples were recorded as the positive sample queue. The red box indicates the anchor sample, and each white box represents an extracted neighborhood sample, all of which together form the sample queue.

We assumed that the dataset D contains a total of n classes of samples, i.e., $D = \{i_1, i_2, \dots, i_n\}$, and the extracted local area is denoted by d . According to the semantic similarity defined in this paper, if the extracted local area contains features of only a single class, e.g., $d_a = \{i_1\}$, then its semantic similarity to the sample $d_n = \{i_1\}$, which also has only features of the same single class, is maximal. If there are multiple types of feature in the extracted local area, e.g., $d_a = \{i_1, i_2, i_3\}$, then its semantic similarity to another sample with one or more of the same feature categories is defined as follows: $d_n = \{i_1, i_2, i_3\} > d_n = \{i_1, i_2\} > d_n = \{i_1\}$, with a larger percentage of shared feature classes corresponding to a larger similarity. Specifically, we denote the anchor sample by x_i , whose size is $H \times W \times C$. With the centroid of the anchor sample as the center, the neighborhood of the anchor sample is defined by a radius R . The image was sampled within the neighborhood radius, and the corresponding image samples, each with a size of $H \times W \times C$, are denoted by $x_{Rj}(j \in N)$. A total of N positive samples were obtained, i.e., the sample queue before relaxed identical sample discrimination contained a total of N positive samples.

2.2.3. Relaxed Identical Sample Discrimination

The relaxed identical sample discrimination strategy was based on anchor sample features and features in the relaxed sample cohort for scoring, for which we defined feature similarity. The feature similarity is an important constraint of the relaxed identical sample

discrimination strategy and allows for greater visual variability in the case of semantic similarity. In contrast, when only traditional data augmentation is used to learn different transformations of a single sample, often the full diversity of possible representations is difficult to learn.

To obtain positive samples with high feature similarity as described above, we used the self-discriminatory ability of the network, i.e., we used the network itself to discriminate and score the samples of the relaxed identical sample cohort. In the early stage of training, due to the weak discriminative ability of the network, we used only a single-branch contrastive learning network; later, we loaded the single-branch network parameters for two-branch network training, adding the network branch responsible for relaxed identity sampling and discrimination. To achieve instance-level to class-level learning, we selected four samples from the scored relaxed identical sample queue that exhibited large differences in feature similarity as positive samples and performed feature fusion. Specifically, the anchor sample is denoted by x_i , and the positive sample cohort is denoted by x_{R_j} ($j \in N$). We computed the similarity between the anchor sample and each sample in the positive sample cohort as shown in Equation (3), where the feature difference between two samples is computed as the mean squared difference, and ranked the samples according to their feature differences.

$$D(x_i, x_{R_j}) = (x_i - x_{R_j})^2 \quad (3)$$

We then selected the top k samples to undergo feature fusion to prevent the impact of false-positive samples on the network, as follows:

$$h_{x'} = \frac{1}{k} \sum_{j=1}^{j \in k} h_{x_{R_j}} \quad (4)$$

where $h_{x_{R_j}}$ denotes the representation extracted from the j -th positive sample in the positive sample queue after the encoder.

2.2.4. InfoNCE Loss

To draw the anchor sample, the relaxed identical samples, and their augmentations closer together in the feature space, the anchor sample was pushed farther from the remaining negative samples in the batch to control the spatial distribution of the features of different instances in the same class. We defined the loss between the anchor sample and the relaxed identical samples as shown in Equation (5). Specifically, we denoted the features of the anchor sample after the projection head by z and the features of the remaining negative samples in the batch after the projection head by z_{x_k} . The anchor sample was approximated by the fused relaxed identical positive samples and pushed away from the other samples in the batch, i.e.,

$$l = -\log \frac{\exp(\text{sim}(z, h_{x'})/\tau)}{\exp(\text{sim}(z, h_{x'})/\tau) + \sum_{k=1}^{2(N-1)} \exp(\text{sim}(z, z_{x_k})/\tau)} \quad (5)$$

where N is the number of samples in the batch and a total of $2N$ samples are obtained after data augmentation, meaning that $2(N-1)$ is the number of samples minus the positive samples and their augmentations, and $\text{sim}(z, h_{x'})$ denotes the cosine similarity between the anchor sample z and the relaxed sampling and discrimination positive sample $h_{x'}$.

3. Experiments

3.1. Datasets

We validated the effectiveness of the method on three remote sensing semantic segmentation datasets: Xiangtan [39], Potsdam [47], and GID [48], dividing the training and validation sets at a ratio of 9:1 and taking 1% of the training data for fine-tuning to verify the results of

the upstream pretrained model effects. The three images had different resolutions, but we uniformly cropped the images to 256×256 , and the data are shown in detail in Table 1.

Table 1. Summary of the datasets. The Potsdam and GID datasets are from an open network, and the Xiangtan dataset contains real data collected from the city of Xiangtan, Hunan Province, China.

Dataset	Resolution	Image Size	Classes	Training	Fine-Tuning	Val
Xiangtan [39]	2 m	256×256	9	16051	160	3815
ISPRS Potsdam [47]	0.05 m	256×256	6	16080	160	4022
GID [48]	1 m	256×256	5	98218	983	10919

3.1.1. Xiangtan

The Xiangtan dataset was collected from the city of Xiangtan, Hunan Province, and captured by the Gaofen 2 satellite with a resolution of 2 m. It contains image data labeled as background, farm land, urban, rural, water, forestland, grass, road, and other, for a total of nine feature classes, of which forestland accounts for a relatively large proportion, and the features show an unbalanced distribution. We cropped 106 images of size 4096×4096 to images of size 256×256 and obtained 16051 training data samples, 160 fine-tuning data samples, and 3815 validation data samples.

3.1.2. ISPRS Potsdam

The ISPRS Potsdam dataset was collected from the city of Potsdam, Germany, with a spatial resolution of 5 cm. Six classes of features are included: water, buildings, vegetation, trees, cars, and other, with water accounting for a large proportion and cars accounting for a very small proportion. We cropped 38 images of size 6000×6000 to images of size 256×256 and obtained 16080 training data samples, 160 fine-tuning data samples, and 4022 validation data samples.

3.1.3. GID

The GID dataset covers more than 60 cities in China and was captured by the Gaofen-2 satellite with a spatial resolution of 1 m. It contains five classes of features: buildings, farmland, forests, grass, and water, which are more balanced than the classes in the other datasets. We cropped 150 images of approximately size 6800×7300 to images of size 256×256 and obtained 98218 training data samples, 983 fine-tuning data samples, and 10919 validation data samples.

3.2. Experiments

We mainly compared the experimental results obtained under the relaxed identity hypothesis with those obtained under the strict identity hypothesis using the following baselines: Random [28], Inpainting [49], Tile2Vec [45], SimCLR [9], MoCo v2 [11], Barlow Twins [50], and FALSE [51]. The specific experimental details are as follows: the pretraining of an instance-level network was conducted first, and after 20 epochs, our network was loaded and trained for 130 epochs. The chosen optimizer was Adam, with the weight decay set to 10^{-5} and the batch size set to 32. In the fine-tuning part, we used 1% of the training data set for training, the number of training epochs was 100, and the Adam optimizer and a batch size of 32 were used again.

- Random [28]: a supervised learning method, using the network Deeplab V3+ and only 1% of the fine-tuning data volume for training.
- Inpainting [49]: a patch-level generative network that learns by computing the loss of the patched image relative to the original image.
- Tile2Vec [45]: a primitive sampling method based on spatial proximity, with samples within the neighborhood treated as positive and samples at a farther distance treated as negative.

- (d) SimCLR [9]: a standard strict-identity-constrained network for which positive samples are obtained through the data augmentation of the anchor samples and negative samples are obtained from the rest of the samples in the same batch.
- (e) MoCo v2 [11]: a standard strict-identity-constrained network that outperforms SimCLR in several respects, for which positive samples are also obtained by augmenting anchor samples and negative samples are stored in a queue.
- (f) Barlow Twins [50]: a self-supervised network with only positive samples, where the positive samples are also augmented by the anchor samples.
- (g) FALSE [51]: a self-supervised network that removes false-negative samples from negative samples in contrastive learning.

Among the methods compared above, the 'Random' method used 1% of the data for supervised training, and the rest of the self-supervised methods used 1% of the data for fine-tuning the network. From Table 2, we can see that our method achieved improved metrics on each dataset compared to the baseline methods, especially on the GID dataset, where the mIOU was improved by 2.77%. On the Xiangtan dataset, the overall metrics of our method showed less improvement compared to the optimal baseline method, while on the GID dataset, our method showed the most improvement compared to the optimal baseline method. Next, we visualized the accuracy for each category of features in the different datasets, and it can be seen from Figure 3 that our method maintained the highest accuracy for each category of features in the three datasets. On the Xiangtan dataset, our method achieved the greatest improvements for urban and water; on the Potsdam dataset, our method showed the most significant improvement in the car category; and on the GID dataset, the improvements were more balanced due to the more balanced nature of the feature categories. In addition, we also visualized the semantic segmentation result graph Figure 4.

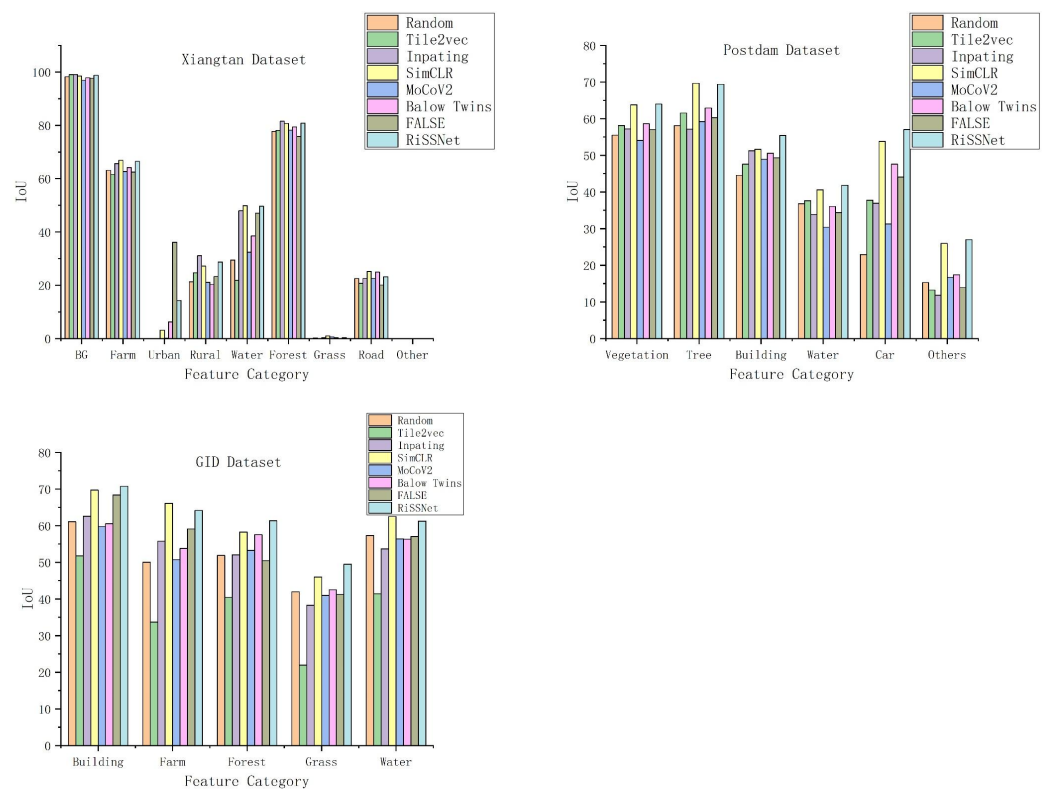


Figure 3. Plots of the IoU for the various classes in the three datasets. Each graph represents a dataset, and the different feature accuracies in the dataset are plotted in different colors.

Table 2. Experimental results comparing our method (RiSSNet) with current advanced methods in terms of three evaluation indicators: Kappa, OA, and mIOU. Bold numbers represent the maximum value under this indicator.

Method	Xiangtan			Potsdam			GID		
	Kappa	OA	mIOU	Kappa	OA	mIOU	Kappa	OA	mIOU
Random [28]	65.31	78.52	34.73	53.78	64.87	38.85	52.61	72.14	52.45
Inpainting [49]	65.40	78.45	34.93	55.74	66.28	35.45	55.94	73.82	52.48
Tile2Vec [45]	64.45	77.73	33.98	56.85	70.57	36.33	32.30	61.19	37.84
SimCLR [9]	70.08	81.12	39.17	63.40	72.02	43.63	65.18	79.40	58.63
MoCo v2 [11]	65.40	78.45	34.93	54.20	65.13	34.37	52.31	71.56	52.25
Barlow Twins [50]	67.55	79.85	36.86	58.59	68.44	39.23	54.80	72.79	54.14
FALSE [51]	69.41	80.99	40.27	55.96	66.31	43.15	60.98	77.52	55.25
RiSSNet	70.16	81.21	40.25	64.81	73.08	44.96	66.10	80.21	61.41

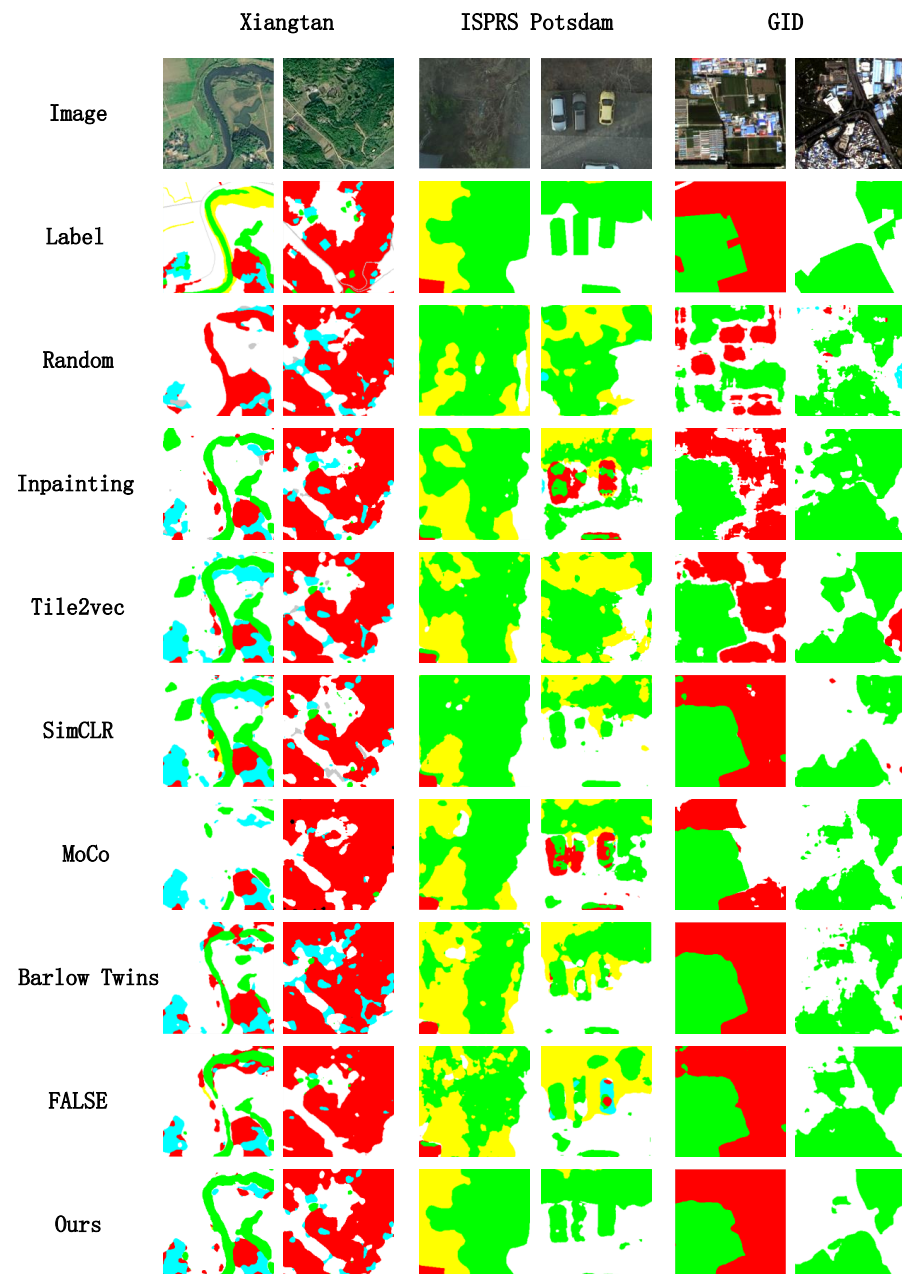


Figure 4. Visualization of semantic segmentation results on different datasets.

3.3. Result Analysis

To verify whether the samples obtained under the semantic similarity and feature similarity constraints of the RiSSNet method were positive samples with respect to the anchor samples, we compared the numbers of positive and negative samples obtained under both constraints with those obtained through random selection while visualizing the images and labels of the selected regions and observed whether our method enriched the number of in-class samples. To verify the invariance of the features learned by the model, we used t-SNE visualizations of the feature distributions to observe the variations between different classes of sample features as learned by the model and illustrate whether samples within the same class were sufficiently closely distributed, and we also calculated the distances between the features within a class.

3.3.1. Relaxed Identity Sampling

To verify whether the semantic similarity was improved under relaxed identity sampling, we calculated and visualized the positive and negative sample statistics for the different datasets. Positive samples were defined as those with the same single feature category if the anchor sample contained only one feature category or as those with the maximum percentage of the same feature categories if the anchor sample contained multiple feature categories.

As seen from the Table 3, compared to the random selection of the positive sample cohort, there was an approximately 8% increase in the number of positive samples and an approximately 18% decrease in the number of negative samples under relaxed identity sampling. In addition, we visualized the images and labels of patches selected under the random selection approach (Rs) and patches selected under the joint semantic similarity approach (Ss) to see whether the samples selected by our method increased the number of positive samples. As shown in the Figure 5, the conclusions obtained from the visualizations were consistent with those obtained in the table.

Table 3. This table shows the numbers of positive and negative samples extracted by the relaxed identity sampling method and through the random selection of positive samples and their trends, where ‘Anchor’ is the anchor sample, ‘N_pos’ represents positive samples, ‘N_neg’ represents negative samples, ‘Random’ is the random extraction method, and ‘Trend’ indicates whether the number of positive or negative samples extracted by our method increased or decreased compared to the random method, with ‘↑’ meaning that our method increased the number of samples compared to the random method and ‘↓’ meaning that our method decreased the number of samples compared to the random method.

Selection Method	Xiangtan			Potsdam			GID		
	Anchor	N_pos	N_neg	Anchor	N_pos	N_neg	Anchor	N_pos	N_neg
Random	16064	41305	22951	7257	39027	25101	98272	308316	84772
Our Method	16064	42395	21733	7257	46180	18076	98272	325000	68088
Trend	-	↑1.02	↓0.94	-	↑1.18	↓0.72	-	↑1.05	↓0.80

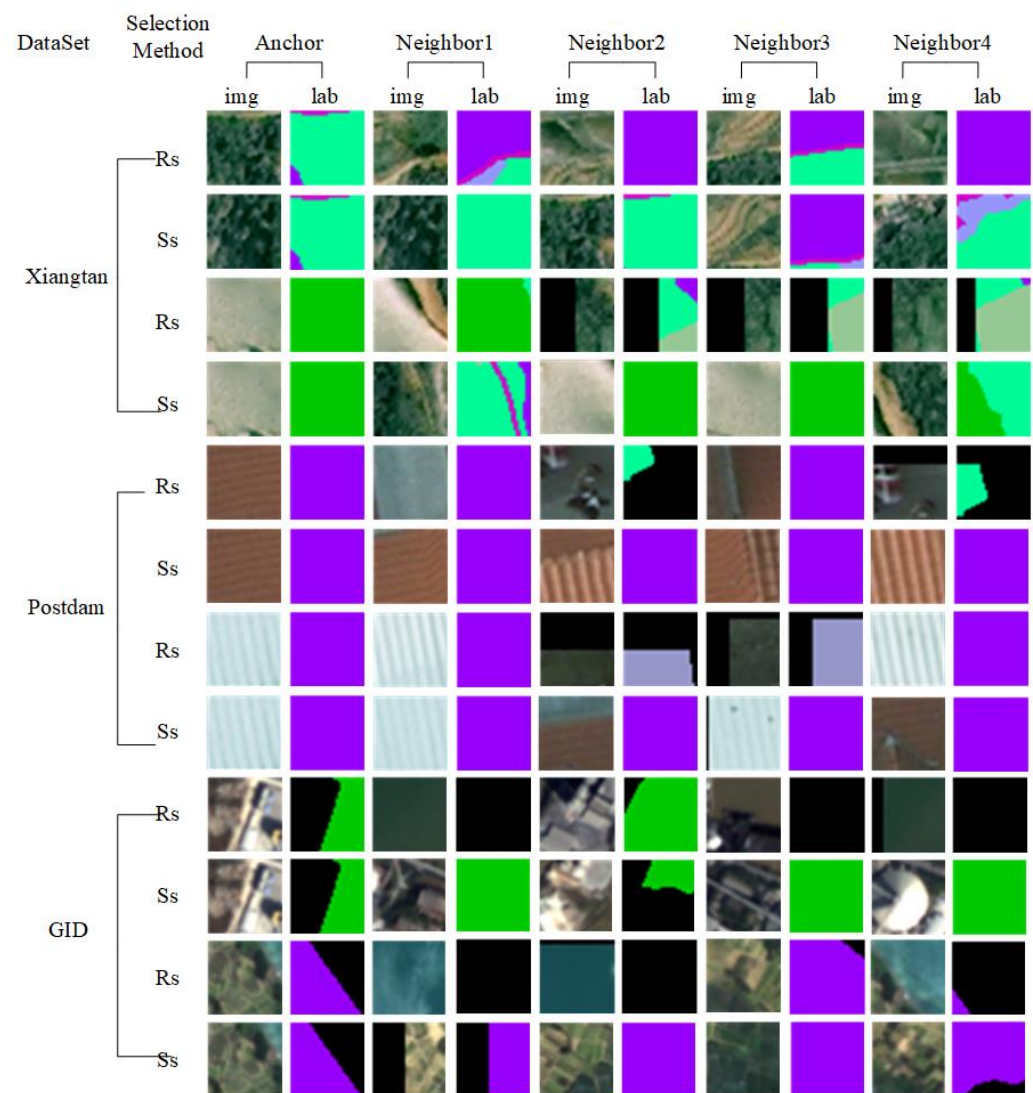


Figure 5. Visualization of the results of relaxed identity sampling and random sampling, where 'Rs' represents the random selection method and 'Ss' represents the method considering semantic similarity. For each sample, the 'img' column presents the cropped image region, and the 'lab' column presents the label map of the cropped region, with different colors representing different feature categories.

3.3.2. Relaxed Identical Sample Discrimination

To verify whether the feature similarity was improved under relaxed identical sample discrimination, we statistically analyzed and visualized the corresponding results on different datasets. Again, we defined positive and negative samples in the same way as in the section on relaxed identity sampling, and the results for the random method were the numbers of positive and negative samples obtained by randomly selecting samples without the relaxed identical sample discrimination strategy.

From Table 4, it can be seen that the number of positive samples extracted by our method after relaxed identical sample discrimination was also increased compared with the random selection method. Additionally, from the image visualization results in Figure 6, it can be observed that there was some variability between the anchor samples and the positive samples obtained by our method, and the improvement brought by this difference could not be obtained through data augmentation. To verify the feature similarity, we again visualized some of the positive samples in the datasets.

Table 4. This table shows the numbers and trends of positive and negative samples extracted with relaxed identical sample discrimination and with the random selection method, where ‘Anchor’ is the anchor sample, ‘N_Pos’ represents positive samples, ‘N_Neg’ represents negative samples, ‘Random’ is the sample selection method, and ‘Trend’ indicates whether the number of positive or negative samples extracted by our method with relaxed identical sample discrimination was increased or decreased compared to the random selection method, with ‘↑’ indicating an increase in the number of samples compared to the sample sampling method and ‘↓’ indicating a decrease in the number of samples extracted by our method compared to the random selection method.

Selection Method	Xiangtan			Potsdam			GID		
	Anchor	N_pos	N_neg	Anchor	N_pos	N_neg	Anchor	N_pos	N_neg
Random	16000	36110	27890	4000	8950	7050	98240	277515	115445
Our Method	16000	37919	26081	4000	10279	5721	98240	308359	84601
Trend	-	↑1.05	↓0.93	-	↑1.14	↓0.81	-	↑1.11	↓0.73

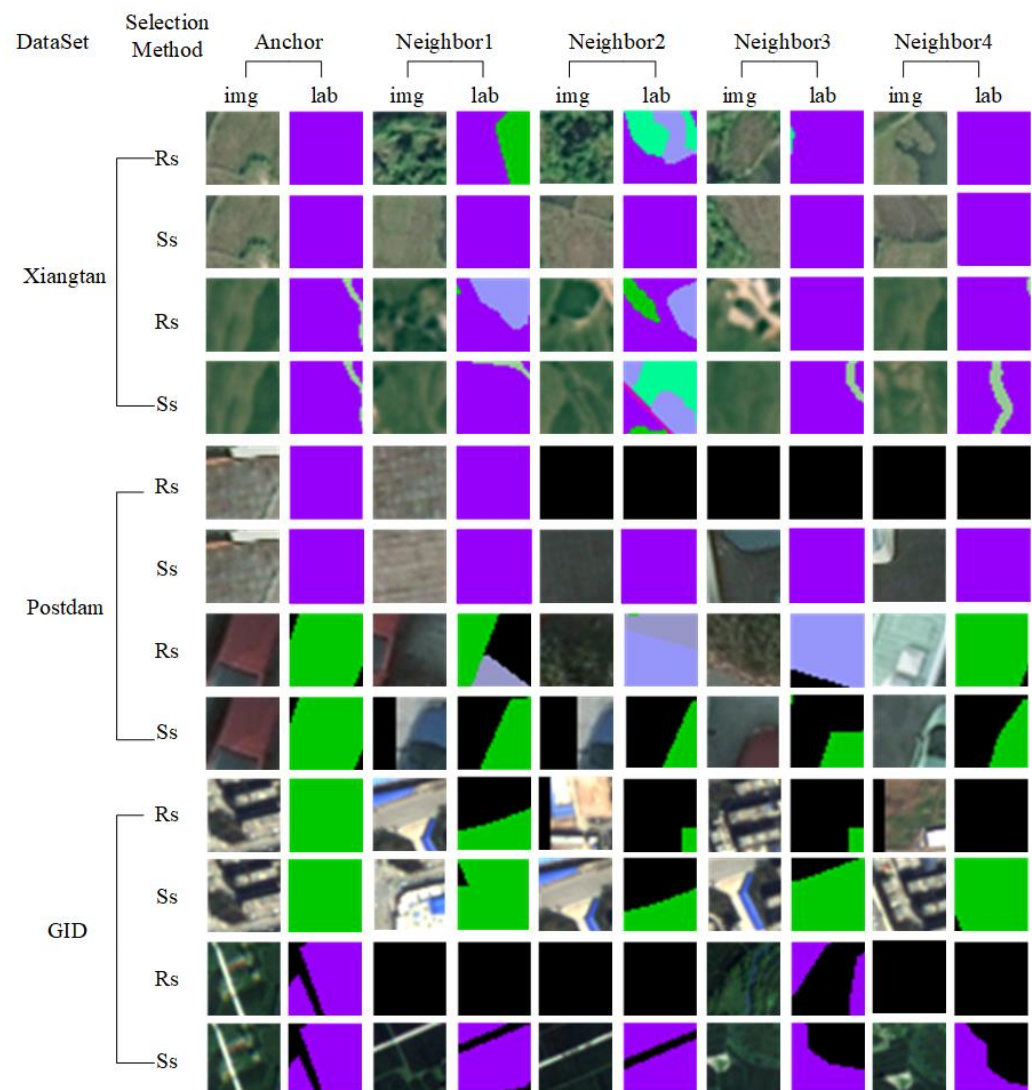


Figure 6. Visualization of the results of relaxed identical sample discrimination, where ‘Rs’ represents the random selection method and ‘Ss’ represents our method considering semantic similarity. For each sample, the ‘img’ column presents the cropped image region, and the ‘lab’ column presents the label map of the cropped region, with different colors representing different feature categories.

3.3.3. Feature Invariance Augmentation Verification Experiments

For the feature invariance augmentation validation experiments, we randomly selected some data from each dataset for feature invariance validation. Some feature categories were missing in the validation experiments for the Xiangtan dataset and Potsdam dataset because of the serious imbalance in the feature categories in these datasets. In addition, we visualized the distribution of the features learned by the different methods on the three datasets, using the GID dataset as an example, as shown in Figure 7.

As seen from this figure, our method yielded more compact features for the GID dataset, reflecting an improvement in the feature invariance for this dataset. To further verify the improvement in the feature invariance in the experimental results of the proposed method, we randomly extracted some of the data from each dataset to calculate the intra-class distances. Due to the uneven distributions of features in the Xiangtan and Potsdam datasets, some categories of features were not extracted, and the corresponding intraclass distances were 0. We calculated the distance between intraclass features as follows, where C_i denotes the i -th feature class, 'Intra-class' is the intraclass distance of the i -th feature class, and 'Best' represents the method with the highest accuracy among the current methods (excluding our method).

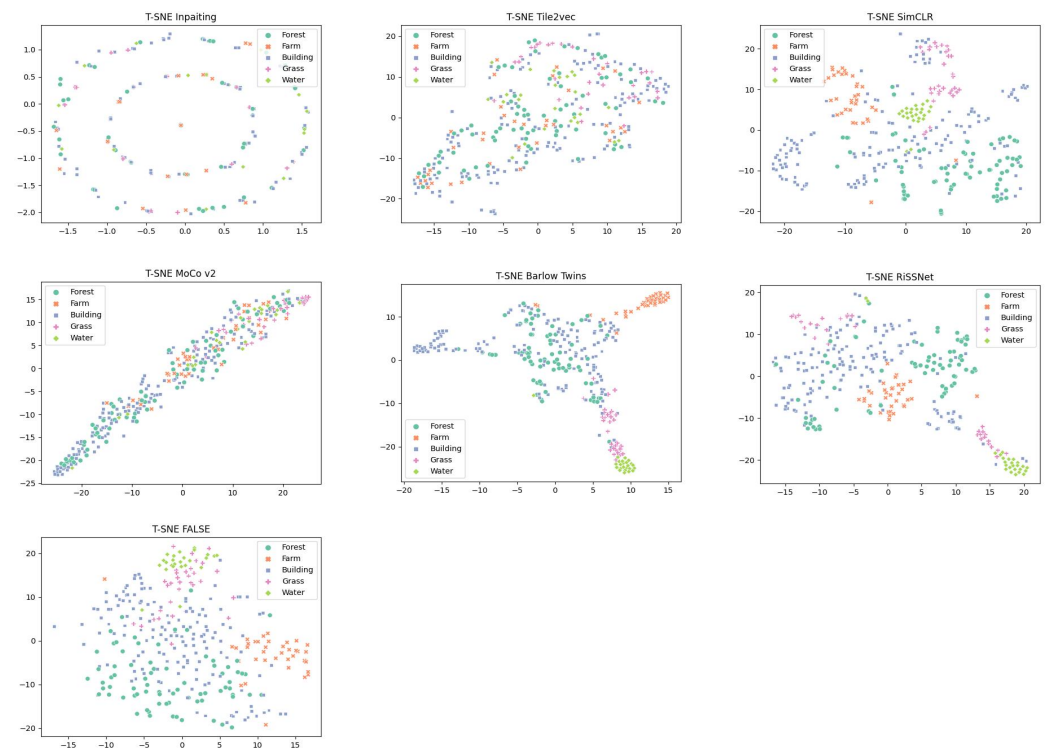


Figure 7. Graphs of t-SNE visualizations of the results, where differently colored dots represent different types of class objects.

The Table 5 shows that our method further reduced the intraclass distances between samples in each dataset compared to the 'Best' method, achieving the closer clustering of similar samples and an improvement in the invariance of the sample characteristics. The improvement in intraclass distance was most obvious in the Potsdam dataset and less obvious in the Xiangtan and GID datasets, but our method did have a narrowing effect for each class of features.

Table 5. This table shows the intraclass distances between features of the same class in different datasets. ‘Best’ represents the intraclass distance of the best method among the current baseline methods excluding our method, and ‘Ours’ denotes our method. ‘ $\rightarrow\leftarrow$ ’ means that our method made the intraclass distance shorter than the best method, and ‘ $\leftarrow\rightarrow$ ’ means that our method made the intraclass distance greater than the best method.

Class	Xiangtan			Potsdam			GID		
	Best	Ours	Trend	Best	Ours	Trend	Best	Ours	Trend
C_1	80.27	127.97	$\leftarrow\rightarrow$	2748	1907	$\rightarrow\leftarrow$	1723	1152	$\rightarrow\leftarrow$
C_2	214.63	118.27	$\rightarrow\leftarrow$	4594	3656	$\rightarrow\leftarrow$	769	237	$\rightarrow\leftarrow$
C_3	0	0	-	3782	2372	$\rightarrow\leftarrow$	1.75	1.77	$\rightarrow\leftarrow$
C_4	0	0	-	1010	684	$\rightarrow\leftarrow$	22.62	18.87	$\rightarrow\leftarrow$
C_5	0.78	0.28	$\rightarrow\leftarrow$	0	0	-	37.41	23.55	$\rightarrow\leftarrow$
C_6	7118	5000	$\rightarrow\leftarrow$	1799	1081	$\rightarrow\leftarrow$	/	/	/
C_7	0	0	-	0	0	-	/	/	/
C_8	0	0	-	/	/	/	/	/	/
C_9	0	0	-	/	/	/	/	/	/

4. Discussion

In this paper, to address the transition from instance-level feature learning to class-level feature learning in contrastive learning, we proposed a relaxed identity hypothesis, validated the effectiveness of the proposed method through experiments on three datasets, and further verified that the relaxed identity hypothesis played a role in learning the invariance of features within classes. Our two-branch network design avoided bias during initial training and guided sample selection to form the relaxed identical sample queue in addition to the sampling and discrimination of the relaxed identical samples by means of a double constraint on the positive sample queue.

An increase in resolution with the same patch size yielded better results under the relaxed identity hypothesis. We experimentally observed that under the condition that the sizes of the training and fine-tuning datasets were the same, a resolution increase led to a greater improvement under the relaxed identity hypothesis; for the example of the Xiangtan and Potsdam datasets, both datasets had training and fine-tuning sets of almost the same size, but the relative index gains were higher for the Potsdam dataset. Furthermore, from the visualizations of anchor samples and relaxed identical samples and from the figure analysis results, we could see that due to the improved resolution, the collection of features contained within a single sample obtained from our positive sample cohort was more uniform and singular, leading to a significant increase in the number of positive samples drawn from the relaxed identical sample cohort, which was a major reason for the improved experimental results.

An increase in the number of training samples for a dataset led to a more significant improvement under the relaxed identity hypothesis. In the comparison between the GID dataset and the Potsdam dataset, it could be seen that the improvement in the mIOU on the GID dataset was more significant compared to the baseline, corresponding to a gain of 2.77%.

The relaxed identity sampling principle was beneficial for helping the network learn feature invariance. As seen from the results of the t-SNE visualization and the calculated intraclass distances, the relaxed identity sampling principle made the intraclass distances smaller, i.e., the classes were more compact. Compared to the experimental results of the best baseline method, our method made the classes more compact.

5. Conclusions

In this paper, we addressed the shortcomings of the strict identity principle used in existing contrastive learning methods, which inherently leads to instance-level feature learning, and proposed a relaxed identity principle for application to RSIs to bridge the gap from instance-level feature learning to class-level feature learning in order to improve

a model's learning of feature invariance. In contrast to previous studies, we switched from a single-constraint focus to a double-constraint focus by defining two constraints on semantic similarity considering relaxed sample identity and feature similarity under relaxed identical sample discrimination to sample and discriminate the relaxed sample cohort. Our method surpassed existing contrastive learning methods under the strict identity hypothesis, such as SimCLR, MoCo v2, and Barlow Twins, and its effectiveness was validated on three remote sensing datasets. However, our relaxed identity sample queue still needs to be further optimized—for example, by considering whether a priori knowledge could be used more effectively for the selection of relaxed identical samples, whether we could establish better criteria and indicators for the discrimination of relaxed identical samples, how samples from different images could be linked in our method, and how best to choose the patch size for images of different resolutions. These issues are all worthy of future study.

Author Contributions: Conceptualization, methodology, software, and writing (original draft preparation): W.J.; supervision: H.L. and J.Q.; software and validation: G.W. and K.W.; data curation: M.S. and P.L.; writing (reviewing and editing): L.L. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: Chongqing Natural Science Foundation Project (cstc2021jcyj-msxmX1203), Chongqing Talent Plan “Contract System” Project (CSTC2021ycjh bgzxm0294), Major Special Project of High-Resolution Earth Observation System (86-Y50G27-9001-22/23).

Data Availability Statement: Data associated with this research are available online. The Potsdam dataset is available at <http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html> (accessed on 20 October 2020). The GID dataset is available for download at <http://captain.whu.edu.cn/repository.html> (accessed on 22 April 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

- Ming, D.; Ci, T.; Cai, H.; Li, L.; Qiao, C.; Du, J. Semivariogram-Based Spatial Bandwidth Selection for Remote Sensing Image Segmentation With Mean-Shift Algorithm. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 813–817. [CrossRef]
- Wang, J.; Qin, Q.; Li, Z.; Ye, X.; Wang, J.; Yang, X.; Qin, X. Deep Hierarchical Representation and Segmentation of High Resolution Remote Sensing Images. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 4320–4323. [CrossRef]
- Liu, X.; Chi, M.; Zhang, Y.; Qin, Y. Classifying High Resolution Remote Sensing Images by Fine-Tuned VGG Deep Networks. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 7137–7140. [CrossRef]
- Ma, W.; Pan, Z.; Guo, J.; Lei, B. Super-Resolution of Remote Sensing Images Based on Transferred Generative Adversarial Network. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 1148–1151. [CrossRef]
- Misra, I.; van der Maaten, L. Self-Supervised Learning of Pretext-Invariant Representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
- Favaro, S.J.P. Self-Supervised Feature Learning by Learning to Spot Artifacts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
- Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9912–9924.
- Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the International Conference on Machine Learning (PMLR), Paris, France, 29 March–1 April 2020; pp. 1597–1607.
- Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; Hinton, G.E. Big self-supervised models are strong semi-supervised learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22243–22255.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
- Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.
- Chen, X.; Xie, S.; He, K. An Empirical Study of Training Self-Supervised Vision Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, online, 11–17 October 2021; pp. 9640–9649.

13. Tao, C.; Qi, J.; Guo, M.; Zhu, Q.; Li, H. Self-supervised remote sensing feature learning: Learning Paradigms, Challenges, and Future Works. *arXiv* **2022**, arXiv:2211.08129.
14. Kim, Y.; Park, W.; Roh, M.C.; Shin, J. GroupFace: Learning Latent Groups and Constructing Group-Based Representations for Face Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5620–5629. [[CrossRef](#)]
15. Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; Isola, P. What makes for good views for contrastive learning? *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6827–6839.
16. Li, B.; Hou, Y.; Che, W. Data augmentation approaches in natural language processing: A survey. *AI Open* **2022**, *3*, 71–90. [[CrossRef](#)]
17. Zhu, J.; Han, X.; Deng, H.; Tao, C.; Zhao, L.; Wang, P.; Lin, T.; Li, H. KST-GCN: A Knowledge-Driven Spatial-Temporal Graph Convolutional Network for Traffic Forecasting. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15055–15065. [[CrossRef](#)]
18. Tao, C.; Qi, J.; Zhang, G.; Zhu, Q.; Lu, W.; Li, H. TOV: The Original Vision Model for Optical Remote Sensing Image Understanding via Self-Supervised Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 4916–4930. [[CrossRef](#)]
19. Nguyen, X.B.; Lee, G.S.; Kim, S.H.; Yang, H.J. Self-supervised learning based on spatial awareness for medical image analysis. *IEEE Access* **2020**, *8*, 162973–162981. [[CrossRef](#)]
20. Minaee, S.; Cambria, E.; Gao, J. Deep Learning Based Text Classification: A Comprehensive Review. *Acm Comput. Surv.* **2021**, *54*, 62. [[CrossRef](#)]
21. Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; Zisserman, A. With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9568–9577.
22. Goyal, P.; Duval, Q.; Seessel, I.; Caron, M.; Misra, I.; Sagun, L.; Joulin, A.; Bojanowski, P. Vision Models Are More Robust And Fair When Pretrained On Uncurated Images Without Supervision. *arXiv* **2022**, arXiv:2202.08360.
23. Ding, Y.; Zhang, Z.; Zhao, X.; Cai, Y.; Li, S.; Deng, B.; Cai, W. Self-Supervised Locality Preserving Low-Pass Graph Convolutional Embedding for Large-Scale Hyperspectral Image Clustering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [[CrossRef](#)]
24. Ding, C.; Zheng, M.; Chen, F.; Zhang, Y.; Zhuang, X.; Fan, E.; Wen, D.; Zhang, L.; Wei, W.; Zhang, Y. Hyperspectral image classification promotion using clustering inspired active learning. *Remote Sens.* **2022**, *14*, 596. [[CrossRef](#)]
25. Jia, X.; Xie, Y.; Li, S.; Chen, S.; Zwart, J.; Sadler, J.; Appling, A.; Oliver, S.; Read, J. Physics-Guided Machine Learning from Simulation Data: An Application in Modeling Lake and River Systems. In Proceedings of the 2021 IEEE International Conference on Data Mining (ICDM), Auckland, New Zealand, 7–10 December 2021; pp. 270–279. [[CrossRef](#)]
26. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
27. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, 18th International Conference, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; Part III, pp. 234–241.
28. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
29. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
30. Wu, Q.; Luo, F.; Wu, P.; Wang, B.; Yang, H.; Wu, Y. Automatic Road Extraction from High-Resolution Remote Sensing Images Using a Method Based on Densely Connected Spatial Feature-Enhanced Pyramid. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3–17. [[CrossRef](#)]
31. Ding, L.; Tang, H.; Bruzzone, L. LANet: Local Attention Embedding to Improve the Semantic Segmentation of Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 426–435. [[CrossRef](#)]
32. Wu, Z.; Xiong, Y.; Yu, S.X.; Lin, D. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3733–3742.
33. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. *arXiv* **2017**, arXiv:1710.09412.
34. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
35. Liu, J.; Liu, B.; Zhou, H.; Li, H.; Liu, Y. TokenMix: Rethinking Image Mixing for Data Augmentation in Vision Transformers. In Proceedings of the Computer Vision—ECCV 2022, 17th European Conference, Tel Aviv, Israel, 23–27 October 2022.
36. Li, H.; Cao, J.; Zhu, J.; Luo, Q.; He, S.; Wang, X. Augmentation-Free Graph Contrastive Learning of Invariant-Discriminative Representations. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–11. [[CrossRef](#)]
37. Tao, C.; Fu, S.; Qi, J.; Li, H. Thick Cloud Removal in Optical Remote Sensing Images Using a Texture Complexity Guided Self-Paced Learning Method. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5619612. [[CrossRef](#)]

38. Peng, X.; Wang, K.; Zhu, Z.; Wang, M.; You, Y. Crafting Better Contrastive Views for Siamese Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16031–16040.
39. Li, H.; Li, Y.; Zhang, G.; Liu, R.; Huang, H.; Zhu, Q.; Tao, C. Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. [\[CrossRef\]](#)
40. Tang, M.; Georgiou, K.; Qi, H.; Champion, C.; Bosch, M. Semantic Segmentation in Aerial Imagery Using Multi-level Contrastive Learning with Local Consistency. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–7 January 2023; pp. 3787–3796. [\[CrossRef\]](#)
41. Wang, X.; Zhu, J.; Yan, Z.; Zhang, Z.; Zhang, Y.; Chen, Y.; Li, H. LaST: Label-Free Self-Distillation Contrastive Learning With Transformer Architecture for Remote Sensing Image Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6512205. [\[CrossRef\]](#)
42. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
43. Li, H.; Zhang, X.; Xiong, H. Center-wise local image mixture for contrastive representation learning. *arXiv* **2020**, arXiv:2011.02697.
44. Ayush, K.; Uzket, B.; Meng, C.; Tanmay, K.; Burke, M.; Lobell, D.; Ermon, S. Geography-Aware Self-Supervised Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10181–10190.
45. Jean, N.; Wang, S.; Samar, A.; Azzari, G.; Lobell, D.; Ermon, S. Tile2vec: Unsupervised Representation Learning for Spatially Distributed Data. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 3967–3974.
46. Jung, H.; Oh, Y.; Jeong, S.; Lee, C.; Jeon, T. Contrastive self-supervised learning with smoothed representation for remote sensing. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8010105. [\[CrossRef\]](#)
47. Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Benitez, S.; Breitkopf, U. The isprs benchmark on urban object classification and 3d building reconstruction. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *I-3*, 293–298. [\[CrossRef\]](#)
48. Tong, X.Y.; Xia, G.S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sens. Environ.* **2020**, *237*, 111322. [\[CrossRef\]](#)
49. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context Encoders: Feature Learning by Inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
50. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In Proceedings of the International Conference on Machine Learning, online, 18–24 July 2021.
51. Zhang, Z.; Wang, X.; Mei, X.; Tao, C.; Li, H. FALSE: False Negative Samples Aware Contrastive Learning for Semantic Segmentation of High-Resolution Remote Sensing Image. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6518505. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.