remote sensing

MDPI

Article

Contrastive Self-Supervised Two-Domain Residual Attention Network with Random Augmentation Pool for Hyperspectral Change Detection

Yixiang Huang [1,2], Lifu Zhang [1,*], Wenchao Qi [1], Changping Huang [1] and Ruoxi Song [1]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, No. 20 Datun Road, Beijing 100101, China; huangyixiang20@mails.ucas.ac.cn (Y.H.); qiwc@aircas.ac.cn (W.Q.); huangcp@aircas.ac.cn (C.H.); songrx@aircas.ac.cn (R.S.)
[2] University of Chinese Academy of Sciences, No. 3 Datun Road, Beijing 100101, China
* Correspondence: zhanglf@radi.ac.cn

Abstract: Hyperspectral images can assist change-detection methods in precisely identifying differences in land cover in the same region at different observation times. However, the difficulty of labeling hyperspectral images restricts the number of training samples for supervised change-detection methods, and there are also complex real influences on hyperspectral images, such as noise and observation directions. Furthermore, current deep-learning-based change-detection methods ignore the feature reusage from receptive fields with different scales and cannot effectively suppress unrelated spatial–spectral dependencies globally. To better handle these issues, a contrastive self-supervised two-domain residual attention network (TRAMNet) with a random augmentation pool is proposed for hyperspectral change detection. The contributions of this article are summarized as follows. (1) To improve the feature extraction from hyperspectral images with random Gaussian noise and directional information, a contrastive learning framework with a random data augmentation pool and a soft contrastive loss function (SCLF) is proposed. (2) The multi-scale feature fusion module (MFF) is provided to achieve feature reusage from different receptive fields. (3) A two-domain residual attention (TRA) block is designed to suppress irrelated change information and extract long-range dependencies from both spectral and spatial domains globally. Extensive experiments were carried out on three real datasets. The results show that the proposed TRAMNet can better initialize the model weights for hyperspectral change-detection task and effectively decrease the need for training samples. The proposed method outperforms most existing hyperspectral change-detection methods.

Keywords: change detection; hyperspectral image; contrastive learning; attention mechanism

Citation: Huang, Y.; Zhang, L.; Qi, W.; Huang, C.; Song, R. Contrastive Self-Supervised Two-Domain Residual Attention Network with Random Augmentation Pool for Hyperspectral Change Detection. *Remote Sens.* **2023**, *15*, 3739. https://doi.org/10.3390/rs15153739

Academic Editor: Pedro Melo-Pinto

Received: 16 May 2023
Revised: 18 July 2023
Accepted: 25 July 2023
Published: 27 July 2023

1. Introduction

Hyperspectral remote sensing is one of the most important remote sensing imaging techniques, and it can capture subtle changes in the Earth's surface with a number of narrow spectral bands [1]. Hyperspectral change-detection techniques aim to identify changes in land cover by comparing images acquired at different times [2], and they have been extensively applied in various research fields, such as farmland inspection, forest-degradation observation, and urban-disaster monitoring [3–5]. Generally, the procedure for existent hyperspectral change-detection methods can be roughly divided into three different stages, namely data preprocessing of hyperspectral images, map-change generation with appropriate change-detection methods, and result evaluation for the change prediction.

According to the appearance period, change-detection technologies can be categorized into traditional change-detection methods and deep learning methods. Hyperspectral change detection using traditional methods can be classified into image-comparison operator-based approaches, transformation-based approaches, and independent image classification approaches. The main idea of image-comparison operator-based approaches

is to infer the change by computing the characteristic space distance between image pairs, but this brings the difficulty of choosing a proper threshold for precisely distinguishing changes, such as with change vector analysis (CVA) [6,7] and similarity measures [8]. For transformation-based methods, the high computational cost of processing the highly dimensional data should be considered, using, for example, principal component analysis (PCA) [9,10] and band selection [11,12], which inevitably lead to a constant information loss. As for the independent image classification, because of two independent classification stages for generating a single change map, the detection result may be affected by the error propagation from both image-classification results [13,14].

Deep learning methods have stimulated a rapid evolution in the hyperspectral change-detection field, and they can be roughly categorized into supervised deep learning methods, semi-supervised deep learning methods, and unsupervised deep learning methods. For supervised deep learning methods, existing methods mainly focus on extracting internal feature expressions using data more effectively with a credible ground truth. These methods intend to propose a convolutional neural network to aggregate multiple direction information by applying 1-D convolution to achieve spectral feature extraction [15], traditional 2-D convolution [16] to obtain spatial information, or constructing a 3-D convolutional network to aggregate spectral–spatial information [17,18]. Zhan et al. combined the characteristics of 1-D and 2-D convolution to extract features from spectral and spatial domains simultaneously [19]. Moustafa et al. proposed a deep CNN semantic-segmentation-based workflow to cope with the complex nature of hyperspectral images and their high dimensionality [20]. However, existing methods capture deep features hierarchically and cannot fully use features from different receptive fields to predict specific tasks. In our previous work, the principal part of our deep model uses convolution operations to achieve stable feature representation as well [21]. Moreover, deep convolutional networks will inevitably meet bottlenecks, which result from ignoring spatial–spectral similarity dependencies in input dual-temporal images. For semi-supervised deep learning methods, the adversarial autoencoder is applied to reconstruct the input spectral vector using a spectral mapping loss function. Based on the reconstructed hyperspectral image pairs, unsupervised methods, i.e., the PCA and Otsu threshold, are used to output classification results and change results [22]. In this instance, deep learning methods are considered as preprocessing and compression tools, which cannot exert the end-to-end detection advantage of deep learning models. In existing deep hyperspectral change-detection methods, the reality of time-consuming labeling and unreliable visual interpretation is often ignored. To address this problem, Ou et al. introduce self-supervised contrastive learning, which is one type of unsupervised learning method, into hyperspectral change detection with random Gaussian noise to obtain a pre-trained model [23]. However, the downstream change-detection task highly relies on both an upstream contrastive training strategy and a sophisticated model design. The appropriate design of the pre-text task should take the actual image reality into consideration, such as noise, direction, and symmetry. However, current self-supervised contrastive methods do not consider that a real hyperspectral image is usually influenced by a set of random and complex data-collection scenarios. Moreover, the effective pre-train model highly depends on the recognition of positive samples from a large number of negative samples. In hyperspectral images, pixels can be similar to spectral information, while contrastive loss may excessively force the model to separate highly similar samples into different types, which makes it difficult for the model to learn the accurate and proper inner structural information of hyperspectral images. Therefore, in the instance of self-supervised learning using limited labeled training samples, (1) the inadequate feature reusage from different receptive fields, (2) the insufficiencies of the extractionof spectral and spatial similarity dependencies, and (3) the influence of complex data-collection scenes burden the improvement in current hyperspectral change detection performance immensely.

To overcome these drawbacks of existing methods, in this paper, a contrastive two-domain residual attention network with a random augmentation pool is proposed for hyperspectral change detection. The proposed method is trained in two self-supervised

phases. The first phase is used to train the deep model with contrastive self-supervised learning, which does not need labeled training samples. In this step, the inner representation of hyperspectral data is extracted to initialize the model weights. By applying a random augmentation pool, the real and complex data-collection scenario can be simulated to generate positive and negative sample pairs. At the same time, it can be seen as a normalization tool to relieve model instability from excessive sample separations. The second phase is arranged in a supervised form with both few labeled training samples and the upstream pre-trained model. During the whole feature-extraction lifecycle, the multi-scale feature fusion module is proposed to enhance the feature reusage from different receptive fields. Meanwhile, a two-domain residual attention block is constructed and applied through the whole deep model, which can effectively capture the long-range dependencies from both spectral and spatial domains while maintaining the model stability and feature consistency.

The contributions of this paper are as follows.

(i) Contrastive learning (CL) with the random data-augmentation pool is introduced into the model-training procedure, which aims to improve the feature extraction in the reality of time-consuming labeling and insufficient training samples. By maximizing the agreement between differently augmented views of the same data example, via a contrastive loss in the latent space, the learned feature representations of positive and negative pairs can effectively improve the downstream change-detection performance. Furthermore, the soft contrastive loss function is proposed to improve the inadequate feature tolerance caused by the hard discrimination of pseudo samples.

(ii) A multi-scale feature-fusion module (MFF) is proposed for a better feature reusage. By storing the features with different resolutions, feature reusage from different receptive fields can be achieved, which improves the hyperspectral change-detection performance with spectral–spatial-related information in shallow layers and class-oriented semantic information in deep layers.

(iii) A two-domain residual attention block (TRA) is contributed to extract long-range dependencies from spectral and spatial domains globally. To effectively obtain the resemblance information from the whole feature-extraction process, TRA is hierarchically applied before every convolutional layer. Moreover, the residual connection is introduced to improve the model's consistency and stability.

2. Related Work

2.1. Contrastive Learning

Contrastive learning belongs to a branch of discriminative approaches that aims to group similar samples closer to each other and diverse samples far from each other [24]. Because of its semi-supervised clustering-like strategy, it can extract inner feature expressions from sample data without ground-truth labels [25]. Therefore, CL is an ideal solution for the situation where labeling is very difficult. Based on the knowledge acquired by CL, the pre-trained model can be transferred to downstream hyperspectral tasks in a supervised manner. To achieve the effectiveness of CL in a downstream hyperspectral classification task, contrastive self-supervised learning is applied to learn spectral–spatial contrastive features with limited hyperspectral labeled samples [26–28]. Hang et al. provide a method of using both hyperspectral and LiDAR (light detection and ranging) data to explore the semantic information and the information on the intrinsic structure, which proves that multimodal data can benefit the effectiveness of CL [29]. Although little research has discussed the potential of CL in hyperspectral change detection, it remains in its early stage, lacking the exploration of complex actual application scenes in CL. Ou et al. have introduced a self-supervised CL framework for hyperspectral change detection. With the Gaussian noise data augmentation, the feature extraction model is pre-trained using a contrastive loss function [23]. Moreover, based on similarity metrics, the contrastive loss function is designed to strictly recognize the positive sample from a large number of negative samples. In the reality of hyperspectral data redundancy, the hard discrimination may cause distortion to the real structural semantic information, which is opposite to the factor

hyperspectral data. Therefore, there is a very urgent need to improve the adaptiveness of self-supervised contrastive learning methods in hyperspectral change detection.

2.2. Attention Mechanism

The attention mechanism is initially presented in the Transformer framework for the natural language process domain, which can address sequence-to-sequence transformation problems and is able to aggregate all resemblance information from the entire sequential input [30]. As the core component, the attention mechanism has been adopted in hyperspectral image processing as well, due to its ability to adaptively suppress task-irrelevant spectral and spatial information. By capturing the rich spatial–spectral information about HIS, the transformer framework now has been studied extensively for hyperspectral classification, such as [31,32]. For hyperspectral change detection, the attention mechanism is used to consider the different contributions from different spectral channels and spatial locations from dual-temporal input image paths, which can emphasize informative channels and locations and adaptively suppress less informative ones [33]. To handle a large amount of irrelevant or noisy spectral and spatial information, adaptive spectral and spatial attention mechanisms with Gaussian distribution can help the model to reduce the sensitivity of patch size in patch-based methods [34]. Based on the attention mechanism, it is effective for extant models to emphasize spectral band and location resemblances while adaptively suppressing unrelated information. However, the existing change methods lack resemblance computation due to feature pooling or convolution in single-feature maps, which leads to inconsistent feature weighting. Moreover, the existing attention mechanism is usually designed as a plug-and-play module, which cannot obtain globally hierarchical information through the whole model. Therefore, it is necessary to make full use of the attention mechanism in hyperspectral change detection.

3. Proposed Method

To address the problems mentioned above and improve the stability and performance of the deep learning model, the two-domain residual attention network (TRAMNet) with a random augmentation pool is proposed for hyperspectral change detection in a contrastive self-supervised pattern. The overall architecture and module arrangement are shown in Figure 1.

The whole training procedure can be divided into two steps. The first step is contrastive self-supervised learning [25]. The hyperspectral image dataset for change detection contains dual-temporal images, which can be described as T1 and T2. And their augmentation versions can be described as T1′ and T2′, which are processed by the augmentation pool. The differencing image patches of dual-temporal images T1 and T2 are inputted into the TRAMNet encoder $f(\cdot)$, and both the original and augmented differencing samples are packed into the sample batch to train their feature representation vectors. The TRAMNet encoder $f(\cdot)$ consists of the attention encoder and the dense convolution block, where the dense convolution block is composed of three successive convolutional bottleneck layers, which are connected densely and can enhance feature reusage for contrastive vector projection or change discrimination. During a specific batch training, the augmented differencing version of the original differencing sample is considered as a positive sample, and the rest of the samples in the batch are considered as negative samples to the selected original sample. The soft contrastive loss function is designed to make a rough discrimination of positive sample pairs from a large number of corresponding negative sample pairs. The second step is the supervised training using model transferring to initialize the model weights effectively and fine-tune the final change-detection map. After these two steps, the trained model can be used to detect the change area of hyperspectral images.

In this section, the proposed method is arranged into four components to give detailed information about every design, namely the random data augmentation pool, the soft contrastive loss function, and the attention encoder, which consists of the muti-scale feature fusion module and the two-domain residual attention block.

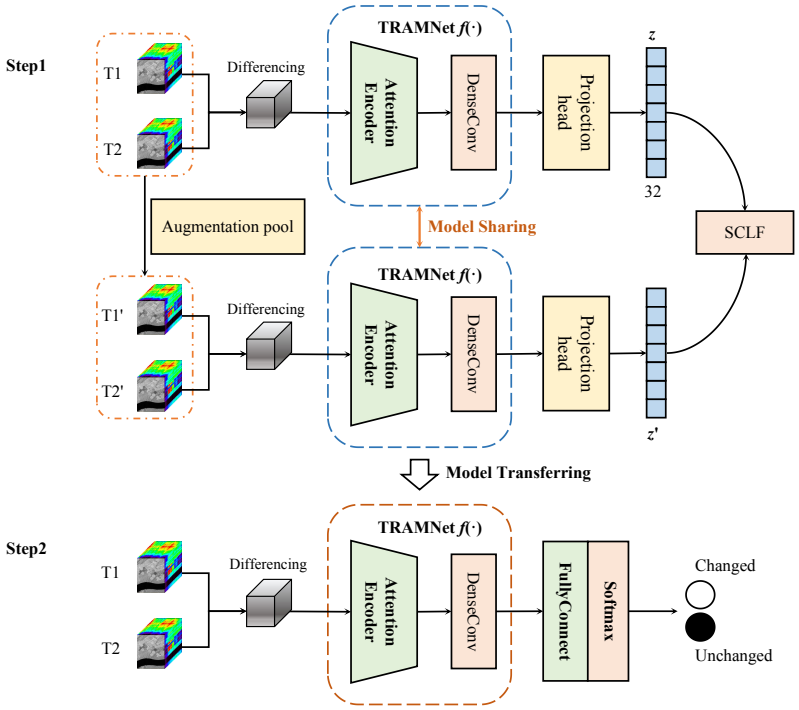


Figure 1. The overall contrastive architecture and module arrangement of the TRAMNet.

3.1. Random Data Augmentation Pool

In real scenarios, hyperspectral images have a significant amount of complicated real noise, and the texture and spectrum information in hyperspectral images includes various directional information as well. Therefore, the change-detection methods would be deeply affected by these real factors. To address these problems, the random data-augmentation pool is proposed to generate an augmented version of training samples. In this article, random Gaussian noise [23,35], the rotate operation, and the flip operation are adopted to simulate these real scenarios.

Among these augmentation methods, Gaussian noise with a zero mean value is often used to simulate real noise, which can distort the original high-frequency features to a certain extent. By adding proper noise into images, the learning stability and robustness of deep networks can be enhanced. The data augmented using the random Gaussian noise can be represented as

$$Input' = Input + \lambda \cdot N, \tag{1}$$

where *Input* is the original patch input, and N indicates the random Gaussian noise. Mathematically, Gaussian noise is a continuous random variable that obeys normal distribution $X \sim N(\mu, \sigma^2)$. The influence of random Gaussian can be adjusted by λ, which is set to $1/25$ [35]. The probability distribution function of random Gaussian noise can be defined as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{2}$$

3.2. Soft Contrastive Loss Function

Self-supervised learning can extract a rough feature representation from a batch of positive and negative samples without a ground truth. After image differencing, both the input patch x and its augmened version x' are trained by the same TRAMNet $f(\cdot)$. To obtain sample feature vectors from feature representations r and r', the projection head is simply developed to project vectors z and z', whose the length is 32. The projection head

is constructed with two layers of fully connected layers and batch-normalization layers. The SELU (scaled exponential linear units) is chosen as the activation function. Based on the similarity metrics, the soft contrastive loss function (SCLF) is constructed, which can be settled to roughly distinguish these positive pairs and negative pairs from the projected feature vectors. Supposing that z_i and z_j are two projections of sample representations, the cosine similarity between example i and example j can be defined as

$$sim(z_i, z_j) = \frac{z_i^\top z_j}{\|z_i^\top\| \|z_j\|}. \tag{3}$$

Because the training samples are trained in batches, the loss function between the positive pair of examples i and j can defined as [25]

$$loss_{i,j} = -\log \frac{\exp\left(sim(z_i, z_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbf{I}_{k \neq i} \exp\left(sim(z_i, z_k)/\tau\right)}, \tag{4}$$

where τ denotes a temperature scale parameter, which can control the influence of cosine sample similarity. And $\mathbf{I}_{k \neq i}$ is an indicator function evaluated as 1 if $k \neq i$. To calculate the final loss, a batch of original samples and augmented samples can be arranged as $Z = [z_0, z_1, z_2, z_3, z_4, z_5, \ldots, z_{2N-2}, z_{2N-1}] \in \mathbf{R}^{32 \times N}$, where the feature vector of the i-th sample is denoted as z_i. And the original sample and its corresponding augmented sample are arranged into adjacent positions. To obtain the final loss, the $loss_{i,j}$ and $loss_{j,i}$ are considered at the same time, which means all the losses for positive pairs in a batch should be computed. Therefore, the final loss can be defined as

$$loss = \lambda \frac{1}{2N} \sum_{k=0}^{N-1} (loss_{2k-2,2k-1}, loss_{2k-1,2k-2}), \tag{5}$$

where λ denotes a soft coefficient to adjust the degree of strictness of the overall positive and negative extracted feature vectors. The larger the value is, the stricter the discrimination of positive and negative pairs is. The SCLF is applied in the first step to pre-train the parameters of TRAMNet, which are saved for later fine-tuning to obtain the final change-detection map in a supervised manner.

3.3. Attention Encoder

In self-supervised contrastive learning, the encoder $f(\cdot)$ is mainly learned in the first step for the second supervised fine-tuning. In this study, the representation learning of input hyperspectral patches is processed by both the attention encoder and the dense convolution block. The feature information stability can be adjusted with dense connections among convolutional bottleneck layers, and the inner structure design of the attention encoder for main feature extraction is shown in Figure 2.

As the top of Figure 2 shows, the attention encoder extracts features using reduplicate residual attention blocks (TRA), residual convolution blocks (RConvBlock), and a muti-scale feature fusion module (MFF). There are no specific restrictions for input patch size due to the feature-shape normalization by linear interpolation. The input patch size would be tailored to $c \times 9 \times 9$, and c indicates the band number of input patches. In the attention encoder, the feature map resolution is adjusted from 9×9 to 3×3 by max pooling operation at every level, which can reserve significant features while decreasing the computational cost to prevent model overfitting.

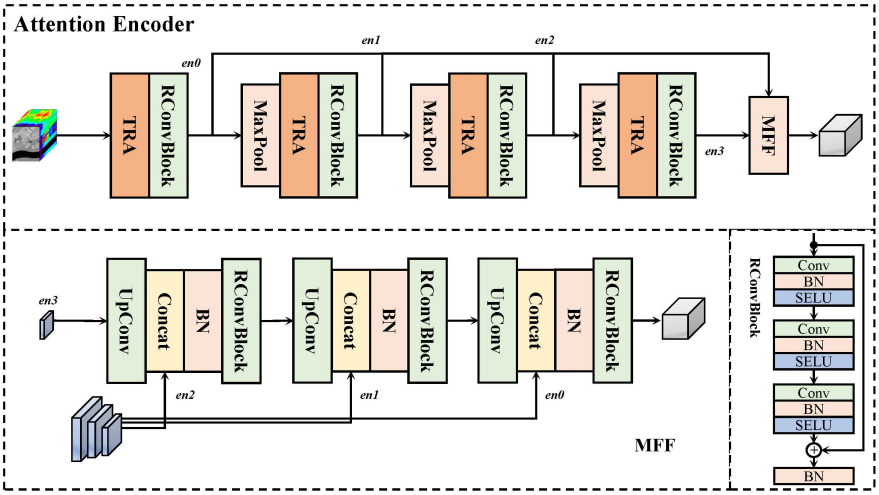


Figure 2. The design of the attention encoder, MFF module, and RConvBlock.

There are four intermediate encoder outputs $\{en_0, en_1, en_2, en_3\}$, which are defined with channels of $[256, 128, 128, 64]$, respectively, to be fused by MFF. The attention encoder intermediate outputs can be defined as

$$en_{i+1} = F_{ResAttention}(F_{RConvBlock}(F_{MaxPool}(en_i))), i = 0, 1, 2, \tag{6}$$

where $F_{MaxPool}(\cdot)$ indicates the max pooling operation, $F_{RConvBlock}(\cdot)$ indicates the residual convolutional block defined below, and $F_{ResAttention}(\cdot)$ denotes the residual attention block. From the input patch x_{in} after preprocessing, the feature map of en_0 is generated with the follow operation:

$$en_0 = F_{ResAttention}(F_{RConvBlock}(x_{in})) \in \mathbf{R}^{256\times 9\times 9}. \tag{7}$$

The RConvBlock is designed with three successive convolutional layers and a skip connection, as shown in the bottom-right part of Figure 2. The convolutional layer can be defined as

$$F_{ConvLayer}(x) = F_{SELU}(F_{BN}(F^{3\times 3}_{Conv}(x))), \tag{8}$$

where $x \in \mathbf{R}^{c\times h\times w}$ indicates the input feature maps among every feature extraction component. $F^{3\times 3}_{Conv}(\cdot)$ indicates the convolution layer with a 3×3 kernel size. $F_{BN}(\cdot)$ denotes the batch normalization layer, and $F_{SELU}(\cdot)$ denotes the SELU activation function. Therefore, the RConvBlock can be represented as

$$F_{RConvBlock}(x) = F_{BN}(x + F^{3}_{ConvLayer}(x)). \tag{9}$$

The input feature map x is processed using convolutional layers and then linked with a skip connection, followed by batch normalization. This can improve both the information flow of convolutional operations and the disorganized data distribution in the encoder representation learning.

3.3.1. MFF Module

Features from different layers have different receptive field scales. Shallow features can easily contain more detailed information due to the small receptive fields, which can benefit the discrimination of the refined change information. On the contrary, deep features contain more semantic information due to wider receptive fields. Therefore, the muti-scale feature-fusion module (MFF) is added to enhance the feature-change representations more effectively by making fusion with shallow features and deep features. The design of the MFF is shown in the downside of Figure 2.

The MFF is designed to fuse encoder feature maps $\{en_0, en_1, en_2, en_3\}$ one by one and constructed with reduplicate up-convolutional blocks (UpConv) and RConvBlock. The UpConv block can be defined as

$$F_{Up}(\cdot) = F_{SELU}(F_{BN}(F_{Conv}^{3\times3}F_{Upsample}(\cdot))) \tag{10}$$

where $F_{Upsample}(\cdot)$ indicates the up-sampling operation to feature maps using the nearest mode. To better explain the process of feature fusion, the formulation of the MFF is defined as follows:

$$x_i^{out} = F_{RConvBlock}(F_{BN}(F_{cat}(en_i, F_{Up}(en_{i+1})))) \in \mathbf{R}^{c\times h\times w}, i = 0, 1, 2, \tag{11}$$

where $F_{cat}(\cdot)$ indicates the concatenation operation.

3.3.2. TRA Block

The attention mechanism has now been extensively studied in many research fields. To extract the long-range dependencies from the spatial and spectral domain while keeping a fluent information flow for cooperation with convolutional blocks, the two-domain residual attention block (TRA) is developed. The TRA blocks are set into every layer of the attention encoder. The detailed design of the TRA is shown in Figure 3.

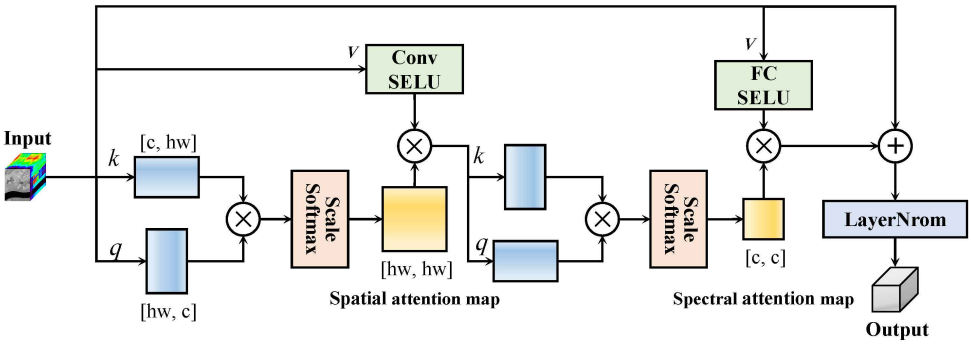


Figure 3. The design for TRA block.

The computation of an attention map plays an important role in weighting feature maps, which can be regarded as variants of the self-attention mechanism. The weighted feature maps for spatial and spectral domains are mainly determined by three parameters, namely query (q), key (k), and value (v). For an input patch feature map $x_{in} \in \mathbf{R}^{c\times h\times w}$, the computation of the spatial attention map can be formulated as follows:

$$AM_{spa} = F_{Softmax}(\frac{q_{spa} \otimes k_{spa}}{\sqrt{d_k}}) \in \mathbf{R}^{hw\times hw}, \tag{12}$$

where $AM_{spa} \in \mathbf{R}^{hw\times hw}$ indicates the attention map for spatial domain, and the $\otimes$ indicates the matrix multiplication. The SoftMax activation function is represented as $F_{Softmax}(\cdot)$. And $q_{spa} \in \mathbf{R}^{hw\times c}$ and $k_{spa} \in \mathbf{R}^{c\times hw}$ represent parameters of spatial query and spatial key, respectively. The parameter of d_k is the dimension of queries and keys, which is considered as the scale factor to improve the numerical stability. After obtaining the spatial attention map, the weighted feature map can be computed by

$$out_{spa} = AM_{spa} \otimes F_{SELU}(F_{conv}^{3\times3}(v_{spa})) \in \mathbf{R}^{c\times hw}, \tag{13}$$

where $out_{spa} \in \mathbf{R}^{c\times hw}$ indicates the weighted attention map in the spatial domain and $v_{spa} \in \mathbf{R}^{c\times h\times w}$ is the parameter of spatial value. As for the computation of the spectral attention map, it can be represented as

$$AM_{spec} = F_{Softmax}(\frac{q_{spec} \otimes k_{spec}}{\sqrt{d_k}}) \in \mathbf{R}^{c \times c}, \tag{14}$$

where $AM_{spec} \in \mathbf{R}^{c \times c}$ indicates the attention map for the spectral domain. The $q_{spec} \in \mathbf{R}^{c \times hw}$ and $k_{spec} \in \mathbf{R}^{hw \times c}$ indicate parameters of spectral query and spectral key, respectively. With the obtained spectral attention map, the final output of the two-domain residual attention mechanism can be represented as

$$out_{attention} = F_{LN}(x_{res} \oplus AM_{spec} \otimes F_{SELU}(F_{fc}(v_{spec}))) \in \mathbf{R}^{c \times h \times w}, \tag{15}$$

where $\oplus$ indicates the matrix plus and $x_{res} \in \mathbf{R}^{c \times h \times w}$ indicates the residual component from original input patch x_{in}. $F_{fc}(\cdot)$ denotes the fully connected layer used to process spectral dimension, and $F_{LN}(\cdot)$ denotes the layer normalization function to process sequence normalization.

4. Experiments

4.1. Description of the Dataset

In this paper, three public dual-temporal datasets for hyperspectral change detection are chosen for verifying the effectiveness of the proposed TRAMNet, and they are all captured using the Earth Observing-1 (EO-1) satellite with Hyperion sensor, which provides a spectral range of 0.4–2.5 µm with 242 spectral bands and a spectral resolution of approximately 10 nm, as well as a spatial resolution of 30 m. In experiments, spectral bands with a low signal-to-noise ratio (SNR) are removed. To distinguish whether or not the land cover area is changed, the binary ground truth maps are obtained through visual analysis and on-the-spot investigation. The first hyperspectral image dataset is the Irrigated Agricultural Dataset [2] captured on 1 May 2004 and 8 May 2007, which illustrates an irrigated agricultural area of Hermiston city in Umatilla County, Oregon, USA. It contains 307 × 241 pixels and 156 bands after omitting no-data bands and removing the noise. The training rate follows its original work, which is approximately 9.7%, which is set as the benchmark rate for training sample analysis. The false-colour map of this dataset is shown in Figure 4. The second dataset is the Wetland Agricultural Dataset [36], which contains images captured on 3 May 2006 and on 23 April 2007. This dataset illustrates a farmland area of Yuncheng City, Zhejiang Province, China. It contains 450 × 140 pixels and 156 bands after removing noise. For convenience, it refers to the training rate of irrigated agriculture dataset and is set to about 9.7%, which is the benchmark for the analysis of training samples as well. Compared to its original work, the number of training samples for the wetland dataset is further decreased. The T1 image, T2 image, and ground truth for this dataset are shown in Figure 5. The last change-detection dataset is the River Dataset [16], which contains images captured on 3 May 2013 and 31 December 2013. This dataset illustrates a river area in Jiangsu Province, China, and it contains 463 × 241 pixels and 198 bands after noise removal. The false-colour images for the river dataset are shown in Figure 6. For the river dataset, to consider the problem of strict class imbalance, the proportion between unchanged samples and changed samples is set as 2:1, which follow the rule of the original work. Eventually, the training rate for the river dataset is 4.03%, which is the benchmark training rate for sample-size analysis. Every dataset is divided into a training set, validation set, and test set. For all three public datasets, stratified random sampling is used to generate random training samples. The details of every dataset are shown in Table 1.

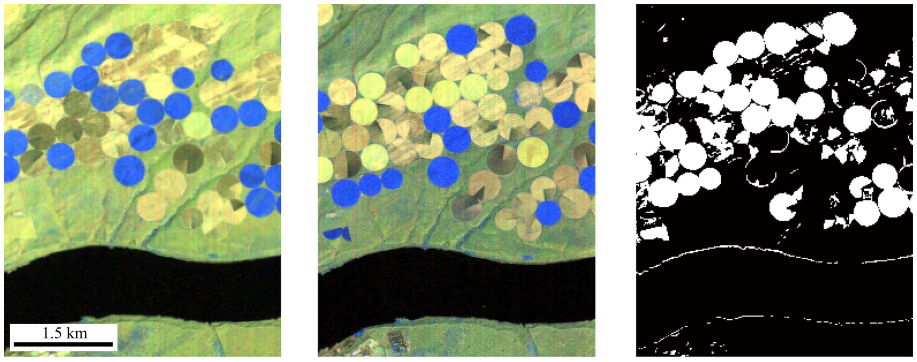


Figure 4. Irrigated agricultural dataset with a false-colour map (bands 134, 90, and 75 as RGB). (**a**) USA farmland image captured on 1 May 2004. (**b**) USA farmland image captured on 8 May 2007. (**c**) Binary ground truth for Irrigated Agricultural Dataset.

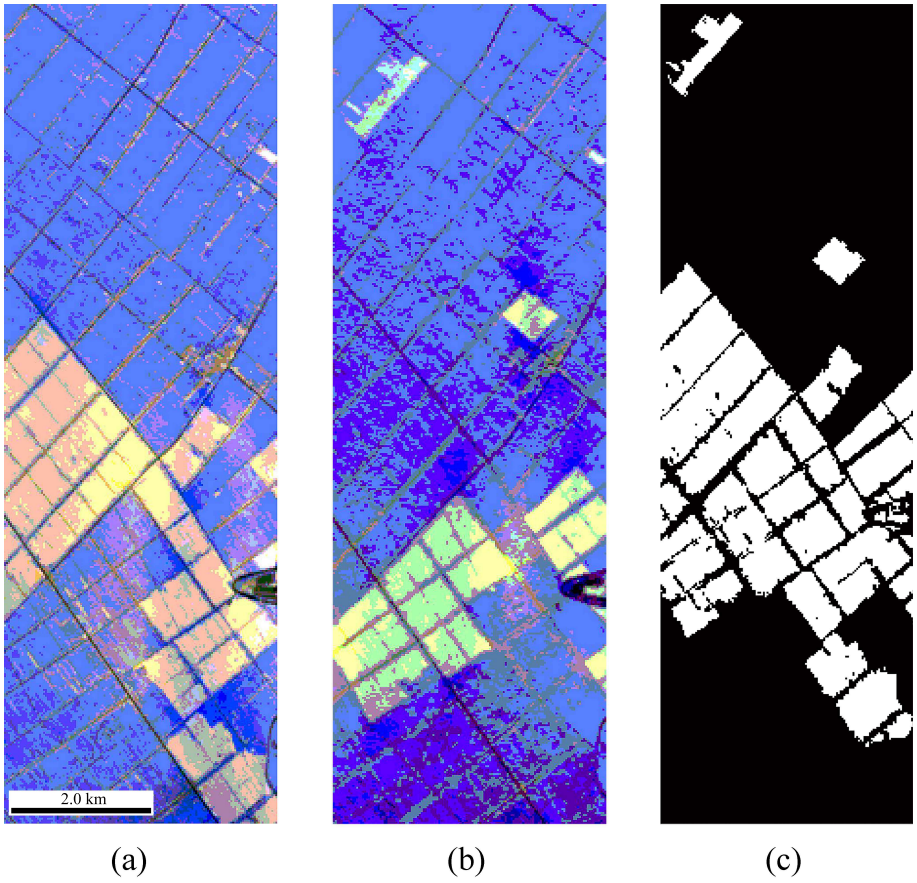


Figure 5. Wetland agricultural dataset with false-colour a map (bands 134, 90, and 75 as RGB). (**a**) China farmland captured on 3 May 2006. (**b**) China farmland captured on 23 April 2007. (**c**) Binary ground truth for Wetland Agricultural Dataset.

Table 1. Details of Three Public Datasets.

Dataset	Spatial Size	Band	Date 1	Date 2	Training Rate	Training Samples
Irrigated	307 × 241	156	1 May 2004	8 May 2007	9.7%	7250
Wetland	450 × 140	156	3 May 2006	23 April 2007	9.7%	6173
River	463 × 241	198	3 May 2013	31 Decemeber 2013	4.03%	4500

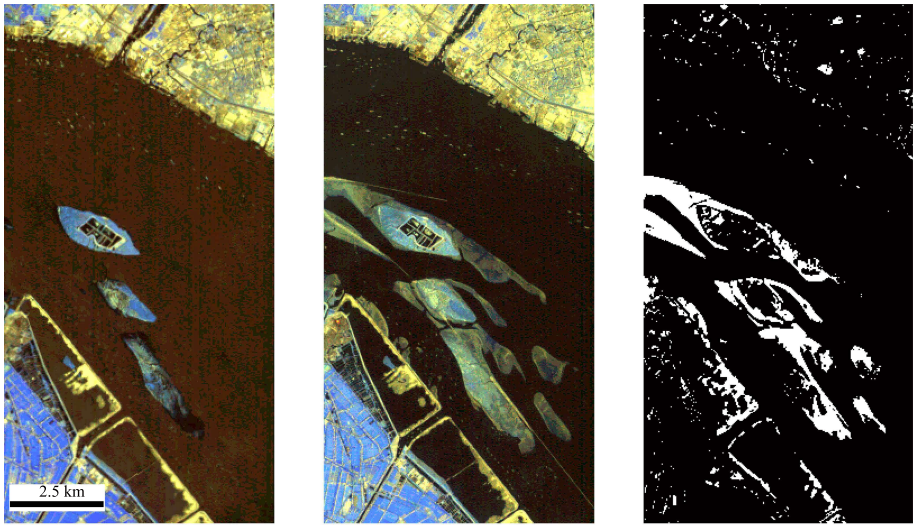


Figure 6. River dataset with false-colour map (band 134, 90, and 75 as RGB). (**a**) China river, captured on 3 May 2013. (**b**) China river, captured on 31 December 2013. (**c**) Binary ground truth for the river dataset.

4.2. Experimental Setup

All experiments were performed on an NVIDIA RTX 3060 (Nvidia Corporation, Santa Clara, CA, USA) with 12 GB GPU memory, and the implementation of this experiment was run on the PyTorch platform. Due to the model's structure, it can handle different patch size inputs without a specific restriction on spatial resolution. In the contrastive training procedure, the SGD was chosen as the optimizer with the proposed soft contrastive loss function, while in the fine-tuning step, the Adagrad was chosen as the optimizer with the cross-entropy loss function. The cross-entropy loss function can be defined as follows:

$$L = \frac{1}{N}\sum_{i} L_i = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic}), \tag{16}$$

where M indicates the number of categories and y_{ic} indicates that the value is equal to 1 or 0 depending on whether it is the true category of sample i. The prediction probability belonging to c of sample i is expressed as p_{ic}.

To show the superior effectiveness of the proposed TRAMNet, several conventional hyperspectral change detection methods and deep learning methods with state-of-the-art performance were chosen to make a comparison. CVA is a traditional method using a difference map of dual-temporal hyperspectral images and an unsupervised Otsu threshold method to segment the change map [37]. The second traditional method for hyperspectral area-change recognition is SVM, which is extensively used in the detection of land-cover changes [38]. As for deep learning methods, the LSConvolution-architecture-based GETNET [16] was chosen as a comparison method. Because the proposed method has the conception of spectral and spatial domain learning, the WCRN [39] was chosen as the two-domain learning comparison method, which was implemented with the CNN architecture. For the transformer-based method, the ChangeFormer was chosen as the comparison method [40]. The TFR-PS^{2}ANet, which is an applied attention mechanism-based module in previous work, was also chosen as a comparison method [21]. The same default parameter settings introduced in the corresponding works are settled for these mentioned deep learning-based methods.

4.3. Evaluation Metrics

For a fair and comprehensive performance comparison among three hyperspectral change-detection datasets, the accuracy (*Acc*), kappa coefficient (*kappa*), *F1-score*, *precision*, and *recall* were selected to evaluate and quantize the models; performance. Every index was calculated based on a confusion matrix, and the larger the value, the better the performance.

Accuracy (*Acc*). In the aspects of pixel-level classification tasks, accuracy is a relatively simple but effective metric to weigh model presentation. The formulation of accuracy calculation is indicated as

$$Acc = \frac{\sum_{i=0}^{c} TP_i}{\sum_{i=0}^{c}(TP_i + FP_i)}. \tag{17}$$

Kappa coefficient (*kappa*). The *kappa* coefficient is another metric usually used for pixel-level classification. According to the formulation of *kappa*, it takes the class imbalance into consideration and can fairly measure model performance in different datasets. The formulation for *kappa* calculation is shown as:

$$kappa = \frac{p_o - p_e}{1 - p_e}. \tag{18}$$

F1-score. *F1-score* (also known as *F1-measure*) is designed to evaluate the performance of the pixel-level binary classification model. It can be considered as the harmonic mean of precision and recall. The *F1-score* can be calculated as

$$F1\text{-}score = 2 \cdot \frac{precision \times recall}{precision + recall}. \tag{19}$$

Precision and Recall. The *precision* indicates that true positives occupy the sum of true-positive and false-positive samples. The *recall* indicates that true positives occupy the sum of true-positive and false-negative samples. The formulations can be represented, respectively, as

$$precision = \frac{TP}{TP + FP}, \tag{20}$$

$$recall = \frac{TP}{TP + FN}. \tag{21}$$

4.4. Experimental Results

Extensive experimentswere conducted on all three hyperspectral change-detection datasets three times. The detailed model comparison results are described below and are explained using mean and standard deviation. And the overall accuracy, kappa, and F1-score comparison results are shown in Figure 7. According to the overall accuracy, kappa, and F1-score, the proposed TRAMNet performs best. However, the WCRN performs worst on irrigated and wetland agricultural datasets according to the overall accuracy and kappa. As traditional change-detection methods, SVM outperforms CVA on the irrigated agriculture dataset. On the wetland agriculture dataset, both methods perform similarly to each other. However, CVA outperforms SVM on the river dataset, probably due to the serious class imbalance.

4.4.1. Experiments on the Irrigated Agricultural Dataset

The model comparison results on the irrigated agricultural dataset can be referred to from Table 2. The proposed TRAMNet outperforms other methods on indices of accuracy, kappa, F1-score and recall. However, the best precision belongs to CVA, which achieves the worst recall of 0.6867, on the contrary. As the transformer-based method, ChangeFormer only achieved 0.8736 accuracy. The best competitor among these comparison methods was TFR-PS^{2}ANet. As a deep learning method, it outperformed other deep learning methods without an attention mechanism. Our proposed TRAMNet showed a superior mean value of accuracy and a more stable standard deviation. The second-best method was SVM,

which achieved 0.9614 overall accuracy, and it was weaker than the proposed TRAMNet, which had around 0.02 accuracy. The supervised GETNET achieved 0.9456 overall accuracy. The WCRN achieved the worst performance in terms of overall accuracy, kappa, F1-score, and so on. According to its unsatisfactory precision result, the unchanged pixels can be badly detected by mistake. This could be influenced by the weakness of WCRN in handling complex agricultural scenes.

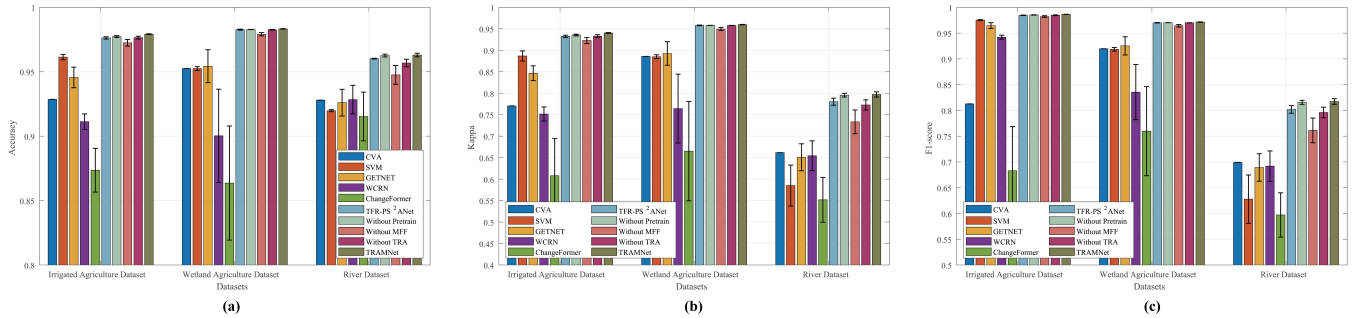


Figure 7. Overall performance comparisons of all methods on three hyperspectral datasets, which are (**a**) accuracy comparison, (**b**) kappa comparison, and (**c**) F1-score comparison.

In the last four rows of Table 2, the comparison results of TRAMNet, TRAMNet without a pre-training step, TRAMNet without an MFF, and TRAMNet without a TRA are shown. Without a pre-training step, the TRAMNet shows a performance descent on the indices of overall accuracy, kappa, F1-score, and recall. The TRAMNet without MFF shows an overall accuracy descent with a large standard deviation. Furthermore, if the TRA blocks are removed from every layer in the feature extractor, a greater performance drop can be observed, which finally comes to 0.9765 accuracy with a large standard deviation as well. And the kappa shows a significant descent from 0.9403 to 0.9329. Although ablation models without pre-training or TRA are inferior to the proposed TRAMNet, they still outperform other state-of-the-art methods, including our previous TFR-PS^{2}ANet, on the irrigated agriculture dataset. In brief, the TRAMNet shows better performance on most metrics than other methods.

Table 2. Model comparison results and module-ablation study results on an irrigated agricultural dataset (repeated 3 times). The best performances are emphasized in bold format.

Models	Acc	Kappa	F1-Score	Precision	Recall
CVA	0.9286	0.7704	0.8127	**0.9953**	0.6867
SVM	0.9614 ± 0.0021	0.8868 ± 0.0117	0.9754 ± 0.0008	0.9657 ± 0.0193	0.9854 ± 0.0184
GETNET	0.9456 ± 0.0080	0.8466 ± 0.0172	0.9646 ± 0.0057	0.9690 ± 0.0091	0.9608 ± 0.0203
WCRN	0.9113 ± 0.0060	0.7516 ± 0.0164	0.9422 ± 0.0039	0.9509 ± 0.0033	0.9336 ± 0.0045
ChangeFormer	0.8736 ± 0.0170	0.6080 ± 0.0863	0.6831 ± 0.0854	0.7874 ± 0.0756	0.6360 ± 0.1579
TFR-PS^{2}ANet	0.9763 ± 0.0009	0.9324 ± 0.0028	0.9846 ± 0.0005	0.9862 ± 0.0029	0.9831 ± 0.0019
Without Pre-training	0.9773 ± 0.0007	0.9356 ± 0.0020	0.9853 ± 0.0004	0.9873 ± 0.0015	0.9833 ± 0.0014
Without MFF	0.9725 ± 0.0026	0.9229 ± 0.0068	0.9821 ± 0.0017	0.9897 ± 0.0014	0.9746 ± 0.0048
Without TRA	0.9765 ± 0.0011	0.9329 ± 0.0034	0.9847 ± 0.0007	0.9860 ± 0.0054	0.9835 ± 0.0056
TRAMNet	**0.9792 ± 0.0003**	**0.9403 ± 0.0010**	**0.9865 ± 0.0002**	0.9862 ± 0.0004	**0.9868 ± 0.0001**

The change-detection maps for the irrigated agricultural dataset are shown in Figure 8. Using a ground truth map as a reference, the proposed TRAMNet shows the most similar visual effects, and the borderline change and farmland change in the circle are easy to distinguish visually. However, the change maps generated by CVA and GETNET failed to detect the borderline change in the river. Although the SVM shows clear overall visual effects, it still misses a number of sporadic and subtle changes, which significantly leads to

terrible kappa and precision. When it comes to the change map of WCRN, the changed area is very cluttered due to visual effects and the borderline of the river almost disappearing, which will inevitably cause awful overall accuracy and kappa indices of the change map. The change map by ChangeFormer also loses most of its detailed information. The TFR-PS^{2}ANet is visually closest to the ground truth map among comparison methods, while the missing data can be observed from the central change in the circle farmland, which leads to it being inferior to our proposed TRAMNet. As for the three ablation methods of TRAMNet without pre-training, without MFF, and without TRA, the main visual difference is sporadic pixel changes on the top left corner of the change map.

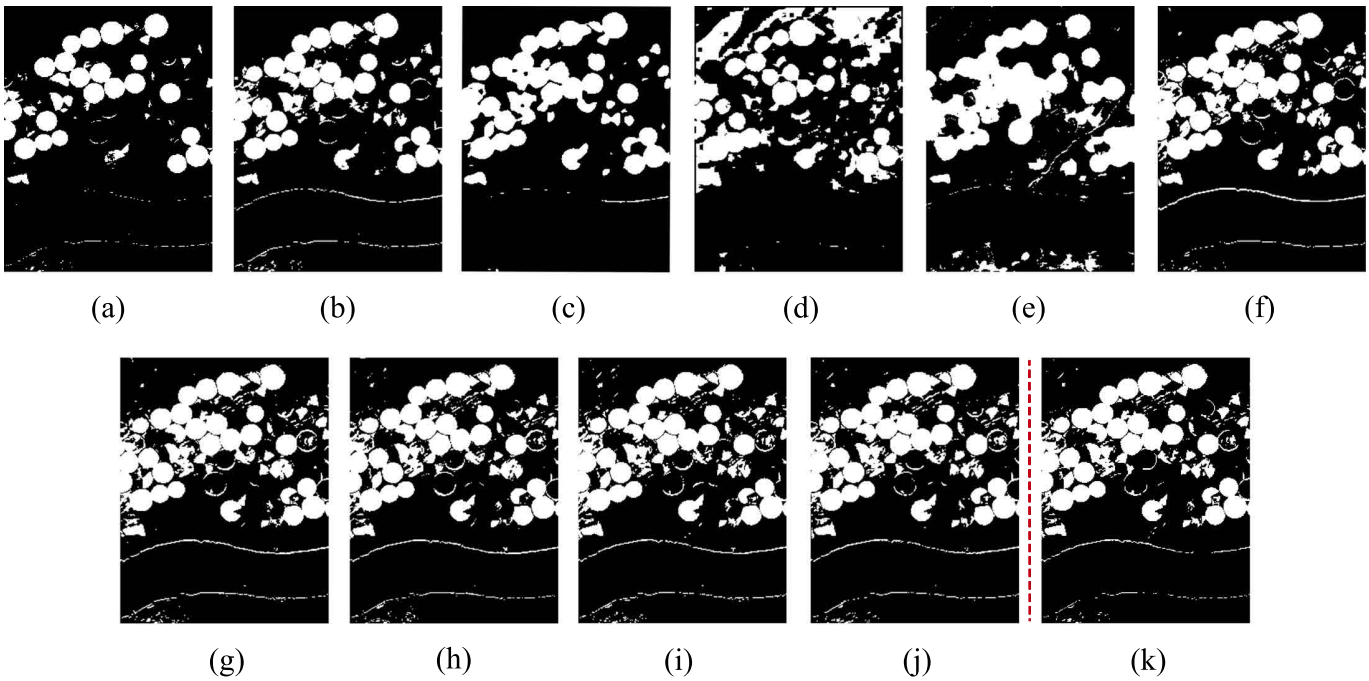


Figure 8. The change map results from different methods on the irrigated agricultural dataset. (**a**) CVA. (**b**) SVM. (**c**) GETNET. (**d**) WCRN. (**e**) ChangeFormer. (**f**) TFR-PS^{2}ANet. (**g**) Without Pre-training. (**h**) Without MFF. (**i**) Without TRA. (**j**) TRAMNet. (**k**) Ground truth.

4.4.2. Experiments on the Wetland Agricultural Dataset

The model comparison results on the wetland agricultural dataset can be found in Table 3. In this dataset, the proposed TRAMNet outperforms other comparison methods on the metrics of overall accuracy, kappa, F1-score, and recall. The best competitor is TFR-PS^{2}ANet, which has a slight descent in overall accuracy compared to the TRAMNet. The best precision was achieved by TFR-PS^{2}ANet, which indicates that the performance on the recall index would be inferior to the TRAMNet. The second-best competitor is GETNET, which achieved 0.9543 overall accuracy, showing an obvious descent of about 0.03 to the TRAMNet, and the large standard deviation indicates the bad model stability. The ChangeFormer achieved the worst accuracy at 0.8636. The performance results of CVA and SVM were close to each other, showing the same accuracy results of 0.9525. The two-domain WCRN only achieves an acceptable accuracy value and kappa results, which are 0.9003 and 0.7643, respectively. This may indicate that WCRN with two-domain learning cannot easily process scenes with a single agricultural land cover change.

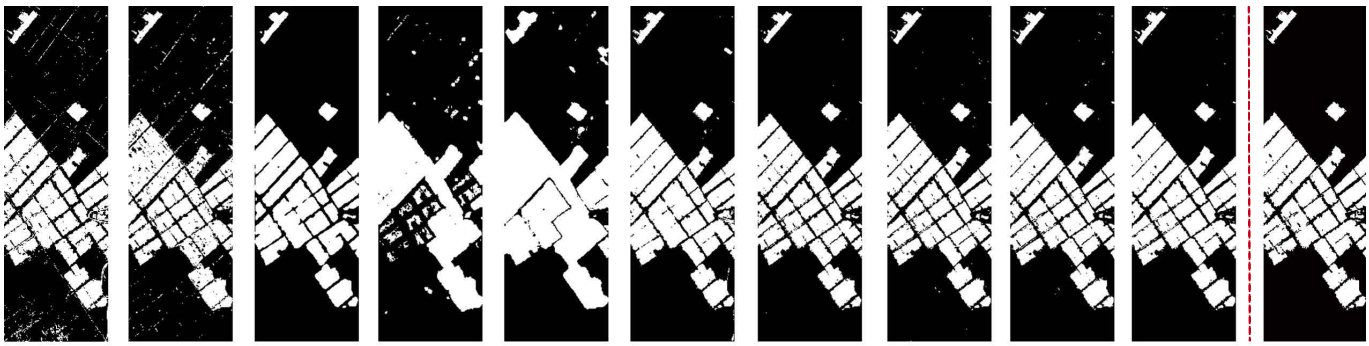
With reference to the last four rows in Table 3, the comparison results of the ablation experiments can be inspected. Compared to the TRAMNet, the ablation model TRAMNet without the pre-training step achieved 0.9828 overall accuracy and a 0.9581 kappa index. The results show its inferiority to the TRAMNet, indicating that the pre-training step can enhance the training effects in a supervised manner. The ablation model without MFF

shows a significant accuracy descent to 0.9791 with a 0.0013 standard deviation. The ablation model without TRA blocks performs worse than TRAMNet without pre-training, and the standard deviation is also larger. This implies that the performance of the feature extractor can be improved by suppressing unrelated information from both spatial and spectral domains. In conclusion, the proposed TRAMNet shows the best performance on the wetland agricultural dataset according to most metrics.

The inference change maps of these methods for the wetland agricultural dataset are shown in Figure 9. The proposed TRAMNet has the closest visual effects to the ground-truth map. Based on the predicted-change map, the subtle changes in the blocks of farmland were successfully distinguished. With reference to the change maps using CVA and SVM, the traditional methods mistakenly detect many borderlines between farmland blocks. As for GETNET, the predicted change map contains overall similarity to the ground truth, while the internal subtle changes in the farmland block are omitted. The two-domain WCRN and ChangeFormer almost lose all the borderline change information in the farmland area, and too many scatters on the top left corner are mistakenly detected. The TFR-PS^{2}ANet is visually closest to the ground-truth map among comparison methods, which reserves the subtle change information. However, the surplus detection on the top-left part burdens the final visual effects. As for the change maps predicted by the three ablation methods, minor differences can be observed in the top-left corner.

Table 3. Model comparison results and module ablation study results on wetland agricultural dataset (repeated 3 times). The best performances are emphasized in bold format.

Models	Acc	Kappa	F1-score	Precision	Recall
CVA	0.9525	0.8859	0.9196	0.9032	0.9366
SVM	0.9525 ± 0.0015	0.8851 ± 0.0045	0.9185 ± 0.0035	0.9150 ± 0.0072	0.9223 ± 0.0144
GETNET	0.9543 ± 0.0128	0.8926 ± 0.0274	0.9253 ± 0.0177	0.8926 ± 0.0548	0.9644 ± 0.0255
WCRN	0.9003 ± 0.0362	0.7643 ± 0.0802	0.8355 ± 0.0535	0.8170 ± 0.0354	0.8579 ± 0.0179
ChangeFormer	0.8636 ± 0.0443	0.6651 ± 0.1157	0.7597 ± 0.0864	0.7646 ± 0.0621	0.7682 ± 0.1530
TFR-PS^{2}ANet	0.9827 ± 0.0004	0.9580 ± 0.0008	0.9701 ± 0.0005	0.9754 ± 0.0081	0.9648 ± 0.0072
Without Pre-training	0.9828 ± 0.0001	0.9581 ± 0.0002	0.9702 ± 0.0001	0.9714 ± 0.0010	0.9691 ± 0.0013
Without MFF	0.9791 ± 0.0013	0.9495 ± 0.0033	0.9643 ± 0.0023	0.9553 ± 0.0077	0.9735 ± 0.0081
Without TRA	0.9826 ± 0.0003	0.9578 ± 0.0004	0.9700 ± 0.0002	**0.9791 ± 0.0048**	0.9609 ± 0.0051
TRAMNet	**0.9833 ± 0.0002**	**0.9598 ± 0.0006**	**0.9714 ± 0.0004**	0.9713 ± 0.0065	**0.9741 ± 0.0023**



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Figure 9. The change map results from different methods on the wetland agricultural dataset. (**a**) CVA. (**b**) SVM. (**c**) GETNET. (**d**) WCRN. (**e**) ChangeFormer. (**f**) TFR-PS^{2}ANet. (**g**) Without Pre-training. (**h**) Without MFF. (**i**) Without TRA. (**j**) TRAMNet. (**k**) Ground truth.

4.4.3. Experiments on the River Dataset

The comparison results of different methods on the river dataset are shown in Table 4. Our proposed TRAMNet achieved the best performance in terms of accuracy, kappa, F1-

score, and precision. Among comparison methods, TFR-PS^{2}ANet is the best competitor to TRAMNet, showing 0.9602 accuracy and 0.7803 kappa index. The second-best competitor is WCRN, which achieved 0.9284 accuracy. This indicates that WCRN can perform well on datasets with random geometry and simple spectral complexity. GETNET shows 0.9260 accuracy and 0.6505 kappa index. The transformer-based ChangeFormer achieved 0.9153 overall accuracy. As the traditional change-detection methods, CVA and SVM had similar performance, achieving 0.928 and 0.9198, respectively. However, the worst performance was obtained by SVM, indicating that traditional machine learning methods cannot effectively extract subtle features between water and land.

With reference to the last four rows in Table 4, the ablation model comparison can be analyzed. If the model is trained without a contrastive pre-training step, the performance on overall accuracy, kappa, F1-score, and precision is slightly weaker than that of TRAMNet, which means the pre-trained model in the current condition can benefit the fine-tuning step to an extent. The ablation model without MFF shows an incredible drop in accuracy to 0.9476, and the standard deviation is also large. When it comes to the model with removed TRA blocks, the performance descent can be clearly observed, which indicates the long-range dependencies can also benefit the feature extraction.

Table 4. Model comparison results and module ablation study results on river dataset (repeated 3 times). The best performances are emphasized in bold format.

Models	Acc	Kappa	F1-score	Precision	Recall
CVA	0.9280	0.6617	0.6992	0.5492	0.9617
SVM	0.9198 ± 0.0008	0.5850 ± 0.0477	0.6278 ± 0.0469	0.5266 ± 0.0092	0.7876 ± 0.0624
GETNET	0.9260 ± 0.0104	0.6508 ± 0.0314	0.6893 ± 0.0267	0.5467 ± 0.0390	0.9368 ± 0.0164
WCRN	0.9284 ± 0.0111	0.6544 ± 0.0346	0.6919 ± 0.0295	0.5579 ± 0.0246	0.9158 ± 0.0175
ChangeFormer	0.9153 ± 0.0189	0.5519 ± 0.0523	0.5971 ± 0.0430	0.5287 ± 0.0908	0.7058 ± 0.0393
TFR-PS^{2}ANet	0.9602 ± 0.0003	0.7803 ± 0.0082	0.8019 ± 0.0074	0.7068 ± 0.0038	0.9267 ± 0.0134
Without Pre-training	0.9626 ± 0.0011	0.7955 ± 0.0044	0.8157 ± 0.0039	0.7147 ± 0.0093	0.9502 ± 0.0064
Without MFF	0.9476 ± 0.0073	0.7333 ± 0.0277	0.7612 ± 0.0240	0.6352 ± 0.0393	0.9529 ± 0.0113
Without TRA	0.9568 ± 0.0029	0.7727 ± 0.0118	0.7960 ± 0.0103	0.6761 ± 0.0182	**0.9682 ± 0.0069**
TRAMNet	**0.9630 ± 0.0014**	**0.7972 ± 0.0061**	**0.8173 ± 0.0054**	**0.7175 ± 0.0096**	0.9495 ± 0.0021

The predicted change maps of different methods on the river dataset are shown in Figure 10. Based on visual effects, the proposed TRAMNet is the closest to the ground truth map. The CVA's change map detects many scatters on the bottom-left corner, which does not correspond with the ground truth. The change map generated by SVM has a smooth area of change, which causes the subtle changes to be missed. Moreover, it stripe noise-like change detection mistakenly appears. GETNET has a similar visual effect, but it mistakenly detects edge information. The results of WCRN show that it detects change areas in the plaque shape, which leads to the detailed information being lost. ChangeFormer loses even more detailed change information. TFR-PS^{2}ANet has very similar visual effects to the ground truth, which loses little information on subtle changes compared to TRAMNet. As for the change maps predicted using three ablation methods, minor differences can be observed from the river course area. In brief, our proposed TRAMNet can achieve the best performance on both comparison models and ablation models.

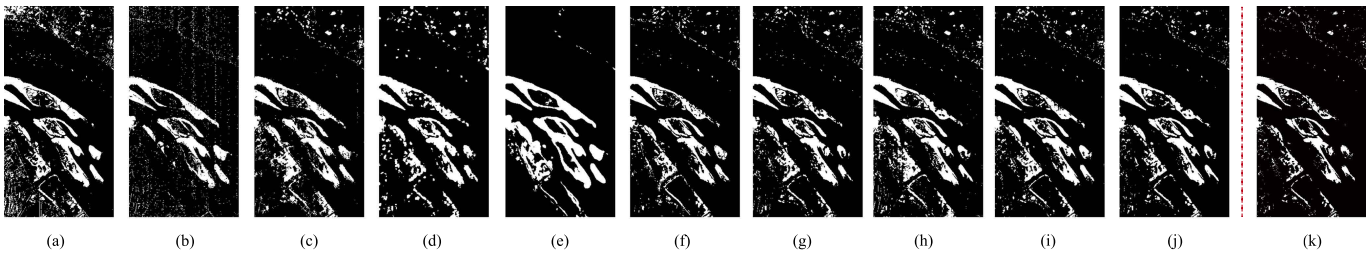


Figure 10. The change map results from different methods on the river dataset. (**a**) CVA. (**b**) SVM. (**c**) GETNET. (**d**) WCRN. (**e**) ChangeFormer. (**f**) TFR-PS^{2}ANet. (**g**) Without Pre-training. (**h**) Without MFF. (**i**) Without TRA. (**j**) TRAMNet. (**k**) Ground truth.

4.5. The Ratio of Training Samples

The main purposes of contrastive learning are to improve the weight initializations using first-step contrastive learning and to reduce the training samples of supervised fine-tuning. Compared to the benchmark training rate of every dataset, which are 9.7%, 9.7%, and 4.03%, the training rate analysis is based on the product of the ratio and benchmark training samples, for example, $0.4 \times 9.7\%$. The impact of the ratio of training samples is shown in Figure 11.

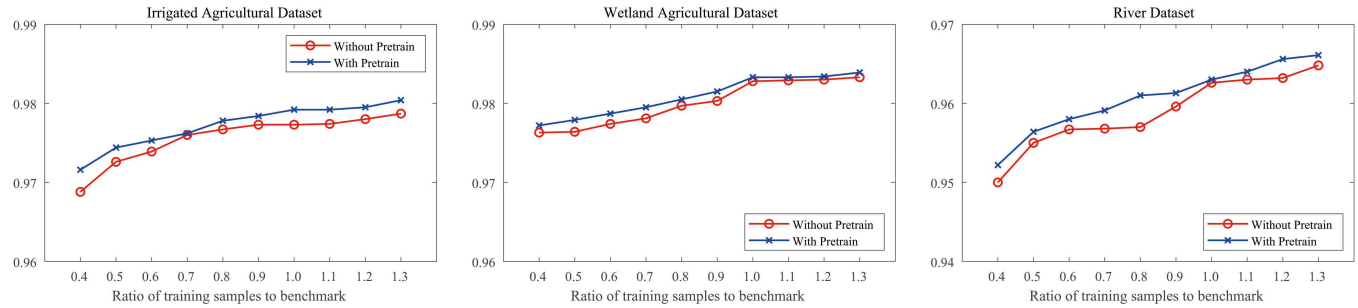


Figure 11. The impact of the training rate on three public datasets. The red line is the overall accuracy of TRAMNet without the pre-training step. The blue line is the overall accuracy with the pre-training step.

As Figure 11 shows, the TRAMNet with the pre-training step showed better performance than the TRAMNet without pre-training, which indicates that contrastive learning with SCLF can effectively improve the weight initialization for fine-tuning. On the irrigated agricultural dataset, the model can perform with over 0.97 accuracy, which is already superior to most state-of-the-art comparison models. Along with the increasing training rate, the overall accuracy can reach almost 0.98 at a rate of 1.3. On the wetland agriculture dataset, the accuracy performance is over 0.975 at a rate of 0.4, which is superior to most comparison methods as well. From the rate of 1.0 to the rate of 1.3, the performance comes to a bottleneck due to visual effects. On the river dataset, the TRAMNet can achieve over 0.95 accuracy at a rate of 0.4, which is superior to most comparison models. When the training rate comes to the rate of 1.3, the overall accuracy can increase to about 0.965, which is largely improved. In brief, self-supervised learning with the soft contrastive loss function can effectively improve the model performance.

4.6. Hyperparameter Analysis

The hyperparameters for the soft contrastive loss function are the similarity temperature τ and the soft coefficient λ, which can determine the discrimination strength of pseudo-positive training samples and negative training samples. The impact of both hyperparameters is shown in Figure 12.

The similarity temperature τ is used to scale the distance measurement of positive sample pairs and negative sample pairs. According to the accuracy performance from the

chosen τ values, the accuracy increases first and achieves the best performance at $\tau = 0.1$ on three public hyperspectral change-detection datasets. Therefore, the value of τ is set as 0.1 by default. When it comes to the analysis of the soft coefficient λ, it is used to directly soften the final loss value obtained from a batch of similarity computation. The accuracy achieves a slow increase with λ getting smaller on the irrigated and wetland agricultural datasets, and it eventually achieves the best accuracy performance at $\lambda = 1 \times 10^{-3}$. The trend is more obvious on the river dataset; the accuracy performance shows a drastic increase from $\lambda = 1$ to $\lambda = 1 \times 10^{-3}$, and it achieves the best accuracy at $\lambda = 1 \times 10^{-3}$ as well. Therefore, the value of λ is set as 1×10^{-3} by default.

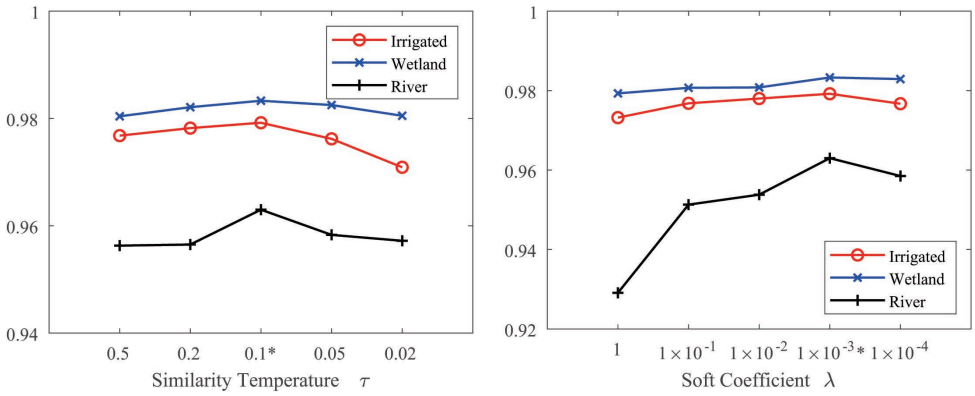


Figure 12. The impact on accuracy due to hyperparameters of a similarity temperature τ and the soft coefficient λ on three public datasets. The asterisk (*) indicates the corresponding hyperparameter value for the best accuracy performance.

5. Discussion

The advantage in terms of the accuracy of the proposed TRAMNet method over the benchmark methods is due mainly to the application of the contrastive training strategy and the deep learning model with a residual spatial–spectral attention mechanism. Moreover, the proposed TRAMNet method is assisted by MFF, which can fuse the feature map with different resolutions. As the comparison results on the three datasets show, the traditional CVA performs stably due to its unsupervised threshold classification. However, the threshold only considers numerical difference and cannot utilize the abundant inner spectral and spatial information, which leads it to perform worse than our proposed TRAMNet. SVM is the traditional supervised method, which needs training samples. In practice, it can be difficult for SVM to fully utilize the local spatial information while finding a proper discriminative property for spectral change. On the contrary, our proposed method can effectively utilize both spatial and spectral information simultaneously with the TRA block.

With reference to the comparison of deep learning methods, our proposed method performs better than the current state-of-the-art method. GETNET is based on the supervised LSConvolution architecture. It can only extract local spatial features from input patches but cannot obtain dependencies from internal spectral and spatial information. ChangeFormer is a Transformer-based method. But its performance is inferior to our proposed method and other deep learning methods. And both WCRN and ChangeFormer show similar visual effects. The reason for this result could be that both WCRN and ChangeFormer are not initially designed for hyperspectral change detection, which inevitably makes it difficult for it to handle abundant and intensive spectral information. TFR-PS^{2}ANet is a method that combines a convolutional encoder–decoder and attention module. It performs better than most deep learning methods, but it is still inferior to our proposed TRAMNet, due to the insufficient global spatial–spectral dependencies.

From the ablation analysis, it can be found that the contrastive self-supervised learning framework can effectively improve the initialization of the model weight. With a reasonable weight-initialization strategy, the proposed model can achieve better detection results with

fewer labeled training samples. In practice, the labels for changed areas are hard to obtain due to the time-consuming ground investigation and visual comparison among dual-temporal hyperspectral images. Our proposed method can obviously decrease the demand for training samples, which can free up time and labor costs.

However, there are also some limitations of the proposed method. First, our proposed method uses different images as training patches in both stage 1 and stage 2. However, remote sensing change detection basically has a time interval, which indicates that the relationship between different times is not considered. To effectively utilize the time features and satisfy future time-series applications, the temporal attention mechanism should be considered and developed. Moreover, the contrastive self-supervised training strategy intends to obtain inner features from image patches, which indicates that the spatial resemblance is still restricted by the input patch size. The pre-trained weights cannot be directly used to make the discrimination of areas of change, and it still needs some training samples to adjust the model to have a change-detection ability. To overcome these issues, in a further study, the similarity or dependencies between patches should be considered. And the pre-text task design can be updated to a pseudo-training-sample strategy and converted to other semi-supervised learning strategies. In our experiments, the method suffered from random sampling due to the serious class imbalance at the beginning. Therefore, the stratified sampling for fine-tuning is applied to generate a few training samples.

There are threats to the internal and external validity of the experiments. (1) The internal validity of the experiments could be influenced by random training samples and multiple module effectiveness. Therefore, we repeated our experiments with random training samples three times to mitigate the chance that it could lead to underestimating or overestimating the model's performance. As for the module's effectiveness, to avoid alternative explanations, we manipulated an independent variable in our study to verify that every module was working normally. (2) The proposed method is expected to be extended to other hyperspectral applications. However, the external validity of the experiments could be influenced by sample features with a single land-cover type. To handle this issue, our contrastive self-supervised learning framework was optimized on three hyperspectral change-detection datasets, which have different time intervals, different coverage areas, different texture information, and different band numbers, to mitigate the threat to external validity.

6. Conclusions

In this article, a deep, contrastive, self-supervised two-domain residual attention network (TRAMNet) is proposed for hyperspectral change detection. First, the hyperspectral image patches were preprocessed with data normalization, and then the patches from two different times were augmented with the proposed random augmentation pool, which contains random Gaussian noise, the rotation operation, and the flip operation. The differencing patches of original patches and augmented patches are sent to the same encoder to learn the feature representation. The projection head is applied to transform the feature representation into feature vectors, which are used to compute the soft contrastive loss among positive sample pairs and negative sample pairs. After the weights of the encoder component are optimized, the pre-trained model is saved for the next fine-tuning to train the change-detection model. The prediction procedure is exerted with the fine-tuning model.

We implemented our algorithm and performed experiments on three public hyperspectral change-detection datasets. Both the visual and quantitative results have shown that the proposed method outperforms most state-of-the-art methods.

When dealing with different change-detection tasks, the TRAMNet can be regarded as a benchmark model with an attention mechanism for other self-supervised hyperspectral change-detection methods. In the future, a pre-text tasks-based supervised model will be considered to decrease the need for training samples and improve the change-detection performance.

Author Contributions: Conceptualization, Y.H.; methodology, Y.H.; software, Y.H.; validation, Y.H.; writing—Y.H.; writing—review and editing, W.Q., C.H. and R.S.; funding acquisition, L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Key Research and Development Projects (grant no. 2022YFF0904400), the National Natural Science Foundation of China (grant no. 41830108, grant no. 42201392), and the China Postdoctoral Science Foundation (grant no. 2022M723222).

Data Availability Statement: The data presented in this study are openly available in Remote Sensing Datasets at https://rslab.ut.ac.ir/data (accessed on 1 June 2021) and River Dataset at http://crabwq.github.io (accessed on 1 June 2021).

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

References

1. Ghamisi, P.; Yokoya, N.; Li, J.; Liao, W.; Liu, S.; Plaza, J.; Rasti, B.; Plaza, A. Advances in Hyperspectral Image and Signal Processing: A Comprehensive Overview of the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 37–78. [CrossRef]
2. Liu, S.; Marinelli, D.; Bruzzone, L.; Bovolo, F. A Review of Change Detection in Multitemporal Hyperspectral Images: Current Techniques, Applications, and Challenges. *IEEE Geosci. Remote Sens. Mag.* **2019**, *7*, 140–158. [CrossRef]
3. Du, P.; Liu, S.; Bruzzone, L.; Bovolo, F. Target-Driven Change Detection Based on Data Transformation and Similarity Measures. In Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, 22–27 July 2012; pp. 2016–2019.
4. Xiao, P.; Sheng, G.; Zhang, X.; Liu, H.; Guo, R. Direction-Dominated Change Vector Analysis for Forest Change Detection. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102492. [CrossRef]
5. Washaya, P.; Balz, T.; Mohamadi, B. Coherence Change-Detection with Sentinel-1 for Natural and Anthropogenic Disaster Monitoring in Urban Areas. *Remote Sens.* **2018**, *10*, 1026. [CrossRef]
6. Liu, S.; Bruzzone, L.; Bovolo, F.; Du, P. Hierarchical Unsupervised Change Detection in Multitemporal Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 244–260. [CrossRef]
7. Bovolo, F.; Marchesi, S.; Bruzzone, L. A Framework for Automatic and Unsupervised Detection of Multiple Changes in Multitemporal Images. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 2196–2212. [CrossRef]
8. Seydi, S.T.; Hasanlou, M. A New Land-Cover Match-Based Change Detection for Hyperspectral Imagery. *Eur. J. Remote Sens.* **2017**, *50*, 517–533. [CrossRef]
9. Ortiz-Rivera, V.; Vélez-Reyes, M.; Roysam, B. Change Detection in Hyperspectral Imagery Using Temporal Principal Components. In *Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2006, Proceedings of the Defense and Security Symposium, Orlando (Kissimmee), FL, USA, 17–21 April 2006*; Shen, S.S., Lewis, P.E., Eds.; SPIE: Bellingham, WA, USA, 2006; p. 623312.
10. Hou, Z.; Li, W.; Li, L.; Tao, R.; Du, Q. Hyperspectral Change Detection Based on Multiple Morphological Profiles. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 21526482
11. Liu, S.; Du, Q.; Tong, X. Band Selection for Change Detection from Hyperspectral Images. In *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXIII, Proceedings of the SPIE DEFENSE + SECURITY, Anaheim, CA, USA, 9–13 April 2017*; Velez-Reyes, M., Messinger, D.W., Eds.; SPIE: Bellingham, WA, USA, 2017; p. 101980T.
12. Liu, S.; Du, Q.; Tong, X.; Samat, A.; Pan, H.; Ma, X. Band Selection-Based Dimensionality Reduction for Change Detection in Multi-Temporal Hyperspectral Images. *Remote Sens.* **2017**, *9*, 1008. [CrossRef]
13. Bruzzone, L.; Prieto, D.F.; Serpico, S.B. A Neural-Statistical Approach to Multitemporal and Multisource Remote-Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **1999**, *37*, 1350–1359. [CrossRef]
14. Bruzzone, L.; Serpico, S.B. An Iterative Technique for the Detection of Land-Cover Transitions in Multitemporal Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **1997**, *35*, 858–867. [CrossRef]
15. Song, B.; Tang, Y.; Zhan, T.; Wu, Z. BRCN-ERN: A Bidirectional Reconstruction Coding Network and Enhanced Residual Network for Hyperspectral Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5510105. [CrossRef]
16. Wang, Q.; Yuan, Z.; Du, Q.; Li, X. GETNET: A General End-to-End 2-D CNN Framework for Hyperspectral Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3–13. [CrossRef]
17. Song, A.; Choi, J.; Han, Y.; Kim, Y. Change Detection in Hyperspectral Images Using Recurrent 3D Fully Convolutional Networks. *Remote Sens.* **2018**, *10*, 1827. [CrossRef]
18. Zhao, C.; Cheng, H.; Feng, S. A Spectral–Spatial Change Detection Method Based on Simplified 3-D Convolutional Autoencoder for Multitemporal Hyperspectral Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 5507705. [CrossRef]
19. Zhan, T.; Song, B.; Sun, L.; Jia, X.; Wan, M.; Yang, G.; Wu, Z. TDSSC: A Three-Directions Spectral–Spatial Convolution Neural Network for Hyperspectral Image Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 377–388. [CrossRef]
20. Moustafa, M.S.; Mohamed, S.A.; Ahmed, S.; Nasr, A.H. Hyperspectral Change Detection Based on Modification of UNet Neural Networks. *J. Appl. Remote Sens.* **2021**, *15*, 028505. [CrossRef]

21. Huang, Y.; Zhang, L.; Huang, C.; Qi, W.; Song, R. Parallel Spectral–Spatial Attention Network with Feature Redistribution Loss for Hyperspectral Change Detection. *Remote Sens.* **2022**, *15*, 246. [CrossRef]
22. Lei, J.; Li, M.; Xie, W.; Li, Y.; Jia, X. Spectral Mapping with Adversarial Learning for Unsupervised Hyperspectral Change Detection. *Neurocomputing* **2021**, *465*, 71–83. [CrossRef]
23. Ou, X.; Liu, L.; Tan, S.; Zhang, G.; Li, W.; Tu, B. A Hyperspectral Image Change Detection Framework With Self-Supervised Contrastive Learning Pretrained Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 7724–7740. [CrossRef]
24. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A Survey on Contrastive Self-Supervised Learning. *Technologies* **2020**, *9*, 2. [CrossRef]
25. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *Proc. Mach. Learn. Res.* **2020**, *119*, 1597–1607.
26. Zhao, L.; Luo, W.; Liao, Q.; Chen, S.; Wu, J. Hyperspectral Image Classification With Contrastive Self-Supervised Learning Under Limited Labeled Samples. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6008205. [CrossRef]
27. Cao, Z.; Li, X.; Feng, Y.; Chen, S.; Xia, C.; Zhao, L. ContrastNet: Unsupervised Feature Learning by Autoencoder and Prototypical Contrastive Learning for Hyperspectral Imagery Classification. *Neurocomputing* **2021**, *460*, 71–83. [CrossRef]
28. Guan, P.; Lam, E.Y. Cross-Domain Contrastive Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5528913. [CrossRef]
29. Hang, R.; Qian, X.; Liu, Q. Cross-Modality Contrastive Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5532812. [CrossRef]
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
31. Hu, X.; Li, T.; Zhou, T.; Liu, Y.; Peng, Y. Contrastive Learning Based on Transformer for Hyperspectral Image Classification. *Appl. Sci.* **2021**, *11*, 8670. [CrossRef]
32. Huang, X.; Dong, M.; Li, J.; Guo, X. A 3-D-Swin Transformer-Based Hierarchical Contrastive Learning Method for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5411415. [CrossRef]
33. Wang, L.; Wang, L.; Wang, Q.; Atkinson, P.M. SSA-SiamNet: Spectral–Spatial-Wise Attention-Based Siamese Network for Hyperspectral Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 18. [CrossRef]
34. Gong, M.; Jiang, F.; Qin, A.K.; Liu, T.; Zhan, T.; Lu, D.; Zheng, H.; Zhang, M. A Spectral and Spatial Attention Network for Change Detection in Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5521614. [CrossRef]
35. Hou, S.; Shi, H.; Cao, X.; Zhang, X.; Jiao, L. Hyperspectral Imagery Classification Based on Contrastive Learning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5521213. [CrossRef]
36. Hasanlou, M.; Seydi, S.T. Hyperspectral Change Detection: An Experimental Comparative Study. *Int. J. Remote Sens.* **2018**, *39*, 7029–7083. [CrossRef]
37. Bovolo, F.; Bruzzone, L. A Theoretical Framework for Unsupervised Change Detection Based on Change Vector Analysis in the Polar Domain. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 218–236. [CrossRef]
38. Nemmour, H.; Chibani, Y. Multiple Support Vector Machines for Land Cover Change Detection: An Application for Mapping Urban Extensions. *ISPRS J. Photogramm. Remote Sens.* **2006**, *61*, 125–133. [CrossRef]
39. Qu, X.; Gao, F.; Dong, J.; Du, Q.; Li, H.-C. Change Detection in Synthetic Aperture Radar Images Using a Dual-Domain Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4013405. [CrossRef]
40. Bandara, W.G.C.; Patel, V.M. A Transformer-Based Siamese Network for Change Detection. *arXiv* **2022**, arXiv:2201.01293.