remote sensing

MDPI

Article

JPSSL: SAR Terrain Classification Based on Jigsaw Puzzles and FC-CRF

Zhongle Ren [1], Yiming Lu [1], Biao Hou [1,*], Weibin Li [1] and Feng Sha [2]

[1] The School of Artificial Intelligence, Xidian University, Xi'an 710071, China; zlren@xidian.edu.cn (Z.R.); xidianlym@stu.xidian.edu.cn (Y.L.); weibinli@xidian.edu.cn (W.L.)

[2] High Resolution Earth Observation System Shaanxi Data and Application Center, Xi'an 710061, China; shafeng0107@126.com

* Correspondence: houbiao@mail.xidian.edu.cn

Abstract: Effective features play an important role in synthetic aperture radar (SAR) image interpretation. However, since SAR images contain a variety of terrain types, it is not easy to extract effective features of different terrains from SAR images. Deep learning methods require a large amount of labeled data, but the difficulty of SAR image annotation limits the performance of deep learning models. SAR images have inevitable geometric distortion and coherence speckle noise, which makes it difficult to extract effective features from SAR images. If effective semantic context features cannot be learned for SAR images, the extracted features struggle to distinguish different terrain categories. Some existing terrain classification methods are very limited and can only be applied to some specified SAR images. To solve these problems, a jigsaw puzzle self-supervised learning (JPSSL) framework is proposed. The framework comprises a jigsaw puzzle pretext task and a terrain classification downstream task. In the pretext task, the information in the SAR image is learned by completing the SAR image jigsaw puzzle to extract effective features. The terrain classification downstream task is trained using only a small number of labeled data. Finally, fully connected conditional random field processing is performed to eliminate noise points and obtain a high-quality terrain classification result. Experimental results on three large-scale high-resolution SAR images confirm the effectiveness and generalization of our method. Compared with the supervised methods, the features learned in JPSSL are highly discriminative, and the JPSSL achieves good classification accuracy when using only a small amount of labeled data.

Keywords: synthetic aperture radar (SAR); self-supervised learning; jigsaw puzzle; terrain classification

check for updates

Citation: Ren, Z.; Lu, Y.; Hou, B.; Li, W.; Sha, F. JPSSL: SAR Terrain Classification Based on Jigsaw Puzzles and FC-CRF. *Remote Sens.* **2024**, *16*, 1635. https://doi.org/10.3390/rs16091635

Academic Editor: Stefano Tebaldini

Received: 28 March 2024
Revised: 29 April 2024
Accepted: 30 April 2024
Published: 3 May 2024

1. Introduction

Synthetic aperture radar (SAR) is an active earth observation system. As an active sensor with high resolution, wide coverage, all-weather imaging capability, and strong penetration capabilities, SAR systems have a wide range of uses, such as disaster monitoring [1] and environmental protection [2]. With the development of satellite technology, SAR systems provide a wealth of accurate earth observation images in various military and civil applications [3]. How to effectively interpret images generated by SAR systems has become the focus of current research. In SAR image interpretation, SAR image terrain classification [4] is an important task. The SAR image terrain classification task refers to effectively distinguishing and labeling different contents in SAR images. It has important value in agricultural detection [5], terrain surface classification [6], and tsunami disaster assessment [7]. Passah et al. [8] conducted a comprehensive study and analysis on SAR image terrain classification, proving the wide application of SAR image terrain classification. With remote sensing deep learning development, SAR image terrain classification has a wider range of application scenarios [9].

A key point in the SAR image terrain classification task is how to extract effective features. Traditional algorithms are mostly machine learning approaches based on handcrafted features. Over the past few decades, many traditional algorithms have been applied to feature extraction from SAR images [10]. Various features can be used to describe SAR images, including color, texture, spatial feature relationship, etc. To combine the characteristics of SAR images, Dai et al. [11] proposed a structure-based multi-level local pattern histogram (MLPH) feature and used it in SAR image classification tasks. Ansari et al. [12] used the correlation of multiple texture features to complete the task of urban change detection. The method based on the gray level co-occurrence matrix (GLCM) [13] is also widely used to extract SAR image features. It constructs statistics such as entropy, contrast, and correlation to describe image texture features by calculating the distribution of pixel pairs with a certain spatial position. Feature extraction methods based on transform domains such as Gabor transform [14] and wavelet transform [15] extract multi-scale and multi-directional features to capture complex local structural information of images. However, traditional algorithms often involve high manual feature extraction and classifier construction costs. Moreover, traditional algorithms usually lack learning capabilities and lack generalization capabilities when faced with complex tasks. This makes it difficult for features to capture implicit patterns and complex relationships between different terrains, which may limit model performance.

How to effectively and automatically extract features with learning capabilities is a hot topic in current research. With the development of artificial intelligence, the successful application of artificial intelligence covers almost all aspects of Earth-observation missions [16]. Artificial intelligence is also of great use in SAR data [17]. Deep learning theory [18] has been widely used in SAR image interpretation tasks due to its advantage of automatically learning high-level semantic features in the image. Su et al. [19] explored the performance of deep learning methods in SAR image interpretation. Wang et al. [20] designed separated convolutional streams to combine the intensity and gradient amplitude features of SAR images. Geng et al. [21] proposed a deep supervised and contractive neural network (DSCNN) for SAR image classification. Atteia et al. [22] proposed a method that integrates the power of autoencoder deep neural networks in mapping input features into representative latent-space features with the feature selection power of the principal component analysis (PCA) algorithm; the findings of this study revealed the superiority of the autoencoder deep learning network in generating latent features. Yue et al. [23] proposed a novel semi-supervised CNN method that can improve the reliability of unlabeled samples. However, some problems exist with using deep learning methods to extract features from SAR images. SAR images contain a variety of ground features. Unlike optical images, the annotation of SAR images often requires the experience of geoscience experts. It requires a lot of manpower, financial resources, and time to annotate SAR images. Therefore, collecting high-quality labeled SAR image data is difficult. However, the deep learning method is data-driven, and the amount of labeled data is positively correlated with the network's performance.

To solve this problem, a new learning paradigm, self-supervised learning (SSL) [24], is proposed. Self-supervised learning methods use a large number of unlabeled data to pre-train a general model and then fine-tune it on downstream tasks using very few labeled data. This method can alleviate the defect of insufficient labeled data and make full use of the information in the data itself. Many scholars have investigated the application of self-supervised learning in the field of computer vision. The development and challenges of self-supervised learning are introduced in [25]. Jing et al. [26] summarized the common deep neural network architectures used for self-supervised learning. To bridge the gap between the progress of SSFL in computer vision and remote sensing, Wang et al. [27] summarized some representative methods of SSFL and analyzed their application in remote sensing tasks. According to the type of supervision acquired, the pretext tasks in self-supervised representation learning methods can be divided into different categories. Tao et al. [28] explored the performance of remote sensing image scene classification under different self-supervised learning signals and proved that self-supervised learning can learn useful features from many unlabeled remote sensing images. Generative-based pretext tasks train the model to reconstruct the original input from a partially corrupted one for feature learning. Considering the differences between remote sensing images and natural images, Sun et al. [29] proposed a basic remote sensing model framework called RingMo and used the strategy of PIMask to reserve some pixels in the masking block randomly. Contrast-based pretext tasks bring different augmented views (positive sample pairs) of the same image closer and separate views (negative sample pairs) of other images. Jung et al. [30] proposed contrastive self-supervised learning of remote sensing smoothed representations based on the SimCLR framework, using multiple neighboring images as positive samples.

Predictive-based pretext tasks focus on learning semantic contextual features. Ji et al. [31] combined rotation prediction and contrastive learning to achieve few-shot scene classification for optical remote sensing images and introduced adversarial model perturbations to enhance generalization. The jigsaw puzzle-based learning signal is a type of predictive learning signal. Jigsaw puzzle learning [32] has a history of hundreds of years since it was proposed. Jigsaw puzzles are associated with learning, and people can help develop intelligence by completing them. It has been shown that jigsaw puzzles can be used to assess visuosp atial processing abilities. Doersch et al. [33] randomly extracted pairs of image patches from an image and predicted the position of the second image patch relative to the first image patch, demonstrating that visual similarity between images can be captured using features learned from the context within images. Noroozi et al. [34] applied the idea of jigsaw puzzles to natural images, where the learned features capture semantic information. Du et al. [35] fine-grained visual classification by progressive multi-granularity training of jigsaw puzzle pieces. Li et al. [36] proposed combining jigsaw puzzles with GAN to form JigsawGAN, a combination of generative and predictive learning signals. By solving the jigsaw puzzle problem, the network can maximize the information in the input data to obtain the characteristics of the image itself.

SAR image annotation is a difficult task, and existing terrain classification methods have certain limitations for SAR images. Also, extracting complex features in SAR images effectively is not easy. Aiming to mitigate these problems, this paper proposes a jigsaw puzzle self-supervised learning framework (JPSSL) for the SAR image terrain classification task. The framework mainly includes the jigsaw puzzle pretext task and the downstream task of SAR image terrain classification. This pretext task can develop the visual–spatial representation of the convolutional neural network context object and mine the representative characteristics of the unlabeled data as supervisory information. Downstream tasks achieve excellent experimental results using only a small number of labeled data.

The main contributions of this paper are as follows:

1. Considering that SAR images contain a variety of terrain types, it is not easy to extract effective features of different terrains from SAR images. A jigsaw puzzle pretext task for SAR images is designed. This task can be learned from the image itself through a large amount of unlabeled data, and the features extracted by the network are more discriminative. Models learned through this task can learn rich data representations that have strong generalization capabilities.
2. A jigsaw puzzle self-supervised learning framework (JPSSL) for the SAR image terrain classification task is proposed. This framework has a low dependence on data. With a few negligible-cost patch-level data, JPSSL can automatically capture image feature representation in the pretext task and effectively transfer it to the downstream task, achieving superior performance compared to supervised methods under the same conditions in terrain classification.
3. The proposed framework in this paper can perform terrain classification on SAR images of different granularities and has achieved excellent experimental results on SAR images of different resolutions and scenes.

The structure of the remainder of this paper is as follows. Section 2 details the self-supervised learning algorithm proposed in this paper. Section 3 details dataset information, experimental settings, experimental results, and analysis. Finally, the whole paper is concluded in Section 4.

2. Method

This paper proposes a SAR image terrain classification method based on self-supervised learning for jigsaw puzzles. A complete framework is shown in Figure 1, in which JPSSL consists of two stages: the pretext jigsaw puzzle task and the downstream terrain classification task. Figure 1a shows the structure diagram of the pretext task. Shuffled image patches are input to predict which permutation is used to shuffle them. The Alexnet network is used in the pretext task. Figure 1b shows the downstream task training part. The pre-trained encoder, consistent with the upstream task, is transferred to the downstream task and connected to the classifier. Downstream task fine-tuning only uses a small amount of labeled data. Figure 1c shows the downstream task testing part, including the SAR image terrain classification and FC-CRF post-processing. Algorithm 1 shows the pseudocode of JPSSL.

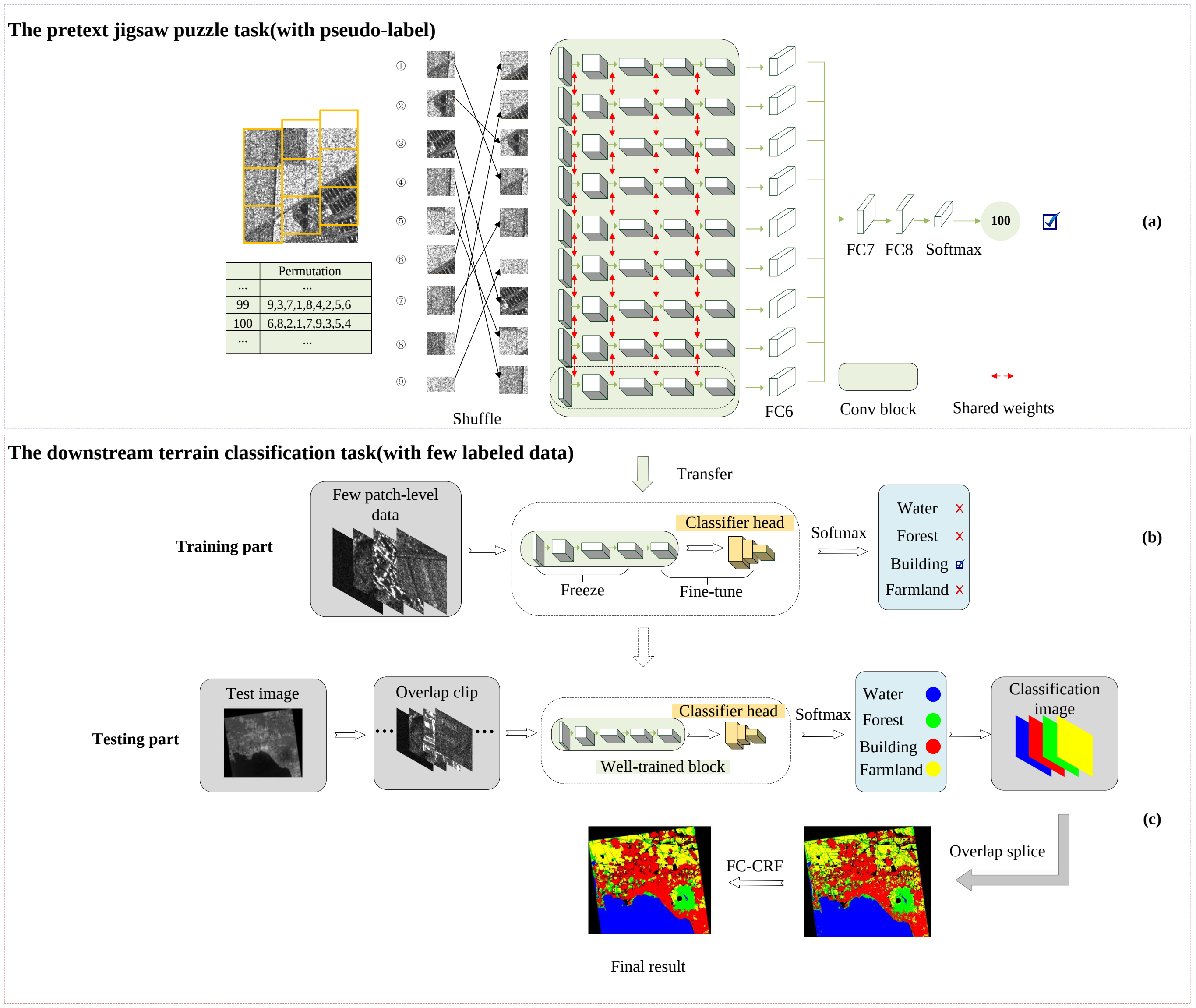


Figure 1. JPSSL framework. JPSSL consists of the pretext jigsaw puzzle task and the downstream terrain classification task. In the pretext jigsaw puzzle task, shuffled image patches are input to predict which permutation is used to shuffle them. The downstream terrain classification task includes the SAR image terrain classification training and testing part.

Section 2.1 describes how to acquire the data and detail the jigsaw puzzle pretext task designed for SAR images. Section 2.2 introduces the downstream task of terrain classification for SAR images and fully connected CRFs (FC-CRFs) for image processing.

2.1. Pretext Task

2.1.1. Data Collection

This paper designs a low-cost data acquisition method to obtain patch-level data using only a small amount of prior knowledge. For large-scene high-resolution SAR images, the natural area covers a large area. There are some large-scale aggregation areas in these natural areas, such as farmland, forest, water, and buildings. After comparing the SAR image with the corresponding optical image, patch-level sampling of different categories in the large-scale aggregation area can be obtained as the experimental data. Figure 2 shows the SAR image and its corresponding optical image. The square area in the figure represents the large-scale aggregation area found in the SAR image.

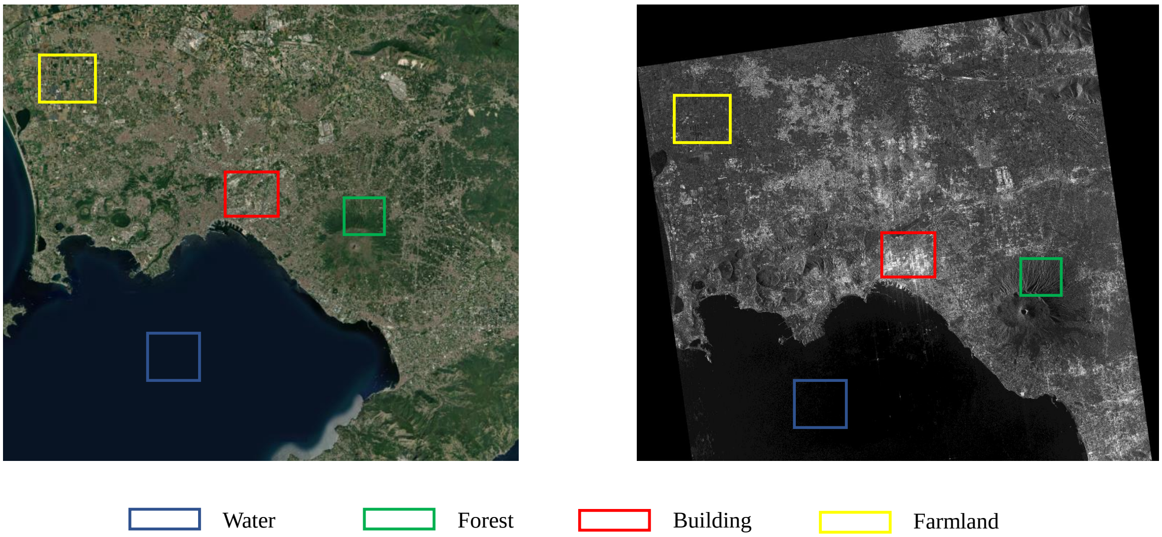
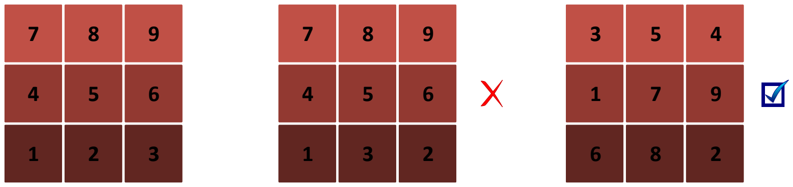


Figure 2. Comparing SAR images and optical images to obtain large-scale accumulation areas. The square area in the figure represents the large-scale aggregation area found in the SAR image.

2.1.2. Pseudo-Label Acquisition

The core of self-supervised learning is pseudo labels, which help the model learn the hidden information in unlabeled data. The jigsaw puzzle permutation set is investigated to obtain pseudo labels.

If the image for the jigsaw puzzle task is divided into 3 × 3 tiles and shuffled, there are 9! = 362,880 species of possible permutations. Different permutations are used as pseudo labels for the pretext task. However, many shuffled permutations are similar to the original permutation, as shown in Figure 3b. These similar permutations only change a small number of image block positions compared with the original permutation, and effective feature representations cannot be learned from these permutations. Therefore, it is necessary to choose a permutation method that is greatly different from the original permutation. The Hamming distance [37] is used to obtain the pseudo-label. The Hamming distance is used in data transmission error-control coding, which indicates the number of different characters in the corresponding positions of two strings of the same length. XOR operation is performed on the two strings and counts the number of 1s as the Hamming distance. The Hamming distance can be divided into maximum Hamming distance, minimum Hamming distance, and average Hamming distance. To select the permutation that is very different from the original permutation, which is shown in Figure 3c, we use the maximum Hamming distance to select permutations. The permutations selected in this way differ significantly from the original permutation. Multiple permutations determined by the maximum Hamming distance are combined to form the permutation set of the jigsaw puzzle task, i.e., the set of pseudo labels of the pretext task. The specific method is to calculate the Hamming distance between the shuffled and original permutations. These permutations are sorted from large to small, and the largest top N permutations are selected to form the permutation set for the pretext task.



(a) (b) (c)

Figure 3. Permutation selection diagram. (**a**) represents the original permutation; (**b**) represents the permutation that only changes the position of few image blocks; (**c**) represents the proposed permutation selection method.

2.1.3. Pretext Task Process

The pretext task based on jigsaw puzzle learning is a predictive–discriminative self-supervised learning task. The network inputs shuffled image blocks to determine the correct permutation and learn effective feature representations.

The network architecture of the jigsaw puzzle pretext task is shown in Figure 1a. The image is divided into nine image blocks according to the oblique clipping method, as shown on the left side of Figure 1a, where the size of each image block is the same. Since the areas in SAR images are very complex, some SAR images contain homogeneous content, such as water and farmland. The images of these areas have no outstanding features. After flatly clipping these nine image blocks, it is challenging to learn the correct permutation through the network, which is not conducive to learning effective feature representation. The nine image blocks clipped using the flat clipping method in some areas of the SAR image are very similar, and it is difficult for the network to determine the differences between them, so it is difficult to identify the correct permutation. Using the flat clipping method significantly increases the difficulty of the pretext task, meaning the pretext task cannot be trained normally. It is hard to train the network using the flat clipping method through experiments. Therefore, we used the oblique clipping method to solve the pretext task effectively, which is equivalent to a shortcut. Specifically, each column was raised to a certain height when clipped, and the pixel values beyond the image were set to 0. Specific differences were observed in the image blocks after clipping, which helped solve the pretext task.

The initial permutation order of the nine image blocks was set to [1, 2, 3, 4, 5, 6, 7, 8, 9], and normalization was performed on the nine image blocks. The permutation set of the pretext task was obtained through the pseudo-label acquisition method designed in Section 2.1.2. When the pretext task experiment was performed, a permutation was randomly selected in the permutation set, such as the 100th permutation [6, 8, 2, 1, 7, 9, 3, 5, 4]. These nine image blocks were sent into the network in order [6, 8, 2, 1, 7, 9, 3, 5, 4] to predict the index of the selected permutation. The pretext task network structure is a modification of AlexNet, in which the number of convolutional kernels in each convolutional layer is fewer. Before the first fully connected layer (fc6), each image block is transmitted in the network with shared weights. They are integrated after the first fully connected layer (fc6) to form the input to fc7. Finally, the features are sent to a softmax classifier to obtain the final output value.

During training, the indices must be assigned to the pre-defined permutation set, and the network returns a vector containing the probability value of each index. The final output of the network can be viewed as a conditional probability of the spatial permutation of objects.

$$p(M \mid C_1, C_2, \ldots, C_N) = p(M \mid F_1, F_2, \ldots, F_N) \prod_{i=1}^{N} p(F_i \mid C_i). \tag{1}$$

M represents the selected permutation, C_i represents the ith image block, and F_i represents the middle feature representation. Our goal was to enable features F_i to identify semantic properties of relative positions between image blocks.

Suppose only one jigsaw puzzle task is generated for a SAR image. In that case, it is possible that the network only learns information about the absolute position and not semantically relevant information. When multiple jigsaw puzzles need to be generated for one image, with M as the location list collection $M = (L_1, L_2, \ldots, L_N)$, then $p(M \mid F_1, F_2, \ldots, F_N)$ can be written as follows:

$$p(L_1, L_2, \ldots, L_N \mid F_1, F_2, \ldots, F_N) = \prod_{i=1}^{N} p(L_i \mid F_i); \tag{2}$$

therefore,

$$p(M|C_1, C_2, \ldots, C_N) = \textstyle\prod_{i=1}^{N} p(L_i|F_i) \prod_{i=1}^{N} p(F_i|C_i) = \prod_{i=1}^{N} p(L_i|C_i). \tag{3}$$

The position L_i of each image block is completely determined by the corresponding feature F_i.

The cross-entropy loss function [38] is used in the network. Cross-entropy is a concept in information theory. Given two probability distributions, p and q, the cross-entropy of p represented by q is as follows:

$$H(p, q) = -\sum_{x} p(x) log q(x). \tag{4}$$

The cross-entropy loss function used in the pretext task is defined as follows:

$$L_{upstream} = -\frac{1}{S} \sum_{j} \sum_{a=1}^{Y} y_{ja} log(p_{ja}), \tag{5}$$

where S is the selected sample size, Y is the number of categories, y_{ja} represents the true distribution reflected by the training set, and p_{ja} represents the predicted probability that the observed sample j belongs to category a.

2.2. Downstream Terrain Classification Task

2.2.1. Task Process

The training stage of the downstream task consists of an encoder and a classification head. The encoder part is identical to the jigsaw puzzle pretext task, so the encoder part can directly use the weights trained on the jigsaw puzzle pretext task. The weights of the convolutional layers are transferred, and the weights before the last convolutional layer are frozen. This means only the last convolutional layer and the fully connected layer are trained in the downstream task, as shown in Figure 1b. A small amount of patch-level image data is used for training. The output of the training phase is the probability of the terrain category of the input image patches, and the model is fine-tuned using a multi-class cross-entropy classification loss. The multi-class cross-entropy classification loss function is defined as follows:

$$L_{downstream} = \mathbb{E}[-\log P(y_i^{'} / y_i)], \tag{6}$$

where $\mathbb{E}$ represents the mathematical expectation, which is the superposition of multiple functions. y_i represents the real terrain category of patch-level data, and $y_i^{'}$ is the corresponding terrain category label.

Unlabeled large-scene high-resolution SAR images need to be preprocessed before the testing phase. The SAR image needs to be cropped into image patches of the same size as the downstream task training data. The overlap cropping method is used to crop image patches so each image patch has a central region. The test image patches are fed into the network to obtain the predicted terrain category, and all pixels in the central region are predicted as the pixel values corresponding to the terrain category. After all the image data have been tested in the test phase, the central region of the image data is spliced to obtain the final result of terrain classification. The overlapping range in the clipping process affects the size of the central region, and this hyperparameter experiment is carried out in the analysis of experimental results in Section 4.

2.2.2. Fully Connected CRFs

This patch-level segmentation method produces errors at the edges of different regions and produces a jagged structure. Therefore, the post-processing method is used to improve classification performance and obtain high-quality results. FC-CRFs are obtained by improving CRFs to combine the relationship between all pixels in the original image to process the classification results obtained by deep learning. This method can optimize the rough and uncertain marks in the classification image and correct the fine misclassified regions to obtain more detailed segmentation boundaries. The efficient probability approximation algorithm [39] is adopted to implement FC-CRFs. A random field $X = \{X_1, X_2, \ldots, X_N\}$ is created for N pixels of an image, where X_j denotes the label assigned to pixel j. Also, the RGB vector of the pixel is defined to form a random field $I = \{I_1, I_2, \ldots, I_N\}$, where I_j represents the RGB vector of pixel j. FC-CRFs follow the Gibbs distribution and can be written as follows:

$$P(x = X \mid I) = \frac{1}{Z(I)} e^{-E(X|I)}. \tag{7}$$

$E(X \mid I)$ is an energy function composed of a unary potential function and a binary potential function.

$$E(X \mid I) = \textstyle\sum_i \psi_u(x_i) + \sum_{i,j} \psi_b(x_i, y_j). \tag{8}$$

The first term, $\psi_u(x_i)$, is the unary potential function produced by the obtained prediction result map, representing the probability distribution of the label assigned to the pixel. The second term, $\psi_b(x_i, y_i)$, is a binary potential function that constrains the relationship between pixels and is defined as follows:

$$\psi_b(x_i, y_j) = u(x_i, y_j) \sum \omega K_G(f_i, f_j), \tag{9}$$

$$K_G(f_i, f_j) = W_1 e^{-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}} + W_2 e^{-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}}. \tag{10}$$

$u(x_i, y_j)$ is a label-compatible item, and energy can only be transmitted under the condition of the same label. $K_G(f_i, f_j)$ is the characteristic function, where p_i and p_j are position vectors, I_i and I_j are color vectors, and θ_α, θ_β and θ_γ are hyperparameters.

This FC-CRF post-processing method can eliminate misclassified points on the edges of different regions and inside and remove the jagged structure that appears in the image. It can alleviate some unavoidable errors when using the patch-level method and obtain high-quality results.

Algorithm 1 The training process of JPSSL.

```
Require:
    Input image $i$, permutation set $p$, training set $t$, and category label $l$.
    The jigsaw puzzle encoder $J$, the jigsaw puzzle classifier $C$, and the classification model $P$.
Ensure:
    Classification model $P$.
The pretext task:
for epochs do
    Randomly choose a permutation $p1$ from the permutation set $p$.
    Divide the input image $i$ into 9 image blocks according to $p1$.
    Send the image blocks into the network to predict the index of the input permutation
    through Formula (3).
    Calculate the loss of the pretext task by Formula (5).
    Update the parameters of $J$, and $C$.
end for
The downstream task:
Select a small number of training data with category label $l$ from the training set $t$.
Transfer the pre-trained encoder $J$ to the classification model $P$.
for epochs do
    Obtain predicted label distribution $l(t)$, and calculate classification loss by
    Formula (6).
    Fine-tuning parameters of the classification model $P$.
end for
Classify, colorize, and stitch images using classification models $P$.
Post-processing with FC-CRF.
```

3. Experiments

In this section, the effectiveness of the present method is demonstrated on a 25-class SAR scene dataset and three large-scene high-resolution SAR images. More precisely, in Section 3.1, the 25-class SAR scene dataset and three large-scene high-resolution SAR images are introduced in detail. Then, the evaluation metrics for pretext and downstream tasks are proposed. In Sections 3.2 and 3.3, the present method is validated on the 25-class SAR scene dataset and the large-scene high-resolution SAR images. All experiments are carried out using Ubuntu 18.04 using a Pytorch 1.8.0 environment with Intel Xeon CPUs and NVIDIA RTX 2080Ti. Intel Xeon CPUs are central processing units produced by Intel Corporation of the United States. NVIDIA RTX 2080Ti is a graphics card produced by Intel Corporation of the United States.

3.1. Experimental Data and Evaluation Indicators

3.1.1. The 25-Class SAR Scene Data

The scene dataset used was constructed by applying a regular grid to several high-resolution SAR images acquired by the TerraSAR-X satellite in HH polarization with spotlight mode. These images were taken over Rosenheim, Toronto, Java, Colorado, Beijing, and Hong Kong airports. The scene dataset contains rich scenes, including water, farmland, forests, residential areas of different densities, etc. These 25 types of scene data are a wasteland, airport runway, three types of water, agriculture, four types of buildings, sparse residences, four types of dense residential buildings with different densities, skyscrapers, two types of Rivers, two types of roads, farmland, forest, grass houses, train tracks, and vegetated farmland mixtures. Each type of scene data contains 400 data, and the size of each image is 200 × 200.

3.1.2. Large-Scene High-Resolution SAR Image Data

Three large-scale high-resolution SAR images from different regions acquired by different satellites were used to conduct SAR image terrain classification experiments. The three SAR images are the SAR images of the Jiujiang area in China, the Napoli area in Italy, and the PoDelta area in Italy. The SAR image of the JiuJiang area in China was taken by the Gaofen-3 satellite. The image size is 8000 × 8000, the imaging mode is DV polarization, and the ground resolution is 3m. The SAR images of the Napoli area and the PoDelta area in Italy were taken by Cosmo-SkyMed satellites. The image sizes are 16,000 × 18,332 and 16,716 × 18,308, respectively; the imaging mode is HH polarization, and the ground resolution is 2.5 m. The original SAR images were stored in 16-bit data, while a standard CNN was used to process the 8-bit image data, so the SAR images needed to be pre-processed. The image pixel values were normalized to the range of 0–255 by truncated linear stretching [40] for all SAR images. Pixels in the above three images were sorted into five terrain categories: water, forest, buildings, farmland, and unknown class. No predictions were made for unknown class regions during training and testing.

3.1.3. Evaluation Indicators

There are certain differences in the evaluation indicators of the pretext task and the downstream task under the two different experiments.

1. Pretext Task
   The evaluation indicators of the pretext tasks under two different experiments are consistent. First, a permutation is randomly selected from the pre-defined permutation set, and image blocks are shuffled according to this permutation. The shuffled image blocks are input into the network to obtain a probability vector. The index value of the largest value in the probability vector is the predicted result, and the accuracy is calculated using the real index value of the permutation. The jigsaw puzzle pretext task can be regarded as a multi-classification problem, and the results of the pretext task are displayed in the form of classification accuracy. The $UP_{accuracy}$ is used to represent the accuracy rate of the pretext task, where TP represents the number of positive classes predicted as positive classes, TN represents the number of negative classes predicted as negative classes, and N represents the total amount of data.
   $UP_{accuracy}$: Pretext task accuracy.

$$UP_{accuracy} = (TP + TN)/N. \tag{11}$$

2. Downstream Task

There are some differences in the evaluation indicators of the downstream tasks of the two experiments.

The downstream task of 25-class SAR scene data classification is a scene classification task, and the results of the final downstream task are presented as the accuracy of a multi-classification problem. The accuracy rate calculation formula is equivalent to Formula (11).

The downstream task of large-scene high-resolution SAR image terrain classification can be regarded as a segmentation task. The pixel accuracy (PA), Kappa coefficient, mean intersection ratio (MIoU), and frequency-weighted intersection ratio (FWIoU) are used to measure the overall classification performance. The check precision rate (CPA), Recall, and the F1 score are used to measure the classification performance of individual categories. The relevant evaluation indicators and their definitions are as follows.

There are $k + 1$ classes (including the unknown class), and $p_{i,j}$ represents the number of pixels that belong to class i but are predicted to class j. The unknown class is not involved in the calculation of the evaluation metric.

CPA: Proportion of correct predictions that are positive to all positive predictions.

$$CPA = \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + p_{ii}}. \tag{12}$$

Recall: Proportion of correct predictions that are positive to all positive data.

$$Recall = \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ji} + p_{ii}}. \tag{13}$$

F1Score: The harmonic mean of recall and CPA.

$$F = \frac{2 \times CPA \times Recall}{CPA + Recall}. \tag{14}$$

PA: Proportion of correctly labeled pixels to total pixels.

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ii}}. \tag{15}$$

Kappa: Used for consistency testing to penalize model bias to obtain a more unbiased model.

$$Kappa = \frac{p_o - p_e}{1 - p_e} p_e = \frac{a_1 \times b_1 + a_2 \times b_2 + \cdots + a_c \times b_c}{n \times n}. \tag{16}$$

p_o represents the proportion of correctly classified data to the total data, which is equivalent to PA; $a_1, a_2, \cdots, a_c$ represent the amount of real data for each class; $b_1, b_2, \cdots, b_c$ represent the amount of predicted data for each class; c represents the number of categories; n represents the total amount of data.

MIoU: Calculates the ratio of the intersection and union of sets of true and predicted values.

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{17}$$

FWIoU: An improvement of MIoU, which sets weights according to the frequency of occurrence of categories.

$$FWIoU = \frac{1}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \times \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}. \tag{18}$$

3.2. 25-Class SAR Scene Data Classification

In this section, the classification experiment of 25-class SAR scene data is detailed. First, the experimental settings for pretext and downstream tasks are introduced. Then, the effects of permutation set size, frozen layers, and the amount of labeled training data are explored. Under the optimal parameter setting, the effectiveness of the proposed method is confirmed compared with the comparative method. Finally, the extracted features are qualitatively evaluated by image retrieval.

3.2.1. Experimental Settings for Pretext and Downstream Tasks

The pretext task uses 25-class scene data; each type of data contains 400 training data, and the size of each image is 200 × 200. The pretext task training and verification sets are divided according to a ratio of 7:3. The images are resized to the size of 255 × 255 and divided into nine image blocks according to Section 2.1.3. The batch size of the training set and the verification set are set to 64 and 32, respectively. The initial learning rate is set to 0.01, and the training epoch is set to 300. The learning rate iteration criterion is that the learning rate becomes 1/2 of the original for every 50 epochs. The optimizer is the SGD optimizer, the momentum is set to 0.9, and the weight decay is set to 0.0001.

The model trained by the jigsaw puzzle pretext task is the pre-trained model for the downstream task. Only the probability of the terrain class of the input image patch is output in the test phase of the downstream task. The downstream task sets the initial learning rate to 0.01 and the training epoch to 100. The learning rate iteration criterion is that the learning rate becomes 1/2 of the original for every 50 epochs. The optimizer is the SGD optimizer, the momentum is set to 0.9, and the weight decay is set to 0.0001.

3.2.2. Analysis of Influencing Factors

1. Permutation Set Size
As the method of pseudo-label acquisition, the permutation set's size affects both the pretext and downstream tasks. To explore the effect of the size of the permutation set on the 25-class SAR image scene classification task, experiments on the permutation set size parameter are conducted. The training data for the downstream task are ten randomly selected data for each category, and the weights of all convolutional layers are frozen. The accuracy rates for the pretext and downstream tasks with different permutation set sizes are shown in Table 1. 0.635 is the highest accuracy for the downstream task, shown in bold in Table 1.

Table 1. Experiments with different permutation set sizes for the pretext task and the downstream task.

Permutation Set Size	50	80	100	120	200	300	500
Pretext task	0.81	0.76	0.72	0.71	0.61	0.52	0.45
Downstream task	0.51	**0.635**	0.63	0.57	0.59	0.43	0.37

It can be seen from the experimental results in Table 1 that as the permutation set size increases, the difficulty of the jigsaw puzzle pretext task increases, which leads to a decrease in the accuracy of the pretext task. The main measure of the performance of the self-supervised task is how well the downstream task performs, and our ultimate goal is to determine the parameters that perform best in the downstream task. As can be seen from the results of the downstream task in Table 1, although the pretext task accuracy is the highest when the permutation set size is set to 50, the accuracy of the downstream task is lower at this point. Better downstream task results are achieved for the permutation set size of 80.

2. Frozen Layers
The model obtained from the pretext task for the downstream task is the transfer learning process. In the experiments shown in Table 1, the transfer learning method is adopted to transfer the weights of all convolutional layers to the downstream task and freeze them, which means the gradient is not updated during the training process. Table 2 shows the downstream task accuracy using different freezing layer methods when the permutation set size is 80. The training data used are consistent with those in Table 1.

Table 2. Experiments with different frozen convolutional layer parameter methods.

How to Freeze Parameters	Permutation Set Size	Downstream Task Accuracy
Method I	80	0.635
Method II	80	**0.726**
Method III	80	0.67
Method IV	80	0.65

In Table 2, Method I freezes all convolutional layer weights, Method II freezes all weights before the last convolutional layer, Method III freezes all weights before the last two convolutional layers, and Method IV freezes all weights before the last three convolutional layers. Bold in Table 2 indicates the highest accuracy. It can be concluded from Table 2 that the downstream task achieves the highest accuracy when all weights before the last convolutional layer are frozen. In the subsequent experiments, the permutation set size was set to 80, and the parameters of all weights before the last convolutional layer were frozen during transfer learning.

3. Amount of Labeled Training Data

With sufficient labeled data, the supervised learning approach is superior to the self-supervised learning approach. Self-supervised learning aims to reduce the reliance on labeled data and achieve better results with a small amount of labeled data. The impact of the amount of labeled training data is explored for the downstream task. The rest of the data, except the training data, are used as the validation set and compared with the supervised method that does not use pre-trained weights.

It can be seen from Table 3 that when the amount of training data for each class is small, the present method can achieve better results than the supervised method. The difference between the two methods gradually decreases as the amount of training data increases. The self-supervised learning method does not show advantages when the amount of training data is further increased. It can be seen from the table that when only 10 training data are selected for each class, the improvement effect of the present method is the largest.

Table 3. Experiments with different amounts of labeled training data for the downstream task.

	Downstream Task Accuracy				
Amount of Data per Class	**3**	**10**	**25**	**35**	**50**
JPSSL(no pretraining)	0.304	0.479	0.715	0.82	0.9
JPSSL	0.501	0.731	0.854	0.874	0.9

3.2.3. Image Retrieval

The extracted features are qualitatively evaluated. Images are manually selected in the 25-class dataset, and the output features from convolutional layers determine the *k*th nearest neighbor of the selected images. The extracted features are qualitatively evaluated using the Euclidean distance. The Euclidean distance is defined as follows:

$$\mathcal{L}_2(x_i, x_j) = \sqrt{\sum_{l=1}^{n}(x_i^l - x_j^l)^2}. \quad (19)$$

The feature space X is an n-dimensional real vector space. Both x_i and x_j belong to X. $x_i = (x_i^1, x_i^2, \ldots, x_i^n)^T$, $x_j = (x_j^1, x_j^2, \ldots, x_j^n)^T$.

The output features of the convolutional layer are used to determine the top three neighbor images through the input image, as shown in Figure 4.

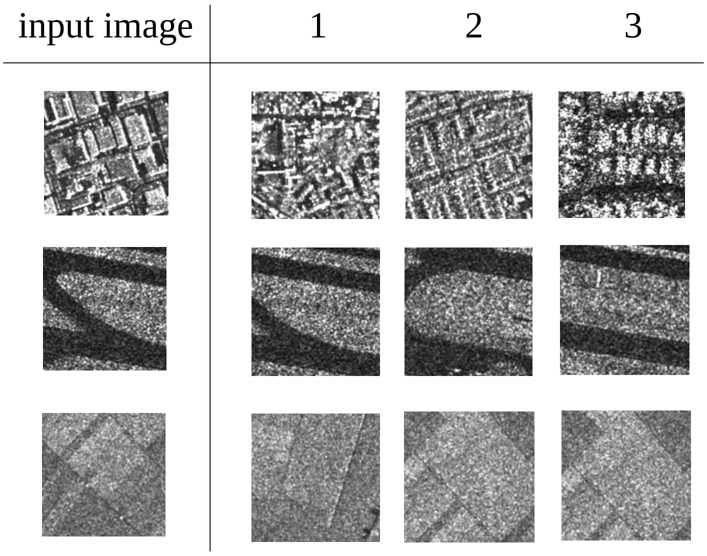


Figure 4. Image retrieval graph for different images. 1, 2, and 3 represent the input image of the nearest neighbor image, the second nearest neighbor image, and the third nearest neighbor image in turn.

It can be seen from Figure 4 that the images retrieved are of the same class as the query image, which means that the features extracted from the convolutional layer are sensitive to similar objects. The effectiveness and discriminativeness of the extracted features can be proven.

3.3. Terrain Classification of High-Resolution Large-Scene SAR Images

In this subsection, the terrain classification experiments of large-scene high-resolution SAR images are introduced in detail. First, the settings of pretext tasks and downstream tasks are introduced. Then the effects of the cut image block size, the permutation set size, the normalization method, the number of labeled training data, the central area, and the selection of labeled data are explored. Section 3.3.2 elaborates the influence of the above factors in the Jiujiang image. The hyperparameters of the Napoli and PoDelta images are determined identically in Section 3.3.3. Section 3.3.3 analyzes the experimental results of three large-scene high-resolution SAR images in detail.

3.3.1. Experimental Settings for Pretext and Downstream Tasks

The SAR image terrain classification data are obtained by the data acquisition method in Section 2.1.1. The three large-scene high-resolution SAR images used are all composed of five categories: water, forest, buildings, farmland, and unknown class. Limited by label data, only four areas are classified: water, forest, building, and farmland. In addition to these areas, there are also some small areas in the three SAR images. These small areas are difficult to distinguish between categories when labeling, so they are marked as unknown classes. We do not consider unknown class areas in the process of terrain classification. The image patches of the same size are cut out for each category. The image used for the jigsaw puzzle task is cut into nine blocks according to the method on the left side of Figure 1, and the pixel value beyond the image is set to 0. The image blocks are sent to the network according to the selected permutation to predict which permutation is selected. The batch size of the training and verification sets are 64 and 32, respectively. The initial learning rate is set to 0.01, and the training epoch is set to 300. The learning rate iteration criterion is that the learning rate becomes 1/2 of the original for every 50 epochs. The optimizer is the SGD optimizer, the momentum is set to 0.9, and the weight decay is set to 0.0001.

In the downstream task, the model obtained by the pretext task is fine-tuned for the downstream task. Twenty data of each terrain category are randomly selected as training data from the patch-level data generated in Section 2.1.1. The training process of the downstream task is shown in Figure 1. The training batch size is set to two due to the small number of training data, the initial learning rate is set to 0.01, and the training epoch is set to 100. The learning rate iteration criterion is that the learning rate becomes 1/2 of the original for every 50 epochs. The optimizer is the SGD optimizer, the momentum is set to 0.9, and the weight decay is set to 0.0001. The testing process of the downstream task is shown in Figure 1. The high-resolution SAR images are cropped into patches of the same size during the training process. Then, the test data are inputted into the network to predict the terrain category. To obtain the pixel-level classification result, the pixel corresponding to the prediction output category is used as the pixel in the central region of the test image patches. After all the test data are tested, the central predicted regions of all the test image patches are stitched together and processed by FC-CRF to obtain the terrain classification results for the complete SAR image.

3.3.2. Analysis of Influencing Factors

1. Cut Image Patch Size
When the low-cost data acquisition method is used, patch-level data need to be selected in each type of large-scale aggregation area in SAR images. The size of the cut image patches affects the results of the pretext and downstream tasks. The size of the JiuJiang SAR image is 8000 × 8000, and the image resolution is 3 m. Image patches that are cut too small cannot effectively provide the context information of

the image, and image patches cut too large lead to the degradation of classification performance. Experiments on the size of the cut image patches are conducted.

Due to the relatively small size of the JiuJiang SAR image, five different cut image patch sizes are used for our experiments. Bold in Table 4 indicates the highest values of different indicators. As can be seen from Table 4, the accuracy of the pretext task is higher when the image patch sizes are 75 × 75, 120 × 120, and 150 × 150. Images can be divided into nine blocks without gaps in these three cases. Meanwhile, images cannot be evenly divided into nine blocks when the image patch sizes are 50 × 50, 150 × 150, and 100 × 100. The performance of self-supervised tasks mainly concerns the performance of downstream tasks. It can be seen from Table 4 that when the cut image patch size is 75 × 75, the PA of the model can reach 85.3%, the MIoU can reach 68.8%, and the overall classification performance is the best. Therefore, the image patch size is set to 75 × 75 for the JiuJiang SAR image.

Table 4. Experiments with different cut image patch sizes for the pretext task and the downstream task on JiuJiang data.

Image Patch Size	Pretext Task	Downstream Task			
	Accuracy	PA	Kappa	MIoU	FWIoU
50 × 50	0.758	0.835	0.754	0.630	0.764
75 × 75	0.88	**0.853**	**0.784**	**0.688**	**0.798**
100 × 100	0.68	0.782	0.689	0.616	0.715
120 × 120	0.913	0.724	0.616	0.567	0.651
150 × 150	0.838	0.754	0.653	0.578	0.698

2. Permutation Set Size and Normalization Method

As the pseudo-label acquisition method of this task, the size of the permutation set has an impact on SAR image terrain classification. The permutations of the pretext task according to the maximum Hamming distance proposed in Section 2.1.2 are selected and combined to form the permutation set. The cut image patches of 75 × 75 pixels are used to conduct experiments on the JiuJiang SAR image. For the training process of the downstream task, 20 labeled training data for each terrain category are selected. For the downstream task training process, the scene classification experiment is used. The performance of scene classification is positively correlated with the performance of final terrain classification and is relatively simpler. Table 5 shows the performance of the pretext and downstream tasks for different permutation set sizes.

Table 5. Experiments with different permutation set sizes for the pretext task and the scene classification task on JiuJiang data.

Permutation Set Size	50	80	100	125	150	175	200	300
Pretext task accuracy	0.935	0.885	0.855	0.844	0.767	0.709	0.68	0.59
Scene classification accuracy	0.948	0.945	0.949	0.944	0.948	0.95	0.944	0.936

It can be seen from Table 5 that as the size of the permutation set increases, the difficulty of the pretext task increases, and the accuracy of the pretext task decreases. However, it can be seen from Table 5 that the accuracy of the downstream task is almost constant, and a slight change in the accuracy of the scene classification task causes little change on the terrain classification task. The accuracy of downstream tasks tends to be consistent under different permutation set sizes through multiple experiments. It can be concluded that the permutation set size has little impact on the terrain classification result. Considering the calculation cost and the final performance, and to be consistent with Section 3.2.2, the permutation set size in the terrain classification experiment is set to 80.

To better solve the SAR image jigsaw puzzle problem, normalization methods for image blocks have been researched. Normalization can reduce internal covariance so the model can be trained effectively. The normalization methods for the nine image blocks used for the jigsaw puzzle task are explored in Table 6, and the results under different normalization methods for the pretext and downstream tasks are shown in Table 7. Bold in Table 7 indicates the highest accuracy for different tasks.

Table 6. Different normalization methods.

Mode	Normalization Methods
Mode I	Individual normalization for each image block
Mode II	All image blocks are normalized using a uniform mean standard deviation
Mode III	Each image block uses a uniform mean and each image blockâĂŹs standard deviation is handled separately
Mode IV	Each image block uses a uniform standard deviation and each image blockâĂŹs mean is handled separately

Table 7. Experiments with different normalization methods for the pretext task and the scene classification task.

Normalization Mode	Mode I	Mode II	Mode III	Mode IV
Pretext task accuracy	**0.937**	0.784	0.933	0.918
Scene classification accuracy	**0.948**	0.933	0.932	0.925

As seen in Table 7, the best results are obtained in both the pretext and downstream tasks when using mode I. Therefore, mode I is used for the jigsaw puzzle task. The same pattern applies to the other data as well.

3. Amount of Labeled Training Data
The amount of labeled training data affects the performance of the final terrain classification task. Self-supervised learning aims to reduce reliance on labeled data and achieve better results with a small amount of labeled data. The effect of the amount of labeled data is explored under the best hyperparameters of the above experiment. For the JiuJiang SAR image, the cut image patch with the size of 75 × 75 pixels is used, and the size of the central area selected when cropping the image from the SAR image is 25 × 25. Table 8 shows the results with and without the transfer pre-trained model.

Table 8. SAR image terrain classification under different amounts of labeled training data.

The Amount of Data per Class	JPSSL (No Pretraining)		JPSSL	
	PA	MIoU	PA	MIoU
10	0.745	0.533	0.827	0.638
20	0.767	0.582	**0.850**	**0.686**
30	0.822	0.641	0.824	0.664

The bold in Table 8 indicates the highest values of PA and MIoU. It can be seen from Table 8 that when 20 labeled training data are selected for each category, the performance of the model is the best, and the classification performance is improved to a certain extent relative to the supervised method. The same pattern applies to other data as well.

4. Central Prediction Area Size
The choice of the central prediction area size of the cut image patch impacts the model's classification accuracy. The larger the central area size, the coarser the classification, but the more efficient it is. The smaller the central area size, the finer the classification, but the less efficient it is. Therefore, choosing the appropriate size of the central prediction area is important. Based on the above experiments, experiments on different central prediction area sizes are carried out using JiuJiang data. Table 9

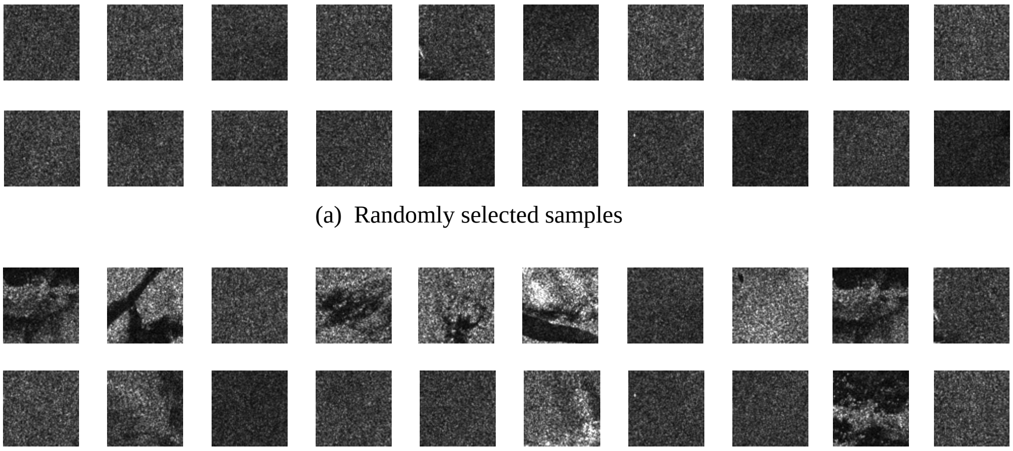
shows the classification performance of different central prediction area sizes. Bold in Table 9 indicates the highest values of different indicators.

Table 9. Experiments with different center prediction area sizes.

Central Prediction Area Size	PA	Kappa	MIoU	FWIoU
15 × 15	**0.870**	**0.808**	**0.705**	**0.812**
25 × 25	0.867	0.803	0.699	0.807
75 × 75	0.834	0.756	0.650	0.767

The experimental results are consistent with the theoretical analysis. When the size of the central prediction area is reduced from 25 × 25 to 15 × 15, the model's classification performance is improved by less than 0.1% but the training time and memory are significantly increased. Considering classification performance and efficiency issues, the central area is set to 25 × 25 pixels.

5. Selection of Labeled Data
When selecting a small amount of labeled data for the downstream task, images from different aggregation areas for each category should be selected evenly. If the selected data are not sufficiently representative, it leads to a decrease in the performance of the model. In the JiuJiang data set, the images of water in different large-scale aggregation areas are slightly different, so we experiment with the selection of labeled data. Figure 5a shows the randomly selected data, and Figure 5b shows the manually selected data. Bold in Table 10 indicates the highest values of different indicators.



(a) Randomly selected samples

(b) Manually selected samples

Figure 5. Randomly selected data and manually selected data.

It can be seen from Table 10 that various performance indicators are improved to a certain extent if manually selected data are used. The selection of representative data can improve the performance of the model. Considering the issue of model performance, more representative data for each category will be selected in subsequent SAR image experiments.

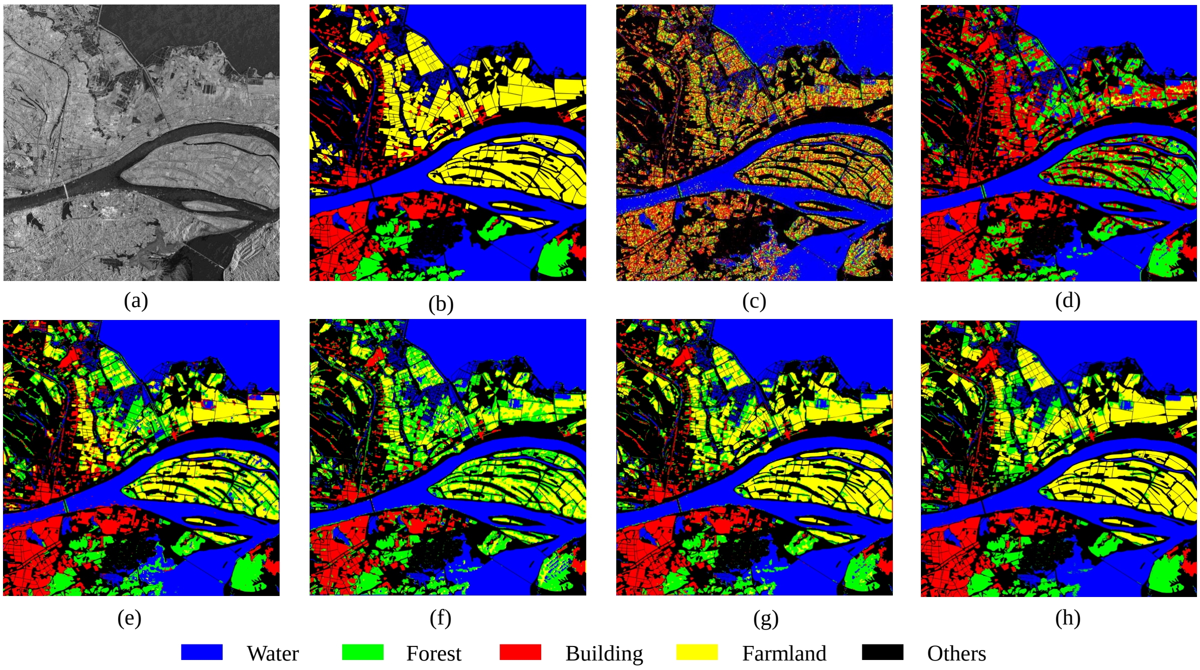
Table 10. Experiments to change training data.

	PA	Kappa	MIoU	FWIoU
randomly selected data	0.853	0.784	0.688	0.798
manually selected data	**0.867**	**0.803**	**0.699**	**0.807**

3.3.3. Large-Scene High-Resolution SAR Images Terrain Classification

By comparing the proposed method with several terrain classification methods on different data, the effectiveness of the proposed method is proven. The Deeplabv3+ method [41], Segformer method [42], SimCLR method [43], and the supervised method that does not use pre-trained weights are used in comparative experiments. Finally, the results of terrain classification combined with the JPSSL framework and FC-CRF are shown. The Deeplabv3+ method achieves advanced classification performance of images by extracting multi-scale features and gradually recovering spatial information. The supervised method does not use pre-trained weights as the baseline.

1. Jiujiang Data
   The experiments are conducted under the best settings explored in Section 3.3.2. Figure 6 shows the visualization results of terrain classification on JiuJiang data for the present method and the comparison methods, and Table 11 shows the evaluation indicators of terrain classification on JiuJiang data for the present method and the comparison methods. Bold in Table 11 indicates the highest values of different indicators. As the deeplabv3+ method requires a large amount of training data to achieve good results, overfitting is serious if a small amount of training data is used, which results in poor final classification performance. The indicators obtained using the JPSSL method improved to a certain extent compared with different comparative experiments. The forest and farmland categories of the JiuJiang image are very similar and cannot easily be distinguished from each other. Compared with the baseline, the present method improves the F1 score indicators of the forest class and farmland class by 17% and 23%, respectively. Comparison of Figure 6f,g shows that the present method can better distinguish forest and farmland. It can be seen from the table that the present method achieves the highest in all indicators. Compared with the baseline, there is a 10% improvement in PA, MIoU, and FWIoU, and a 14% improvement in Kappa indicators. After using FC-CRF post-processing, the overall metrics are improved by approximately 2%, and the classification effect of forest and farmland is improved. A comparison of Figure 6b,h shows the excellent result achieved by the present method.


(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 6. Visualization of terrain classification results with different methods on Jiujiang data. (**a**) SAR image. (**b**) Ground truth. (**c**) Deeplabv3+. (**d**) Segformer. (**e**) SimCLR. (**f**) JPSSL (no pre-training). (**g**) JPSSL (pre-training). (**h**) JPSSL (pre-training + FC-CRF).

Table 11. Terrain classification results with different methods on JiuJiang data.

Method	PA	Kappa	MIoU	FWIoU	F1score			
					Water	Forest	Building	Farmland
Deeplabv3+	0.610	0.432	0.352	0.501	0.884	0.187	0.411	0.403
Segformer	0.652	0.511	0.402	0.516	0.941	0.372	0.622	0.078
SimCLR	0.827	0.748	0.645	0.752	0.956	0.480	0.859	0.747
JPSSL (no pre-training)	0.767	0.663	0.582	0.702	0.963	0.338	0.867	0.603
JPSSL (pre-training)	0.867	0.803	0.699	0.807	0.964	0.503	**0.899**	0.834
JPSSL (pre-training + FC-CRF)	**0.884**	**0.827**	**0.725**	**0.829**	**0.973**	**0.567**	0.892	**0.858**

2. Napoli Data
The size of the SAR image in the Napoli area of Italy is 18,332 × 16,000, a large image with a higher resolution than the SAR image of JiuJiang. Due to the difference in image size, resolution, and imaging method, the hyperparameters suitable for the JiuJiang SAR image may not be suitable for the Napoli SAR image. The cut image patch size is one of the most significant points. Based on the hyperparameter experiment of the cut image patch size of the JiuJiang SAR image, the experiment is conducted on the cut image patch size of the Napoli image, and supplements are made accordingly. Table 12 represents the impact of different cut image patch sizes for the pretext task and the downstream terrain classification task. Bold in Table 12 indicates the highest values of different indicators.

Table 12. Experiments with different cut image patch sizes in Napoli, Italy, data.

Image Patch Size	Pretext Task Accuracy	Downstream Task			
		PA	Kappa	MIoU	FWIoU
50 × 50	0.728	0.788	0.710	0.618	0.674
75 × 75	0.856	0.786	0.707	0.620	0.675
100 × 100	0.656	0.784	0.709	0.631	0.679
120 × 120	**0.894**	**0.804**	**0.733**	**0.649**	**0.697**
150 × 150	0.805	0.787	0.711	0.630	0.675
200 × 200	0.740	0.789	0.712	0.625	0.673
255 × 255	0.890	0.768	0.682	0.596	0.644

It can be seen from Table 12 that when the cut image patch size is 120 × 120 pixels, both the pretext task and the downstream terrain classification task achieve the highest indicators. For the SAR image in Napoli, Italy, the cut image patch size is set to 120 × 120 pixels, and the rest of the parameters are universal and consistent with the corresponding parameters of the JiuJiang SAR image.

The experiments are conducted using the best parameter settings presented above. Figure 7 shows the visualization results of terrain classification on the Napoli data for the present method and the comparison methods, and Table 13 shows the evaluation indicators of terrain classification on Napoli data for the present method and the comparison methods. Bold in Table 13 indicates the highest values of different indicators. The Deeplabv3+ method is seriously overfitted when only 20 training data are used for each class. The indicators obtained using the JPSSL method improved to a certain extent compared with different comparative experiments. Compared with the baseline, the F1 score indicators of the forest and building categories increased by 20% and 10%, respectively, which shows the present method improves the discriminative performance of forest and building features. The present method achieves the highest indicators in terms of overall indicators. Compared with the baseline, PA shows approximately 6% improvement, and Kappa, MIoU, and FWIoU show roughly 8% improvement. After using FC-CRF post-processing, the overall metrics hare improved by approximately 1.5%, and the classification effect of forest and farmland improved overall. A comparison of Figure 7b,h shows the excellent result achieved by the present method.

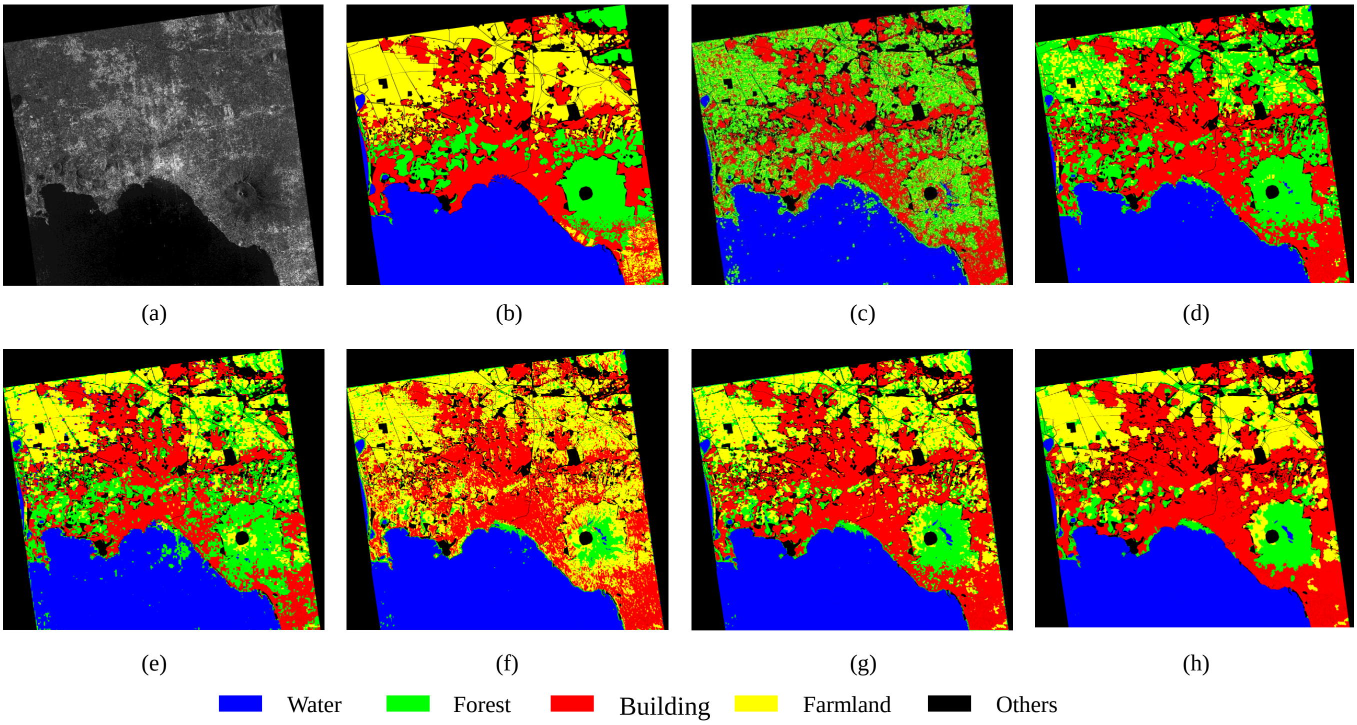


Figure 7. Visualization of terrain classification results with different methods on Napoli data. (**a**) SAR image. (**b**) Ground truth. (**c**) Deeplabv3+. (**d**) Segformer. (**e**) SimCLR. (**f**) JPSSL (no pre-training). (**g**) JPSSL (pre-training). (**h**) JPSSL (pre-training + FC-CRF).

Table 13. Terrain classification results obtained using different methods on Napoli data.

Method	**PA**	**Kappa**	**MIoU**	**FWIoU**	**F1score**			
					Water	**Forest**	**Building**	**Farmland**
Deeplabv3+	0.660	0.548	0.498	0.545	0.952	0.393	0.785	0.321
Segformer	0.740	0.654	0.577	0.620	0.974	0.531	0.861	0.384
SimCLR	0.770	0.693	0.622	0.664	0.954	0.525	0.825	0.684
JPSSL (no pre-training)	0.745	0.650	0.562	0.620	0.978	0.301	0.778	0.648
JPSSL (pre-training)	0.804	0.733	0.649	0.697	0.981	0.501	**0.872**	0.688
JPSSL (pre-training + FC-CRF)	**0.820**	**0.754**	**0.665**	**0.719**	**0.986**	**0.540**	0.853	**0.731**

3. PoDelta Data

The data of the PoDelta region of Italy and the Naples region of Italy are both obtained from the same Cosmo-SkyMed satellite with the same resolution, imaging method, and imaging band. The size of the PoDelta SAR image is 18,308 × 16,716, which is close to the size of the Napoli SAR image. Therefore, the experiment is conducted on the cut image patch size of the PoDelta image based on Table 12.

Bold in Table 14 indicates the highest values of different indicators. It can be seen from Table 14 that when the cut image patch size is 120 × 120 pixels, the downstream terrain classification task achieves the highest indicators. For the PoDelta, Italy, SAR image, the cut image patch size is set to 120 × 120 pixels, and the rest of the parameters are universal and consistent with the corresponding parameters of the JiuJiang SAR image.

The experiments are conducted under the best parameter settings presented above. Figure 8 shows the visualization results of terrain classification on the PoDelta data for the present method and the comparison methods, and Table 15 shows the evaluation indicators of terrain classification on PoDelta data for the present method and the comparison methods. Bold in Table 15 indicates the highest values of different indicators. The Deeplabv3+ method is seriously overfitted when only 20 training data are used for each class and works poorly in the forest and building classes. The indicators obtained using the JPSSL method improved to a certain extent compared with different comparative experiments. Compared with the baseline, the F1 score indicators of building and farmland increased by 45% and 22%, respectively, which shows the present method improves the discriminative performance of building and farmland features. The present method achieves the highest indicators in terms of overall indicators. Compared with the baseline, PA, Kappa, and FWIoU show an improvement of approximately 8%, and MIoU shows an improvement of roughly 14%. After using FC-CRF post-processing, PA and FWIoU are improved by approximately 2%, and Kappa and MIoU are greatly improved overall. By observing the results and indicators, it is found that the classification effect of the building category is effectively improved. A comparison of Figure 8b,h shows the excellent result achieved by the present method.

Table 14. Experiments with different cut image patch sizes in PoDelta, Italy, data.

Image Patch Size	Pretext Task		Downstream Task		
	Accuracy	PA	Kappa	MIoU	FWIoU
50 × 50	0.755	0.907	0.800	0.522	0.881
75 × 75	0.873	0.911	0.808	0.520	0.887
100 × 100	0.678	0.908	0.799	0.522	0.879
120 × 120	0.873	**0.929**	**0.846**	0.588	**0.904**
150 × 150	0.822	0.913	0.813	0.526	0.885
200 × 200	0.748	0.920	0.828	**0.595**	0.897
255 × 255	**0.918**	0.913	0.813	0.571	0.885

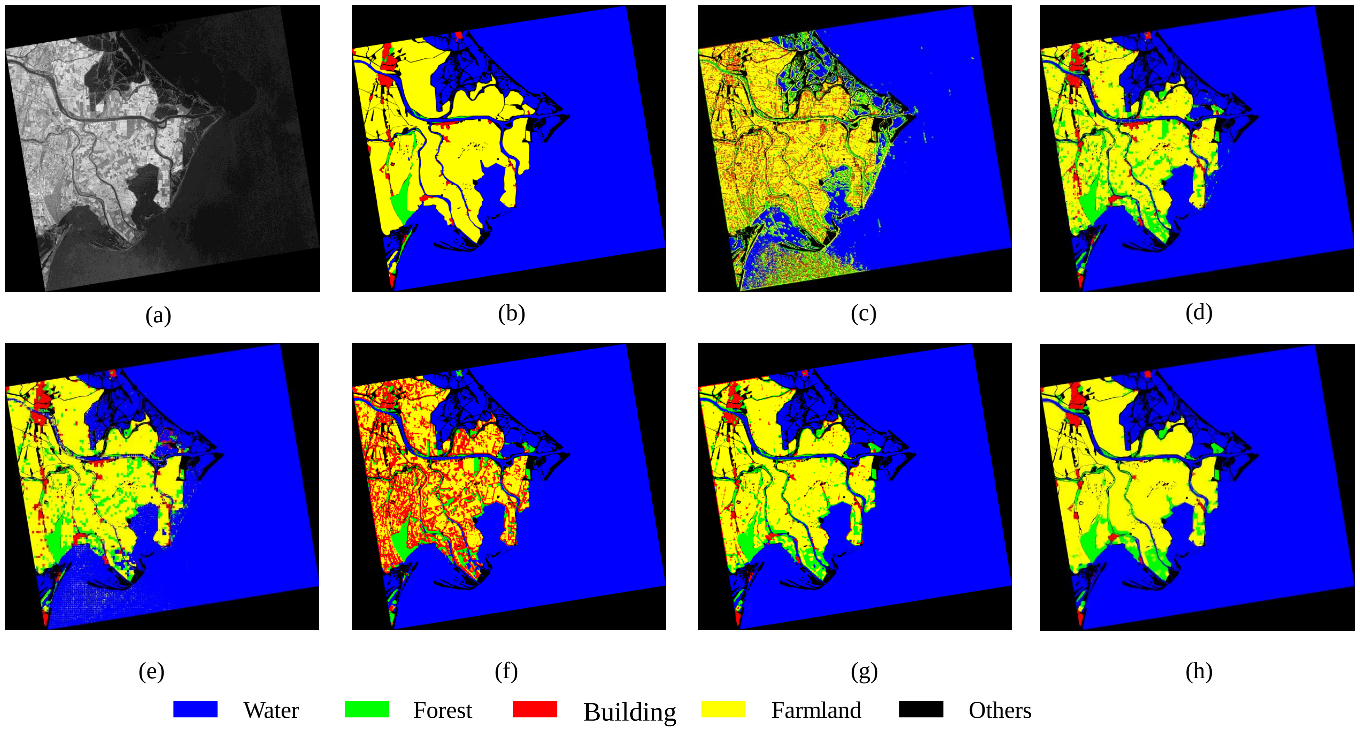


Figure 8. Visualization of terrain classification results with different methods on PoDelta data. (**a**) SAR image. (**b**) Ground truth. (**c**) Deeplabv3+. (**d**) Segformer. (**e**) SimCLR. (**f**) JPSSL (no pre-training). (**g**) JPSSL (pre-training). (**h**) JPSSL (pre-training + FC-CRF).

Table 15. Terrain classification results with different methods on PoDelta data.

Method	PA	Kappa	MIoU	FWIoU	F1score			
					Water	Forest	Building	Farmland
Deeplabv3+	0.773	0.575	0.373	0.727	0.900	0.073	0.123	0.727
Segformer	0.925	0.838	0.603	0.901	**0.989**	0.287	0.665	0.868
SimCLR	0.923	0.834	0.595	0.889	0.980	0.323	0.627	0.870
JPSSL (no pre-training)	0.842	0.672	0.442	0.814	0.987	**0.365**	0.125	0.664
JPSSL (pre-training)	0.929	0.846	0.588	0.904	0.988	0.299	0.582	0.881
JPSSL (pre-training + FC-CRF)	**0.948**	**0.885**	**0.662**	**0.924**	0.988	0.337	**0.766**	**0.918**

The method proposed in this paper shows the best performance on three different SAR images. Using this method can achieve better results with a small amount of labeled data and alleviate the drawback of difficulty in obtaining labeled data. At the same time, using this method can improve the discriminability of features and make it easier to distinguish different features in similar areas.

4. Conclusions

In this paper, a JPSSL framework for SAR image terrain classification was proposed. A jigsaw puzzle task for SAR images was designed to obtain relevant information from SAR images by shuffling the image blocks to determine the correct permutation. Applying the pre-trained model obtained from the pretext task to the downstream terrain classification task can reduce the dependence of the deep learning model on labeled data and alleviate the problem of model overfitting concerning small data. The pretext task designed in this paper can learn from the information of the image itself, and the learned features are more discriminative and can distinguish complex features. The framework designed in this paper is suitable for images of different granularities and achieved excellent results on SAR images of different resolutions and different scenes.

The method to achieve the best terrain classification accuracy for SAR images with different resolutions and sizes was explored in this paper. The results of the scene classification experiment and the terrain classification experiment show that this method achieved excellent results when using a small amount of labeled data. Meanwhile, better classification accuracy can be obtained than with fewer labeled data than other supervised comparison methods. In future work, we will explore more suitable self-supervised learning methods for SAR images with different polarization modes.

Author Contributions: Conceptualization, Z.R. and Y.L.; Data curation, W.L. and F.S.; Formal analysis, Z.R. and Y.L.; Investigation, Z.R. and Y.L.; Methodology, Y.L.; Project administration, B.H.; Resources, B.H. and W.L.; Software, Y.L.; Validation, Z.R. and Y.L.; Writing—original draft, Y.L.; Writing—review and editing, Z.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62101405, 62171347, 62371373, 62271377, and the 111 Project. It was also supported by a research project of the Shaanxi Coalfield Geological Group Co., Ltd SMDZ-2023CX-14, Shaanxi Province Water Conservancy Science and Technology, leading the Breakthrough Project LKJ2024-06, and key scientific research projects linked to the Shaanxi Provincial Department and Municipal Government 2022GD-TSLD-61-3.

Data Availability Statement: The data presented in this study are not publicly available. Some or all data, models, or codes generated or used during the study are available from the corresponding author by request. (houbiao@mail.xidian.edu.cn).

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

References

1. Zhang, W.; Zheng, L.; Wang, J.; Wang, G.; Qi, J.; Zhang, T. Application of Flood Disaster Monitoring Based on Dual Polarization of Gaofen-3 SAR Image. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 3382–3385.
2. Souza, W.d.O.; Reis, L.G.d.M.; Ruiz-Armenteros, A.M.; Veleda, D.; Ribeiro Neto, A.; Fragoso, C.R., Jr.; Cabral, J.J.d.S.P.; Montenegro, S.M.G.L. Analysis of environmental and atmospheric influences in the use of sar and optical imagery from sentinel-1, landsat-8, and sentinel-2 in the operational monitoring of reservoir water level. *Remote Sens.* **2022**, *14*, 2218. [CrossRef]
3. Gao, G.; Yao, L.; Li, W.; Zhang, L.; Zhang, M. Onboard information fusion for multisatellite collaborative observation: Summary, challenges, and perspectives. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 40–59. [CrossRef]
4. Zhang, Z.; Yang, J.; Du, Y. Deep convolutional generative adversarial network with autoencoder for semisupervised SAR image classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 4000405. [CrossRef]
5. Wang, H.; Magagi, R.; Goita, K. Polarimetric decomposition for monitoring crop growth status. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 870–874. [CrossRef]
6. Tombak, A.; Turkmenli, I.; Aptoula, E.; Kayabol, K. Pixel-based classification of SAR images using feature attribute profiles. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 564–567. [CrossRef]
7. Bai, Y.; Gao, C.; Singh, S.; Koch, M.; Adriano, B.; Mas, E.; Koshimura, S. A framework of rapid regional tsunami damage recognition from post-event TerraSAR-X imagery using deep neural networks. *IEEE Geosci. Remote Sens. Lett.* **2017**, *15*, 43–47. [CrossRef]
8. Passah, A.; Sur, S.N.; Paul, B.; Kandar, D. SAR Image Classification: A Comprehensive Study and Analysis. *IEEE Access* **2022**, *10*, 20385–20399. [CrossRef]
9. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
10. Hong, Z.Q. Algebraic feature extraction of image for recognition. *Pattern Recognit.* **1991**, *24*, 211–219. [CrossRef]
11. Dai, D.; Yang, W.; Sun, H. Multilevel local pattern histogram for SAR image classification. *IEEE Geosci. Remote Sens. Lett.* **2010**, *8*, 225–229. [CrossRef]
12. Ansari, R.A.; Buddhiraju, K.M.; Malhotra, R. Urban change detection analysis utilizing multiresolution texture features from polarimetric SAR images. *Remote Sens. Appl. Soc. Environ.* **2020**, *20*, 100418. [CrossRef]
13. Xiang, D.; Tang, T.; Zhao, L.; Su, Y. Superpixel generating algorithm based on pixel intensity and location similarity for SAR image classification. *IEEE Geosci. Remote Sens. Lett.* **2013**, *10*, 1414–1418. [CrossRef]
14. Yao, J.; Krolak, P.; Steele, C. The generalized Gabor transform. *IEEE Trans. Image Process.* **1995**, *4*, 978–988. [PubMed]
15. Lu, C.S.; Chung, P.C.; Chen, C.F. Unsupervised texture segmentation via wavelet transform. *Pattern Recognit.* **1997**, *30*, 729–742. [CrossRef]
16. Xu, Y.; Bai, T.; Yu, W.; Chang, S.; Atkinson, P.M.; Ghamisi, P. Ai security for geoscience and remote sensing: Challenges and future trends. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 60–85. [CrossRef]
17. Datcu, M.; Huang, Z.; Anghel, A.; Zhao, J.; Cacoveanu, R. Explainable, physics-aware, trustworthy artificial intelligence: A paradigm shift for synthetic aperture radar. *IEEE Geosci. Remote Sens. Mag.* **2023**, *11*, 8–25. [CrossRef]
18. Dong, S.; Wang, P.; Abbas, K. A survey on deep learning and its applications. *Comput. Sci. Rev.* **2021**, *40*, 100379. [CrossRef]
19. Su, S.; Cui, Z.; Guo, W.; Zhang, Z.; Yu, W. Explainable Analysis of Deep Learning Methods for SAR Image Classification. In Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022.
20. Wang, N.; Wang, Y.; Liu, H.; Zuo, Q.; He, J. Feature-fused SAR target discrimination using multiple convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1695–1699. [CrossRef]
21. Geng, J.; Wang, H.; Fan, J.; Ma, X. Deep supervised and contractive neural network for SAR image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2442–2459. [CrossRef]
22. Atteia, G.; Collins, M.J.; Algarni, A.D.; Samee, N.A. Deep-Learning-Based Feature Extraction Approach for Significant Wave Height Prediction in SAR Mode Altimeter Data. *Remote Sens.* **2022**, *14*, 5569. [CrossRef]
23. Yue, Z.; Gao, F.; Xiong, Q.; Wang, J.; Huang, T.; Yang, E.; Zhou, H. A novel semi-supervised convolutional neural network method for synthetic aperture radar image recognition. *Cogn. Comput.* **2021**, *13*, 795–806. [CrossRef]
24. Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-supervised learning Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 857–876. [CrossRef]
25. Ericsson, L.; Gouk, H.; Loy, C.C.; Hospedales, T.M. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Process. Mag.* **2022**, *39*, 42–62. [CrossRef]
26. Jing, L.; Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4037–4058. [CrossRef] [PubMed]
27. Wang, Y.; Albrecht, C.M.; Braham, N.A.A.; Mou, L.; Zhu, X.X. Self-supervised learning in remote sensing: A review. *arXiv* **2022**, arXiv:2206.13188.
28. Tao, C.; Qi, J.; Lu, W.; Wang, H.; Li, H. Remote sensing image scene classification with self-supervised paradigm under limited labeled samples. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 8004005. [CrossRef]

29. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5612822. [CrossRef]
30. 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]
31. Ji, H.; Gao, Z.; Zhang, Y.; Wan, Y.; Li, C.; Mei, T. Few-shot scene classification of optical remote sensing images leveraging calibrated pretext tasks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5625513. [CrossRef]
32. Markaki, S.; Panagiotakis, C. Jigsaw puzzle solving techniques and applications a survey. *Vis. Comput.* **2023**, *39*, 4405–4421. [CrossRef]
33. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1422–1430.
34. Noroozi, M.; Favaro, P. Unsupervised learning of visual representations by solving jigsaw puzzles. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 69–84.
35. Du, R.; Chang, D.; Bhunia, A.K.; Xie, J.; Ma, Z.; Song, Y.Z.; Guo, J. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 153–168.
36. Li, R.; Liu, S.; Wang, G.; Liu, G.; Zeng, B. Jigsawgan: Auxiliary learning for solving jigsaw puzzles with generative adversarial networks. *IEEE Trans. Image Process.* **2021**, *31*, 513–524. [CrossRef] [PubMed]
37. Du, W.S. Subtraction and division operations on intuitionistic fuzzy sets derived from the Hamming distance. *Inf. Sci.* **2021**, *571*, 206–224. [CrossRef]
38. Ruby, U.; Yendapalli, V. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng* **2020**, *9*, 5393–5397.
39. Krähenbühl, P.; Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 109–117.
40. Zhang, S.; Xing, J.; Wang, X.; Fan, J. Improved YOLOX-S Marine Oil Spill Detection Based on SAR Images. In Proceedings of the 2022 12th International Conference on Information Science and Technology (ICIST), Kaifeng, China, 14–16 October 2022; pp. 184–187.
41. 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.
42. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
43. 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, Virtual, 13–18 July 2020; pp. 1597–1607.

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.