remote sensing

MDPI

Article

Change Detection Based on Existing Vector Polygons and Up-to-Date Images Using an Attention-Based Multi-Scale ConvTransformer Network

Shengli Wang [1,2], Yihu Zhu [2,3], Nanshan Zheng [1,*], Wei Liu [3], Hua Zhang [1], Xu Zhao [4] and Yongkun Liu [5]

[1] School of Environment and Spatial Informatics, China University of Mining and Technology, Xuzhou 221116, China; shengliwang@cumt.edu.cn (S.W.); zhhuacumt@cumt.edu.cn (H.Z.)
[2] Jiangsu Geologic Surveying and Mapping Institute, Nanjing 211102, China; 2020150184@jsnu.edu.cn
[3] School of Geography, Geomatics and Planning, Jiangsu Normal University, Xuzhou 221116, China; liuw@jsnu.edu.cn
[4] School of Earth Science and Engineering, Hohai University, Nanjing 211100, China; zx0065@hhu.edu.cn
[5] Aerospace Information Research Institute (AIR), Chinese Academy of Sciences (International Research Center of Big Data for Sustainable Development Goals), Beijing 100094, China; liuyongkun19@mails.ucas.ac.cn
* Correspondence: znshcumt@cumt.edu.cn; Tel.: +86-159-5216-7986

Abstract: Vector polygons represent crucial survey data, serving as a cornerstone of national geographic censuses and forming essential data sources for detecting geographical changes. The timely update of these polygons is vital for governmental decision making and various industrial applications. However, the manual intervention required to update existing vector polygons using up-to-date high-resolution remote sensing (RS) images poses significant challenges and incurs substantial costs. To address this, we propose a novel change detection (CD) method for land cover vector polygons leveraging high-resolution RS images and deep learning techniques. Our approach begins by employing the boundary-preserved masking Simple Linear Iterative Clustering (SLIC) algorithm to segment RS images. Subsequently, an adaptive cropping approach automatically generates an initial sample set, followed by denoising using the efficient Visual Transformer and Class-Constrained Density Peak-Based (EViTCC-DP) method, resulting in a refined training set. Finally, an enhanced attention-based multi-scale ConvTransformer network (AMCT-Net) conducts fine-grained scene classification, integrating change rules and post-processing methods to identify changed vector polygons. Notably, our method stands out by employing an unsupervised approach to denoise the sample set, effectively transforming noisy samples into representative ones without requiring manual labeling, thus ensuring high automation. Experimental results on real datasets demonstrate significant improvements in model accuracy, with accuracy and recall rates reaching 92.08% and 91.34%, respectively, for the Nantong dataset, and 93.51% and 92.92%, respectively, for the Guantan dataset. Moreover, our approach shows great potential in updating existing vector data while effectively mitigating the high costs associated with acquiring training samples.

Keywords: change detection (CD); deep learning; fine-grained scene classification; vector polygons; high-resolution remote sensing (RS) image

check for updates

Citation: Wang, S.; Zhu, Y.; Zheng, N.; Liu, W.; Zhang, H.; Zhao, X.; Liu, Y. Change Detection Based on Existing Vector Polygons and Up-to-Date Images Using an Attention-Based Multi-Scale ConvTransformer Network. *Remote Sens.* **2024**, *16*, 1736. https://doi.org/10.3390/rs16101736

Academic Editors: Naoto Yokoya, Chen Wu, Wen Yang and Pengyuan Lv

Received: 7 April 2024
Revised: 10 May 2024
Accepted: 13 May 2024
Published: 14 May 2024

1. Introduction

Land cover change detection plays a crucial role in understanding dynamics on the Earth's surface, with it being indispensable for applications such as land use analysis, environmental assessment, monitoring of human development, and disaster response [1–5]. The increasing availability of high-resolution remote sensing (RS) images has revolutionized the field of change detection, enabling detailed long-term monitoring of land cover dynamics [6]. The combination of these images with advancements in deep learning techniques offers unprecedented opportunities for automating the change detection process, thereby facilitating the timely updating of critical geographic datasets. However, despite these advancements, current methods used for change detection still face several challenges. The processing involved in updating existing vector polygons with new high-resolution RS images is time-consuming and labor-intensive, with manual intervention remaining a bottleneck, leading to inefficiencies and high costs. Furthermore, traditional methods for change detection through image differencing often struggle to handle noisy data and require large quantities of high-quality training samples, limiting their scalability and applicability in real-world scenarios [7–10]. Therefore, there is an urgent need for innovative approaches to overcome these limitations and enhance the efficiency and accuracy of change detection in high-resolution RS images.

To address the challenge of difficult sample acquisition, Zhao et al. [11] utilized linear spectrum hybrid analysis and spectral index method to extract ground object samples in an attempt to tackle sample selection difficulties in deep learning classification. Cui et al. [12] determined the optimal focus radius for different land types based on focus statistics and unique phenological characteristics, and subsequently proposed an approach for the automatic generation of training samples using an enhanced distance measure. Cao et al. [13] introduced a comprehensive fused cross-task transfer learning method (FFCTL) that effectively utilizes crowdsourced building data and high-resolution satellite images to address the issue of expensive real samples in building change detection. Lv et al. [7] presented an iterative training sample augmentation (ITSA) strategy combined with a deep learning neural network to enhance change detection performance. Li et al. [14] developed a label-noise active learning sample collection method for multi-temporal land cover classification. However, these studies did not consider the uncertainty inherent in the training data, which may introduce errors into the final results. Therefore, addressing the denoising problem associated with training samples is crucial for accurate change detection [15].

In addition, feature extraction plays a crucial role in object-based change detection. Deep learning is widely employed in RS image change detection due to its robust feature extraction and modeling capabilities. For instance, Zhang et al. [16] proposed a deep learning change detection framework to detect newly built buildings by paying more attention to the overall features and contextual associations of the change object instances. Gu et al. [17] introduced a multi-scale convolutional layer feature fusion network to achieve high-precision image change detection by addressing pseudo-changes and reducing the loss of details in the detection process. Despite the excellent performance of Convolutional Neural Networks (CNNs) in extracting relevant multi-scale features from images, they have limitations in establishing long-range dependencies of self-attention within images. The Transformer model has made significant advancements in image recognition and computer vision due to its efficient processing of contextual information and global modeling ability [18–22]. Recent studies have applied Transformer-based architectures to RS tasks, such as multi-modal image processing [23] and scene classification, which is carried out utilizing a Vision Transformer (ViT) [24].

Despite the enhanced perspectives ViT offers for image modeling, it frequently faces challenges related to high computational complexity and memory usage. Addressing this concern, Chen et al. [20] devised an efficient dual-path Transformer architecture. This innovative approach has led to achieving state-of-the-art accuracy on benchmark building extraction datasets. Furthermore, integrating the strengths of CNNs and Transformers to enhance change detection capability has become a current research focus and has yielded significant results [25–28]. While these advanced deep learning methods hold significant promise in image analysis, their application in change detection encounters challenges due to the inherent complexity of high-resolution RS images. These challenges include, but are not limited to, diverse ground objects, image noise, and seasonal changes. Specifically, difficulties arise in high label cost, accurate detection of multi-scale ground objects, data imbalances, and the substantial demand for computing resources. Consequently, many current methods for multi-temporal image change detection, which are based on deep learning, encounter difficulties in direct applicability to practical tasks. While recent developments in RS foundation models, such as SAM-enhanced [29], RingMo [30], and RSPrompter [31], have exhibited excellent generalization ability and zero-shot learning capability, they still confront challenges related to reducing human intervention, decreasing computing resource requirements, and accommodating diverse downstream task applications [21,32].

The effective utilization of prior knowledge, such as vector data, is crucial for change detection. However, currently, few studies integrate vector and image data for change detection methods. Vector data contain boundary and categorical information of ground objects, with a large number of valuable vector polygons amassed by land survey projects and other endeavors. Unlike image data, it offers essential support for image segmentation, sample annotation, and classification tasks. Zhang et al. [33] proposed a change detection method based on vector data and the isolated forest algorithm. This method utilizes vector constraints with category information from an old time phase to finely segment a new time-phase RS image, obtaining the plot, and applies the isolated forest to calculate the change index of the plot. However, its efficacy hinges on the assumption that the proportion of various ground object change plots is small, and it imposes high computational demands. Wei et al. [34] replaced the historical RS image with vector data, and introduced a new method to detect changes from single-phase RS image and vector data using the outlier index of texture feature space. However, the texture-based approach limits the widespread adoption of the method.

The automatic generation of labeled samples based on vector boundary constraints has garnered significant attention from scholars in recent years, holding considerable promise for application in vector data-aided change detection within the field [35–37]. However, due to registration errors, semantic gaps among ground objects, land cover changes, variations in annotation personnel, and other factors, label noise inevitably exists when directly utilizing vector attributes for sample annotation within RS datasets. Consequently, this leads to an incomplete capture of the image's semantic content by the training model and a subsequent decline in generated feature discrimination ability [35,38–42]. Currently, denoising methods for samples have been initially developed. Li et al. [43] employed superpixel segmentation on the image to delineate image objects and extracted spectral and texture features. They then analyzed the distribution of homogeneity characteristic values within cultivated land and applied a box plot anomaly detection method to eliminate noisy samples, which was employed in monitoring the non-agriculturalization of cultivated land. Kang et al. [44] utilized an energy constraint minimization criterion for hyperspectral image target detection and effectively corrected training samples. However, these studies only considered abnormal removal from pixel values and texture features and did not integrate the actual needs of change detection. In summary, the theory and technology behind the change detection methods integrating vector data and images are still immature, and the automatic generation of samples and efficient unsupervised denoising methods are lacking, requiring further research.

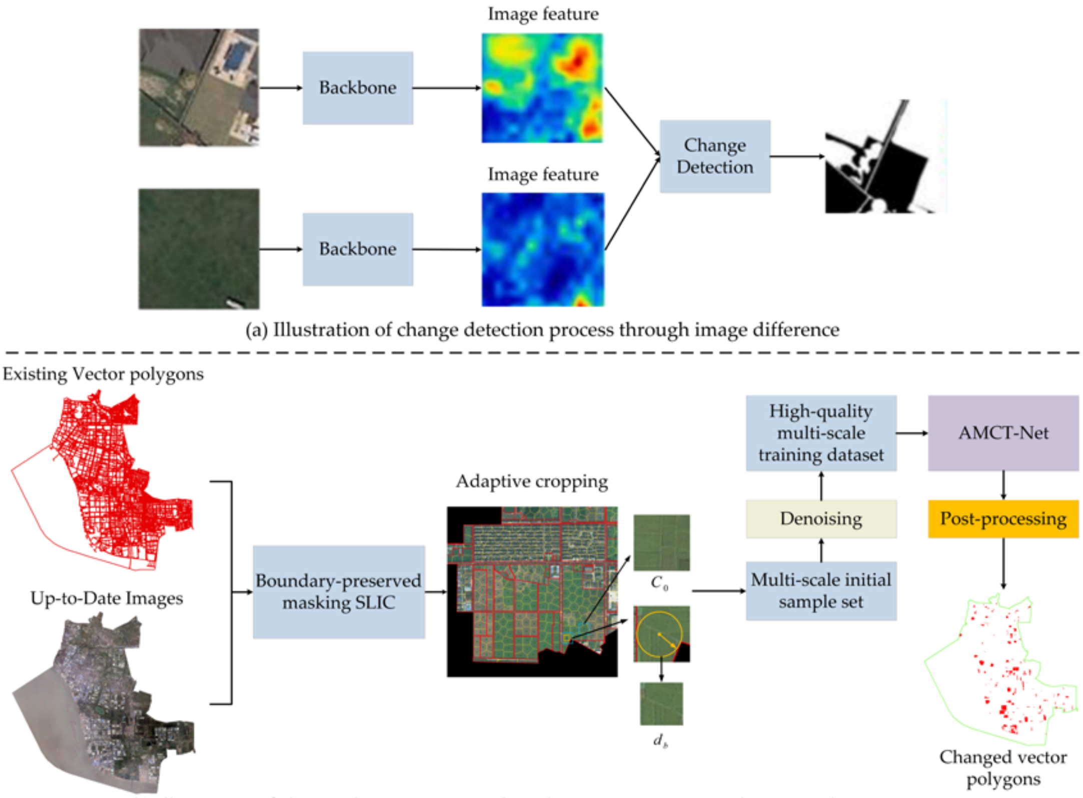
To address the aforementioned issues, this paper proposes a novel change detection method for detecting changed land cover vector polygons using high-resolution RS images and deep learning. Figure 1 shows the differences between our method and conventional change detection methods.

The key contributions of this work are delineated as follows:

- We introduce a framework for detecting changes in vector polygons utilizing single-temporal high-resolution RS images and deep learning. This framework enables end-to-end application, encompassing image preprocessing through change detection, requiring solely up-to-date images and corresponding land cover vector data from the previous time image. This method offers a comprehensive bottom-up solution.
- For sample construction, we propose boundary-preserved masking Simple Linear Iterative Clustering (SLIC) for generating superpixels. These are then combined with land cover vector data to create an adaptive sample cropping scheme. To address noise, we introduce an efficient Visual Transformer and class-constrained Density Peak-based (EViTCC-DP) method for noisy label removal, followed by the transformation of noisy

samples into representative ones using k-means clustering, resulting in the automatic generation of a high-quality multi-scale sample set.

- To enhance fine-grained scene classification precision, we employ an improved attention-based multi-scale ConvTransformer network (AMCT-Net) for superpixel cropping unit classification. By integrating a CNN structure and Transformer, along with the attention mechanism module, we achieve a more discriminative feature representation, enhancing the model's classification accuracy. Additionally, we introduce a change decisionmaker with various rules, which synergistically combines and post-processes sample predictions with land cover vector data to effectively extract changed vector polygons.



(a) Illustration of change detection process through image difference

(b) Illustration of change detection process based on existing vector polygons and images (ours)

Figure 1. The comparison of our method with other approaches. (**a**) Illustration of change detection process through image difference. (**b**) Illustration of change detection process based on existing vector polygons and images. Currently, most change detection research adopts image differencing methods.

The work is organized as follows. Section 2 introduces the proposed methodology in detail, consisting of five main parts: Density Peak (DP) clustering, automated generation of initial samples with vector boundary constraints, initial samples denoising based on DP clustering algorithm, AMCT-Net, and vector polygons change detection analysis based on confidence rules. Section 3 provides details of the datasets, describes the experiments, and analyzes the results. Section 4 presents the discussion. Finally, Section 5 concludes this paper.

2. Methodology

Most existing change detection algorithms rely heavily on image contrast and require high-quality training samples, which are often challenging to acquire. This limitation impedes their effectiveness in large-scale practical change detection scenarios. In this work, we propose a novel framework for detecting changes in vector polygons using high-resolution RS images, as illustrated in Figure 2. Our framework comprises an end-to-end processing architecture consisting of four primary stages. Initially, we employ a boundary-preserved masking SLIC [45] algorithm for image superpixel segmentation, leveraging high-resolution RS images and land cover vector data as inputs. Subsequently, adaptive cropping is performed on each superpixel unit to generate initial samples, which undergo denoising via the EViTCC-DP method, and the inclusion of representative training samples (RTS) is introduced to construct the final high-quality training sample set. These refined samples are then subjected to fine-grained image classification through an improved attention-based multi-scale ConvTransformer network (AMCT-Net). Finally, a change decisionmaker applies various rules to detect and locate changes within vector polygons. A detailed elaboration of each module will be provided subsequently.

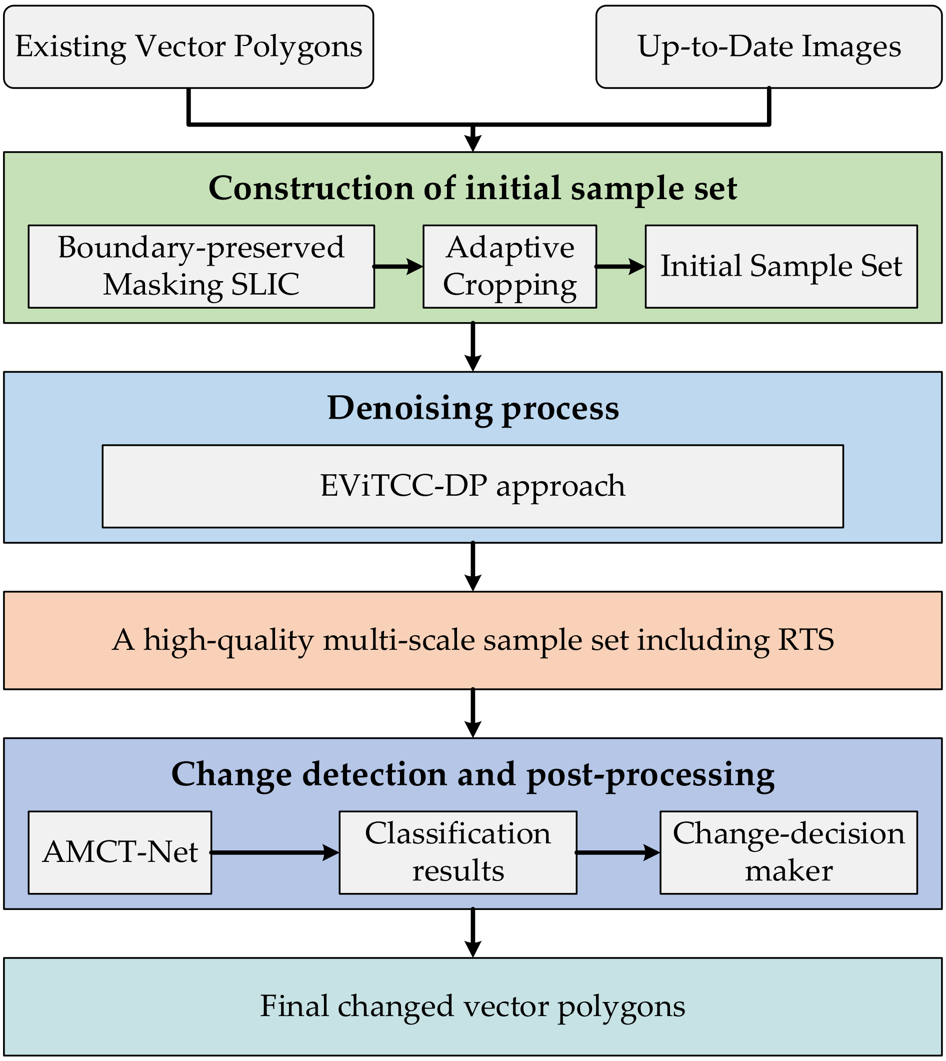

Existing Vector Polygons
Up-to-Date Images
Construction of initial sample set
Boundary-preserved Masking SLIC
Adaptive Cropping
Initial Sample Set
Denoising process
EViTCC-DP approach
A high-quality multi-scale sample set including RTS
Change detection and post-processing
AMCT-Net
Classification results
Change-decision maker
Final changed vector polygons

Figure 2. The overall workflow of the proposed methodology.

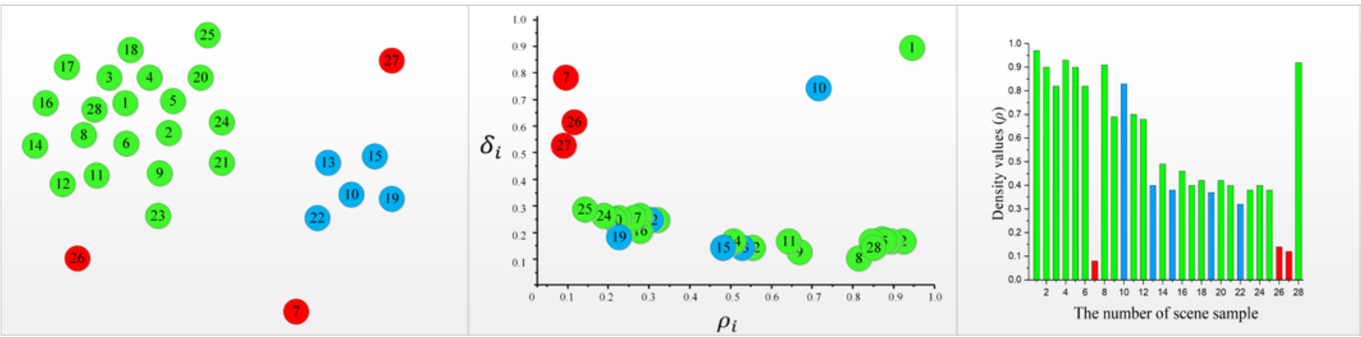
2.1. Density Peak Clustering

In contrast to the k-means algorithm [46,47], which necessitates the specification of the cluster number, the Density Peak (DP) algorithm [48] can identify cluster centers without iterative procedures. Operating non-iteratively, it adheres to a straightforward principle and has showcased outstanding clustering efficacy in handling both spherical clusters and non-convex datasets. The fundamental principles are as follows:

1. The density around the cluster center should be relatively high;
2. The cluster center should be situated at a considerable distance from points with higher surrounding density.

The DP clustering algorithm is based on the aforementioned concept to achieve data clustering, where local density $\boldsymbol{\rho}_i$ and relative distance δ_i are two crucial factors that significantly influence the outcomes of the DP algorithm.

Figure 3 provides an example that illustrates the fundamental concept of the DP clustering algorithm. The cluster centers (designated as 1 and 10) are encompassed by samples from the same class, which exhibit lower local densities compared to the cluster centers (see Figure 3). Additionally, samples located far from the class centers typically exhibit very low local densities.



(**a**) (**b**) (**c**)

Figure 3. Graph example illustrating the fundamental concept of the DP clustering algorithm. (**a**) Distribution of the scene samples belonging to different classes. (**b**) The relationship between relative distances and local densities of the scene samples. (**c**) Local densities of the scene samples. Different colors represent different classes of the scene samples.

The main steps of the DP clustering algorithm are outlined as follows. Given a dataset $\mathbf{X} \in \mathbb{R}^M$, where M represents the number of samples, the Euclidean distance $\boldsymbol{d}_{ij}$ between the samples $\mathbf{X}_i$ and $\mathbf{X}_j$ can be calculated as follows:

$$\boldsymbol{d}_{ij} = \left\|\mathbf{X}_i - \mathbf{X}_j\right\|_2^2. \tag{1}$$

The local density $\boldsymbol{\rho}_i$ of each sample can be obtained as follows:

$$\boldsymbol{\rho}_i = \sum\nolimits_{j \neq i} \chi\left(\boldsymbol{d}_{ij} - d_c\right) \tag{2}$$

where d_c is the cut-off distance and is constrained by a parameter P, which takes a value between 10 and 30 [48]. In Equation (2), $\chi(u) = \begin{cases} 1, u < 0 \\ 0, u \geq 0 \end{cases}$.

After computing $\boldsymbol{\rho}_i$, δ_i is defined as follows:

$$\delta_i = \begin{cases} \max\limits_{j}\left(\boldsymbol{d}_{ij}\right), & if\ \boldsymbol{\rho}_i = max(\rho) \\ \min\limits_{j:\rho_j > \rho_i}\left(\boldsymbol{d}_{ij}\right), & Otherwise. \end{cases} \tag{3}$$

Generally, δ_i refers to the minimum distance between sample i and other samples with a higher density than sample i. In special cases, δ_i refers to the maximum distance between sample i and samples other than sample i. Samples with relatively higher $\boldsymbol{\rho}_i$ and δ_i values will be identified as clustering centers. Therefore, an index γ_i is defined as follows:

$$\gamma_i = \boldsymbol{\rho}_i \times \delta_i. \tag{4}$$

The samples with higher γ_i values are more likely to be clustering centers. Therefore, the clustering center in the dataset can be easily found using a sorting algorithm. In Equation (2), outlier samples are detected by a hard index constraint, such as d_c. In accordance with previous studies [39,40], this paper adopts a soft Gaussian kernel function to detect the outlier probability of samples, which is defined as follows:

$$\rho_i = \sum_j e^{-\left(\frac{d_{ij}}{d_c}\right)^2}. \tag{5}$$

The advantage of the soft Gaussian function is that it can reduce the negative impact of statistical errors caused by fewer samples in some categories.

2.2. Automated Generation of Initial Samples with Vector Boundary Constraints

2.2.1. Automatic Generation of Initial Samples

The quality of sample generation and the accuracy of vector polygons change detection in subsequent stages are directly influenced by the segmentation results of images. While the SLIC algorithm can rapidly generate uniform and compact superpixels, it does not guarantee a perfect fit between superpixels and ground object boundaries, potentially leading to boundary exceedance situations [49].

To address this issue, we propose a novel approach termed boundary-preserved masking SLIC. By incorporating prior knowledge of vector boundaries, our approach significantly enhances the accuracy of image segmentation, as illustrated in Figure 4.

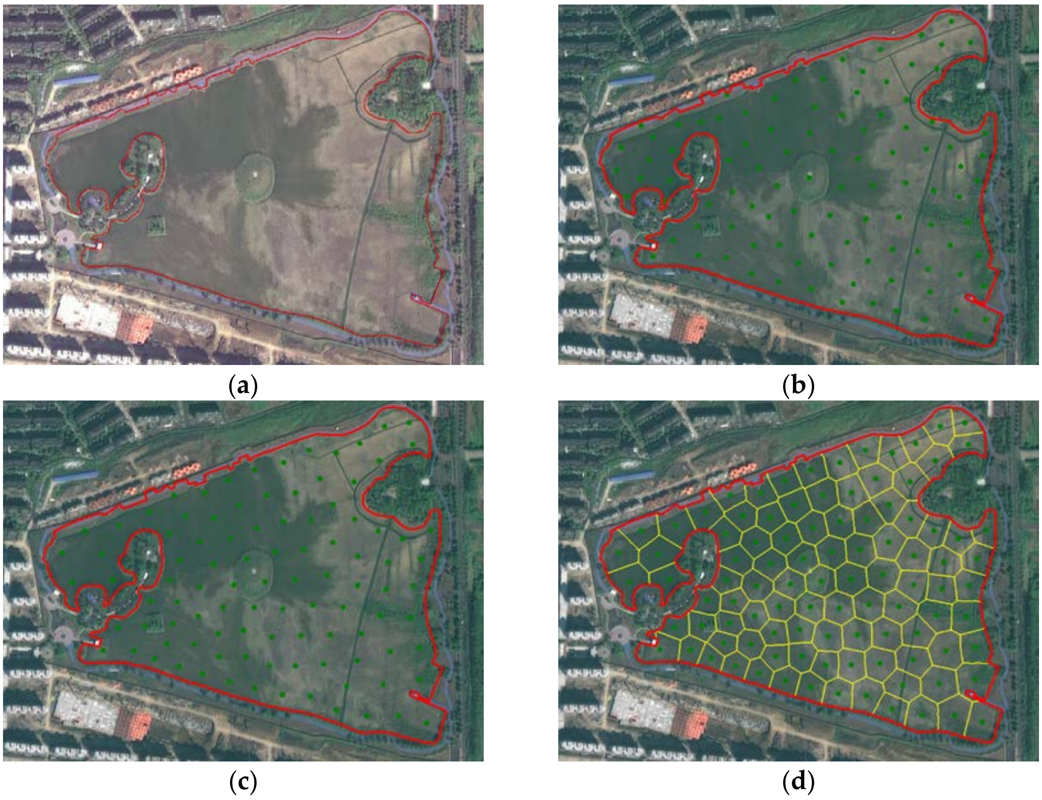


Figure 4. Steps of boundary-preserved masking SLIC. (**a**) Original image and corresponding land cover vector polygon. (**b**) Seed points are initialized using the farthest point sampling strategy. (**c**) Seed point positions are fine-tuned based on the k-means algorithm to ensure a more uniform distribution. (**d**) Pixels within the patch are clustered based on the color and spatial distance to the nearest seed point.

The details of the algorithm are as follows:

1. Given the number of seed points (N), to distribute them uniformly across the region, each seed point is positioned furthest from the boundary and from any other seed point. The seed point (P) is calculated as follows:

$$P = \underset{x}{\operatorname{argmax}}(\|x - y\|_2), \tag{6}$$

where y represents a point on the boundary or a seed point and x represents another point within the region. The boundary of the region is derived from a polygon. To ensure a rational spatial distribution of seed points, N is calculated using a ratio, as depicted in (7), where K denotes the total number of pixels in the mask and C_0 denotes the preset cropping size for the sample. Assuming each superpixel is uniformly square, the edge length is defined as S, where $S = m \times C_0$. The parameter m is introduced to preserve specific proximity information while ensuring the integrity of the features in the generated data. Typically, $0 < m \leq 1$.

$$N = K\Big|(m \times C_0)^2 \tag{7}$$

2. The image is converted from RGB color space to CIELAB color space to facilitate the measurement of color differences. The weighted distance d between color and space, used in clustering, can assess the similarity between pixels, as expressed in Equation (9), where w is the weight balancing the color distance (d_c) and spatial distance (d_s). To minimize repeated searches, a circular area with the seed point as the center and a radius of $2S$ is defined as the search range during iterative clustering. Subsequently, any isolated small clusters are merged with adjacent larger clusters based on lightness similarity $d_l = (\mu - \mu_m)^2$, where μ and μ_m are the average lightness of the small cluster and the neighboring cluster, respectively.

$$d_c = \sqrt{\sum_{n \in l,a,b} (n_j - n_i)^2} \tag{8}$$

$$d = \sqrt{(d_c)^2 + (d_s/w)^2} \tag{9}$$

The segmentation accuracy at the image boundary is effectively enhanced by this method, as demonstrated in Figure 5. Following image segmentation, we employed an adaptive cropping method to automatically generate initial samples. Figure 6 illustrates the process of adaptive image cropping. With the initial cropping size C of the sample, superpixels are utilized as processing units for adaptive cropping, relying on the results of image superpixel segmentation to generate the sample. Nevertheless, increasing the cropping scale raises the likelihood of generating samples that contain features from other land cover types, resulting in a decrease in sample quality. Additionally, when cropping samples along the edges of imagery, areas extending beyond the image boundaries may result in null values (Nodata), as shown in Figure 6a. To address these issues, we propose a sample-adaptive cropping method with a formula for determining the cropping size C as follows:

$$C = \begin{cases} C_0, d_b > C_0 \\ d_b, d_b \leq C_0 \end{cases}, \tag{10}$$

where C_0 is the preset size, and d_b is the shortest distance between the center and boundary. Finally, pure multi-scale samples were generated based on the constraints of vector polygons boundaries.

2.2.2. Source of Noise Samples

The sources of noise samples primarily encompass the following aspects. Firstly, they encompass discrepancies between vector boundaries and actual ground object boundaries due to variations in operating personnel and standards contribute to noise generation, as shown in Figure 7. Secondly, the formation of vector survey results often takes a certain time period, leading to disparities between land class attributes derived from vector polygons and those obtained from high-resolution RS images (see Figure 8). Consequently, it is not feasible to directly extract labeled samples for training solely based on vector polygons depicted on high-resolution RS images. In view of this, DP clustering is introduced for initial samples denoising with the aim of addressing the issue of automatic sample set construction.

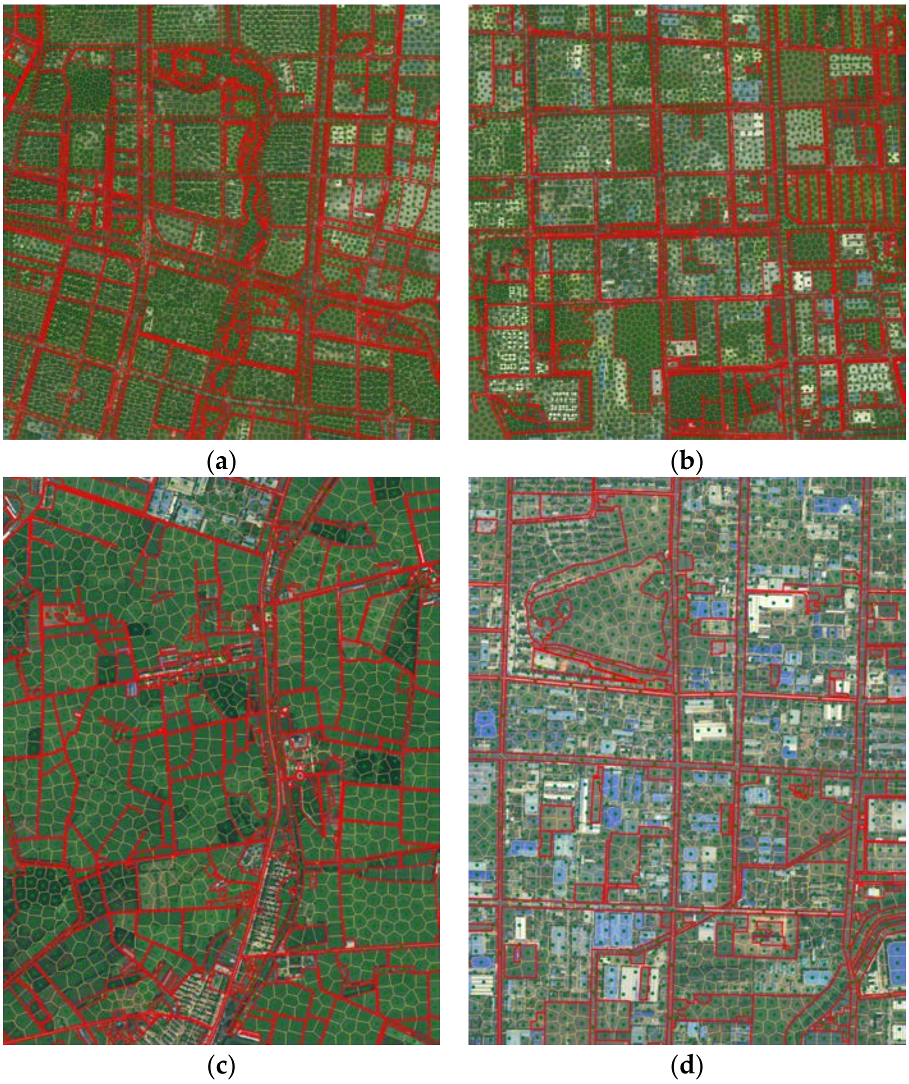


Figure 5. Segmentation results of the boundary-preserved masking SLIC. Parts (**a**,**b**): results of the Nantong dataset. Parts (**c**,**d**): results of the Guantan dataset.

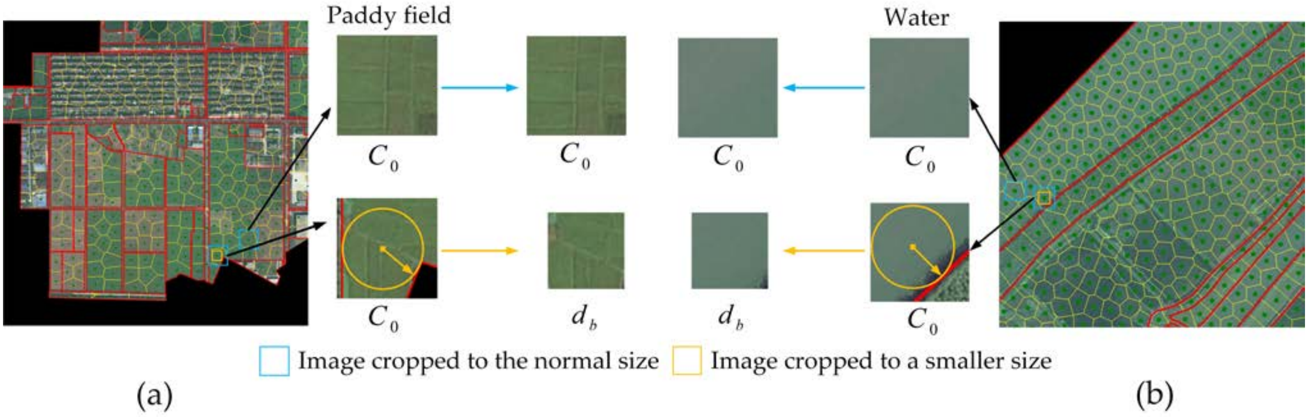

Paddy field
Water
C_0
C_0
C_0
C_0
C_0
d_b
d_b
C_0
Image cropped to the normal size
Image cropped to a smaller size
(a)
(b)

Figure 6. Adaptive cropping of images. Part (**a**) signifies that when the boundary of the image is cropped, any areas that exceed the original image dimensions will result in null values. Part (**b**) suggests that as the cropping size expands, the likelihood of the sample incorporating additional land cover classes also increases. Normal size cropping results (**upper**) and reduced size cropping results (**lower**).

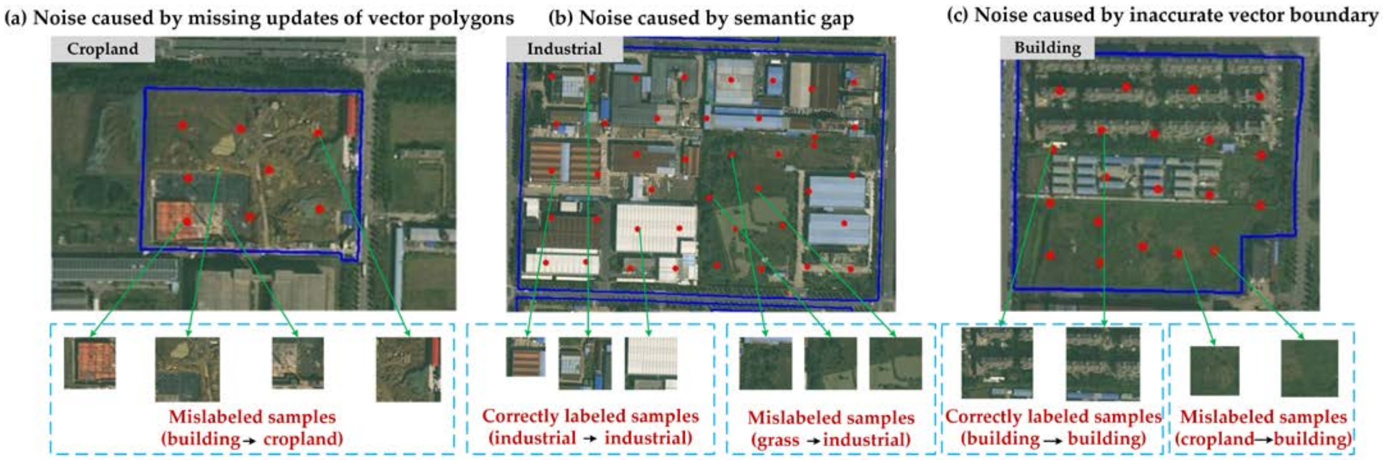

(a) Noise caused by missing updates of vector polygons
Cropland
Mislabeled samples (building→ cropland)
(b) Noise caused by semantic gap
Industrial
Correctly labeled samples (industrial → industrial)
Mislabeled samples (grass →industrial)
(c) Noise caused by inaccurate vector boundary
Building
Correctly labeled samples (building → building)
Mislabeled samples (cropland→building)

Figure 7. Several types of noise in the initial samples. (**a**) Missing updates of vector polygons. (**b**) Semantic gap. (**c**) Inaccurate vector boundary.

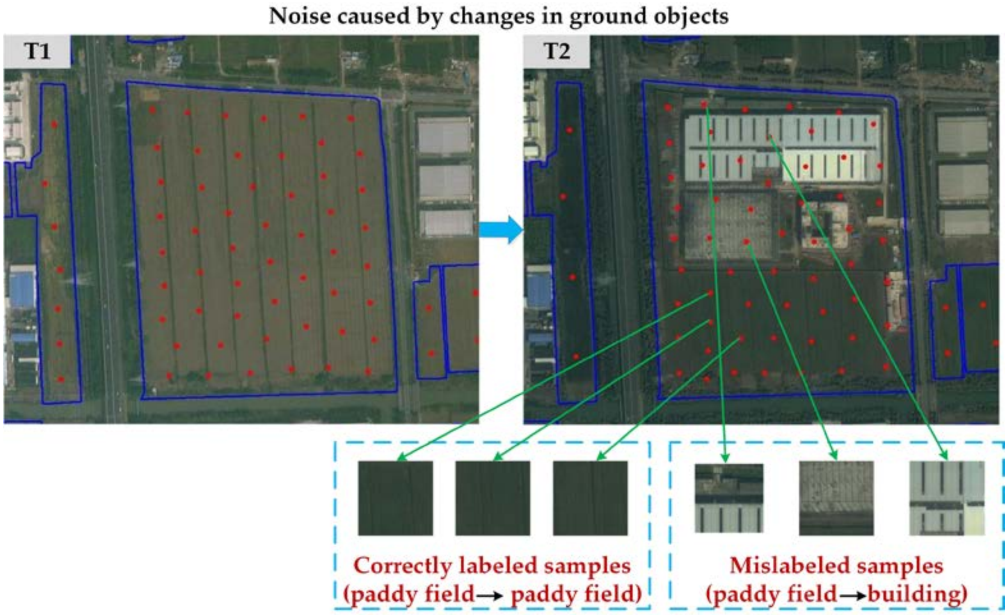

Noise caused by changes in ground objects
T1
T2
Correctly labeled samples (paddy field→ paddy field)
Mislabeled samples (paddy field→building)

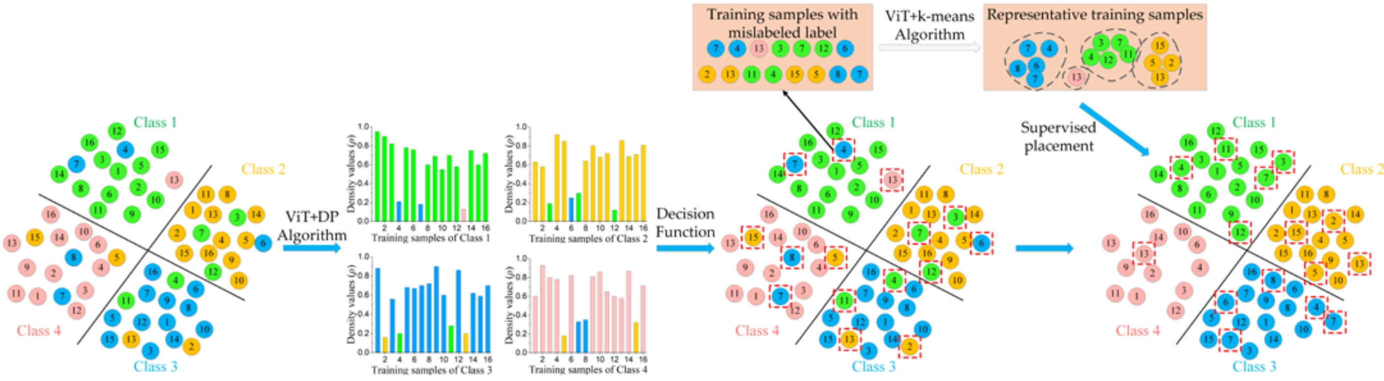
Figure 8. Noise samples caused by changes of ground objects on images in different periods.

2.3. Initial Samples Denoising Based on DP Clustering Algorithm

Sample denoising plays a crucial role in acquiring high-quality training samples, thereby enhancing the classification accuracy of the model [50,51]. Consequently, this paper introduces the EViTCC-DP approach, as illustrated in Figure 9. The approach comprises the following steps:

1. Employing boundary constraints of vector polygons and adaptive cropping of RS images to automatically generate initial samples and train ViT models.
2. Utilizing the pre-trained ViT to extract features from scene samples and inputting them into the DP clustering algorithm according to class constraints to achieve the purpose of denoising.

The approach uses the vector polygon attribute to classify the samples, thereby eliminating the need to define the cluster center in the DP clustering algorithm, and can construct a high-quality multi-scale sample set with minimal manual intervention. In this paper, we introduce the DP clustering algorithm for removing noise samples in RS image scene classification for the first time. Additionally, this paper improves classification accuracy in subsequent stages by transforming noise samples into RTS under the supervision of k-means clustering. The k-means algorithm surpasses random selection in terms of selecting optimal clustering centers while ensuring consistency between cluster numbers and label category data, thus obviating the requirement for manual definition.


Training samples with mislabeled label
ViT+k-means Algorithm
Representative training samples
Supervised placement
Class 1
Class 2
Class 3
Class 4
ViT+DP Algorithm
Density values (ρ)
Training samples of Class 1
Training samples of Class 2
Training samples of Class 3
Training samples of Class 4
Decision Function

The scene sample set with mislabeled training samples

The local density values of different scene samples

Detection abnormal density values of different scene samples

The high quality training set with representative training samples

Figure 9. Graph example illustrating the principle of EViTCC-DP approach. Different colors represent different labels of the training samples.

The implementation steps are described as follows:

1. Calculating the distance of training samples

Firstly, image feature embeddings are extracted for each sample using a pre-trained ViT model. Then, let $\mathbf{X} = \left\{\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^M\right\}$ refer to the feature embeddings of the initial samples, where M denotes the number of classes and X^m refers to the training samples in the mth class. For two training samples belonging to the mth class, i.e., $\mathbf{X}_a^j$ and $\mathbf{X}_b^j$, the distance d_{ab}^m between two samples can be measured. In this paper, Euclidean distance [39] is used as the distance measurement between training samples, with the distance $\boldsymbol{d}_{ab}^j$ between $\mathbf{X}_a^j$ and $\mathbf{X}_b^j$ being able to be calculated as follows:

$$\boldsymbol{d}_{ab}^j = \left\|\mathbf{X}_a^j - \mathbf{X}_b^j\right\|_2^2. \tag{11}$$

Through calculating the distances among the nth scene sample and the scene samples in the mth class, a distance array $\boldsymbol{d}_n^m$ of the nth scene sample can be constructed as follows:

$$\boldsymbol{d}_n^m = \left[\boldsymbol{d}_{n1}^m, \boldsymbol{d}_{n2}^m, \cdots, \boldsymbol{d}_{nN_m}^m\right]^T \tag{12}$$

where N_m refers to the number of samples in the mth class. In this way, a distance matrix $\boldsymbol{D}^m$ can be constructed as $\boldsymbol{D}^m = \left\{\boldsymbol{d}_1^m, \boldsymbol{d}_2^m, \cdots, \boldsymbol{d}_{N_m}^m\right\}$.

2. Calculating the local density of the training samples

The cut-off distance d_c^m can be calculated as follows:

$$d_c^m = \mathbf{S}^m(t)\text{s.t.}t = \langle \frac{N_m \cdot (N_m - 1)}{100} \cdot P \rangle \tag{13}$$

where $\mathbf{S}^m$ is a matrix that sorts the non-zero elements in the upper triangular matrix of $\boldsymbol{D}^m$ from the smallest to the largest elements, P is a free parameter that will be analyzed in Section 4.1, and $< \cdot >$ refers to the round operation.

With the above-obtained d_c^m, the local densities $\boldsymbol{\rho}^m = \left\{\boldsymbol{\rho}_1^m, \boldsymbol{\rho}_2^m, \cdots, \boldsymbol{\rho}_{N_m}^m\right\}$ of the samples in the mth class can be calculated as follows:

$$\boldsymbol{\rho}^m = \sum e^{-\left(\frac{D^m}{d_c^m}\right)^2}. \tag{14}$$

3. Detecting the mislabeled samples in each class

Once the local densities of the training samples in different classes are obtained, mislabeled samples can be easily detected and removed as follows:

$$\mathbf{Y}_i^m = \begin{cases} \mathbf{X}_i^m \ if \ \rho_{N_m}^m \geq \lambda \cdot \overline{\boldsymbol{\rho}}^m \\ \varnothing \ \ \text{Otherwise} \end{cases} \tag{15}$$

where $\mathbf{Y} = \left\{\mathbf{Y}^1, \mathbf{Y}^2, \cdots, \mathbf{Y}^M\right\}$ refers to the resulting training set, in which the noisy labels are detected and removed. λ is a free parameter.

4. Converting the mislabeled samples to representative samples

The scene sample $\mathbf{X}_a^j$ represents noise for the mth class; however, it may be a valuable representative sample for the nth class. In this paper, the final high-quality sample set can be automatically constructed as follows:

$$\hat{\mathbf{Y}}_i^m = \mathbf{Y}_i^m + \mathrm{k}\varnothing_{class}^j \tag{16}$$

where $\hat{\mathbf{Y}} = \left\{\hat{\mathbf{Y}}^1, \hat{\mathbf{Y}}^2, \cdots, \hat{\mathbf{Y}}^M\right\}$ refers to the final high-quality training scene sample set in which the noisy labels are supervised and placed into the correct class using the k-means clustering algorithm, and $\mathrm{k}\varnothing_{class}^j$ represents the k-means clustering algorithm.

2.4. Attention-Based Multi-Scale ConvTransformer Network, AMCT-Net

2.4.1. Overview of the Proposed AMCT-Net

The samples are generated using vector polygons and high-resolution RS images, where each element requires fine-grained image classification [52]. Traditional methods for image classification encounter challenges in capturing the intricate features of an image. However, the fusion of the Convolutional Neural Network (CNN) and Transformer enables the extraction of both global and local features, thereby enhancing classification accuracy [17,23,27,53]. In this article, we introduce an attention-based multi-scale ConvTransformer network (AMCT-Net), depicted in Figure 10. AMCT-Net integrates a spatial attention mechanism and a multi-scale feature extraction module into the Transformer architecture to enhance fine-grained classification accuracy.

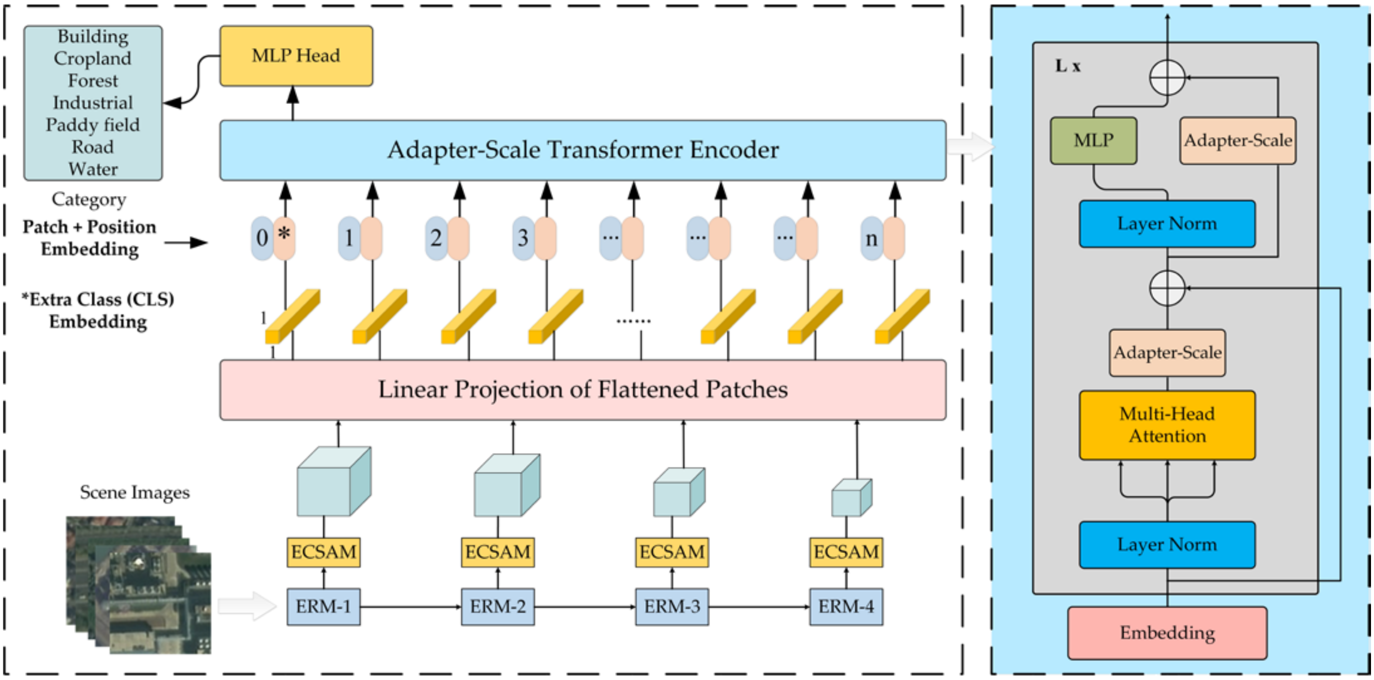


Figure 10. Example architecture of the proposed AMCT-Net.

The proposed AMCT-Net adopts a Siamese structure for local feature extraction, comprising four Encoder Residual Modules (ERMs) that generate feature maps of varying sizes with channel dimensions of 16, 32, 64, and 128, respectively. Subsequently, to enhance model performance, we incorporate an attention mechanism known as the Efficient Convolution Spatial Attention Module (ECSAM) at the backend of ERM. ECSAM prioritizes channel features and effectively leverages low- and high-level features [54]. Finally, ECSAM feature maps of different sizes are transformed into flattened feature patch tokens with uniform shapes. Similar to the original Transformer [55], we flatten the patches and map them to D dimensions using a trainable linear projection. We refer to the output of this projection as the patch embeddings. Position embeddings are added to the patch embeddings to retain positional information. The resulting sequence of embedding vectors is utilized as input to the Adapter-Scale Transformer Encoder for classification, facilitating the extraction of global features. Following an overview of the general motivation and architecture of our proposed method, we proceed to describe each main improved module in detail.

2.4.2. Module Details

1. Efficient Convolution Spatial Attention Module

Generally, spatial attention mechanisms are employed to assist models in focusing on crucial image regions. However, computing weights between features across all positions entails significant computational effort. In this study, drawing inspiration from the concept of “coordinate separation” [56], we propose the ECSAM to effectively capture cross-channel relationships and long-range dependencies while incorporating specific positional information.

The detailed architecture of ECSAM is illustrated in Figure 11. First, the input feature map, denoted as $F \in \mathbb{R}^{H\times W\times C}$, undergoes horizontal avg-pooling of size $(\mathrm{H}, 1)$ and vertical avg-pooling of size $(1, \mathrm{W})$ for each channel of the feature map, respectively. These operations are defined as follows:

$$F_c^H(h) = \frac{1}{W}\sum_{0\leq x<W} F_c(h, x) \tag{17}$$

$$F_c^W(w) = \frac{1}{H}\sum_{0\leq y<H} F_c(y, w) \tag{18}$$

where $F_c^H(h)$ represents the horizontal avg-pooled feature of the c-th channel at height h, and $F_c^W(w)$ represents the vertical avg-pooled feature of the c-th channel at width w.

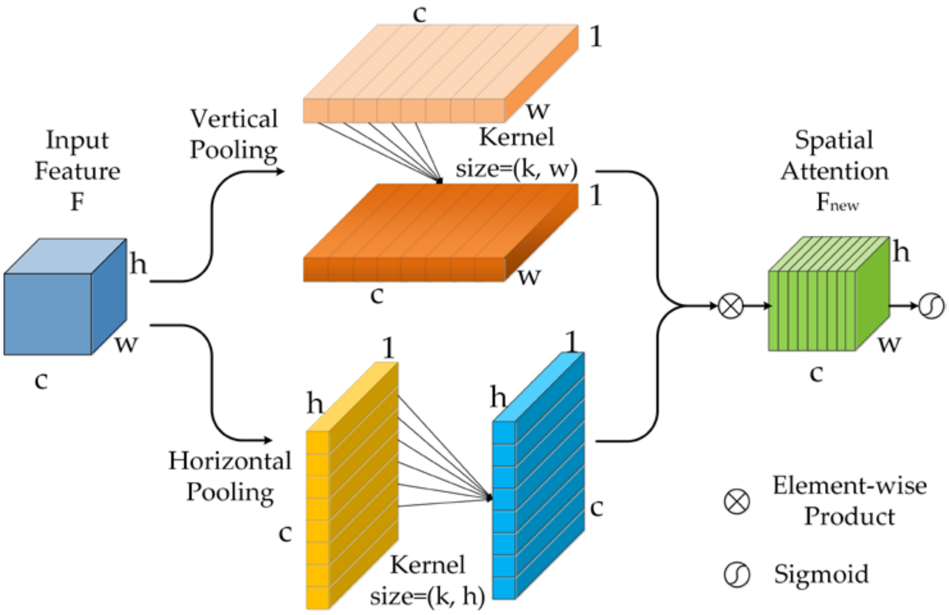


Figure 11. Architecture of the proposed ECSAM.

To prevent reduction in channel dimensionality during cross-channel interactions, ECSAM employs 1D convolution extended to 2D with an adaptive kernel size for generating attention weights along two spatial directions, respectively. These operations are formulated as

$$\hat{F}_c^H(h) = ReLU(C2D_k(F_c^H(h))) \tag{19}$$

$$\hat{F}_c^W(w) = ReLU(C2D_k(F_c^W(w))) \tag{20}$$

where $\hat{F}_c^H(h)$ denotes the horizontal feature vector of the c-th channel at height h, $\hat{F}_c^W(w)$ denotes the vertical feature vector of the c-th channel at width w, and $C2D_k$ represents 2D convolution with kernel size of k, where k can be calculated as

$$\mathrm{k} = \left|\frac{\log_2(C)+1}{2}\right|_{odd} \tag{21}$$

where $|\mathrm{t}|_{odd}$ represents the nearest odd number of t, and C denotes the number of channels in F.

2. Encoder Residual Module

The residual module structure of the encoder is shown in Figure 12a. Initially, the input features traverse a convolutional layer (Conv), followed by batch normalization (BN) and rectified linear unit (ReLU) activation. Subsequently, another combination of Conv and BN is introduced. Ultimately, the output is derived by adding the outcome of the second BN to the outcome of the initial Conv, followed by another ReLU. The Conv kernels are sized at 3 × 3. Employing residual connections and activation function layers serves to accelerate network learning and mitigate the problem of gradient vanishing [25].

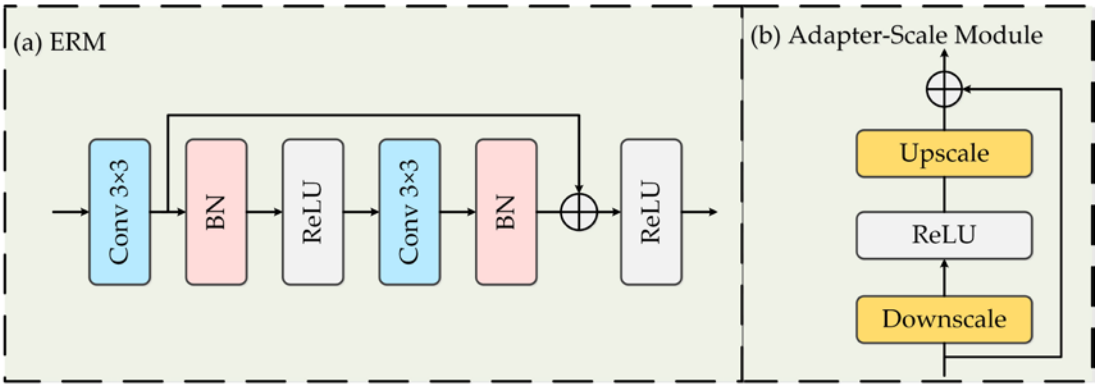


Figure 12. Architecture of the embedded modules. (**a**) ERM. (**b**) Adapter-Scale Module.

3. Adapter-Scale Module

The architecture of the Adapter-Scale Module (ASM) is depicted in Figure 12b. It comprises three components: Downscale, ReLU, and Upscale. The Downscale segment utilizes a single Multi-Layer Perceptron (MLP) layer to reduce the dimensionality of the embedding. Subsequently, the ReLU activation function is applied, and the embedding is restored to its original dimensionality through another MLP layer in the Upscale segment. Two ASMs are integrated into the ViT block. The first is positioned before the multi-head attention blocks and residual connections, while the second is embedded within the residual structure of the MLP. Additionally, a scale factor of 0.5 is applied to each adapter.

2.5. Vector Polygons Change Detection Analysis Based on Confidence Rules

The set of image polygons generated by the post-processing unit R_i after interpretation is denoted as $\{P_N^i\}$, where N represents the total number of image polygons. Let y_N^i refer to the classification attribute confidence of P_N^i, with y_N^i being able to be calculated by counting the occurrences of $\{x\}$, where $\{x\}$ is the set of class attribute values. The attribute $X_{P_N^i}$ of

P_N^i is subsequently computed based on the maximum confidence rule. The operation is formulated as follows:

$$X_{P_N^i} = argmax\left\{y_N^i\right\}. \tag{22}$$

Finally, the vector polygon attribute value Y can be obtained with

$$Y = argmax\{S_x\} \tag{23}$$

where x is the attribute of different types of processing units after interpretation, and S_x is the area corresponding to x.

The threshold k for change detection can be calculated with

$$k = \frac{S_Y}{S_{total}} \times 100\% \tag{24}$$

where S_Y is the area corresponding to Y, and S_{total} is the total area of the vector polygon.

In this experiment, the threshold k was set at 0.9. The changed vector polygons were identified based on the "prediction category <> initial category" and a developed change decisionmaker, as illustrated in Figure 13. A detailed explanation of the change decisionmaker will be provided in Section 3.2.1.

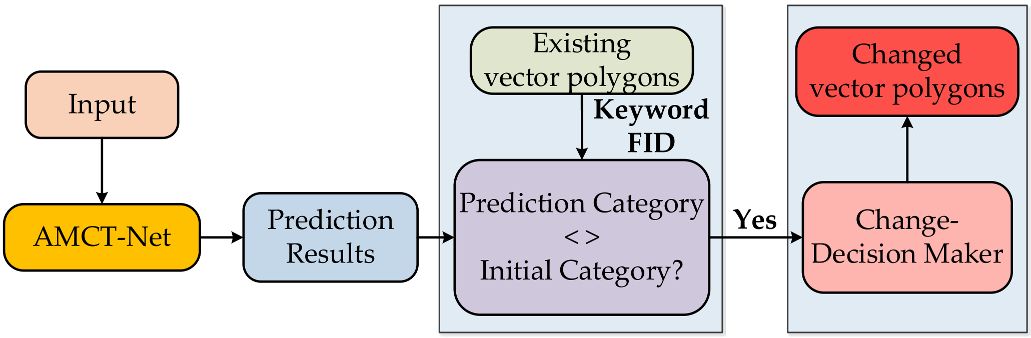


Figure 13. Flowchart of change detection.

3. Experiments and Results

3.1. Description of Data Sources and Research Scheme

The RS images and vector data used in this study were provided by the land change survey project conducted by the Jiangsu Institute of Geology and Surveying. The Nantong dataset is situated in Nantong Development Zone, Nantong City, Jiangsu Province, China (see Figure 14). The existing vector data depicted in Figure 14a were collected during the third land survey conducted in November 2021. The images presented in Figure 14b,c were obtained from the Beijing-2 satellite with a spatial resolution of 0.8 m in October 2021 and October 2022. Similarly, the Guantan dataset is located in Guantan Town, Xuyi County, Huai'an City, Jiangsu Province (refer to Figure 15). Moreover, the existing vector data shown in Figure 15a were also collected during the third land survey conducted in November 2021. The images presented in Figure 15b,c were acquired from the GF-2 satellite with a spatial resolution of 1m, with images also being taken in October 2021 and October 2022. Temporal image details are illustrated in Figures 14d and 15d, with land use types within these areas primarily comprising buildings, cropland, forest land, industrial land, paddy fields, roads, and water bodies. Considering the spatial resolution of the images, the standard sample size for both datasets in the experiment was set at 96 pixels, with a minimum size of 32 pixels. The final multi-scale sample sizes range from 32 × 32 to 96 × 96 pixels, as depicted in Figures 14e and 15e.

After automatically generating the initial sample set, we obtained a high-quality training sample set using the EViTCC-DP method. Specifically, the sample categories and quantities for the two datasets are as shown in Table 1.

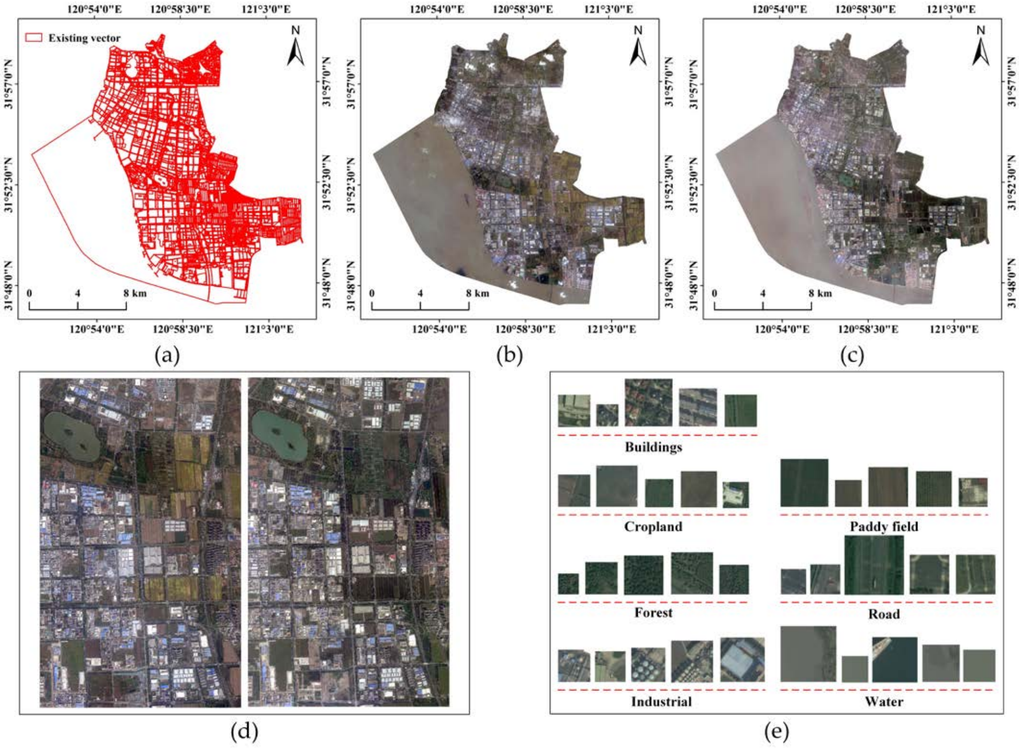


Figure 14. The Nantong dataset obtained from the Nantong Development Zone, Jiangsu, China. (**a**) Existing vector updated in November 2021. (**b**) Image acquired from BeiJing-2 in October 2021. (**c**) Image acquired from BeiJing-2 in October 2022. (**d**) Detailed presentation of two phases of images. (**e**) Examples of generated samples in the Nantong dataset.

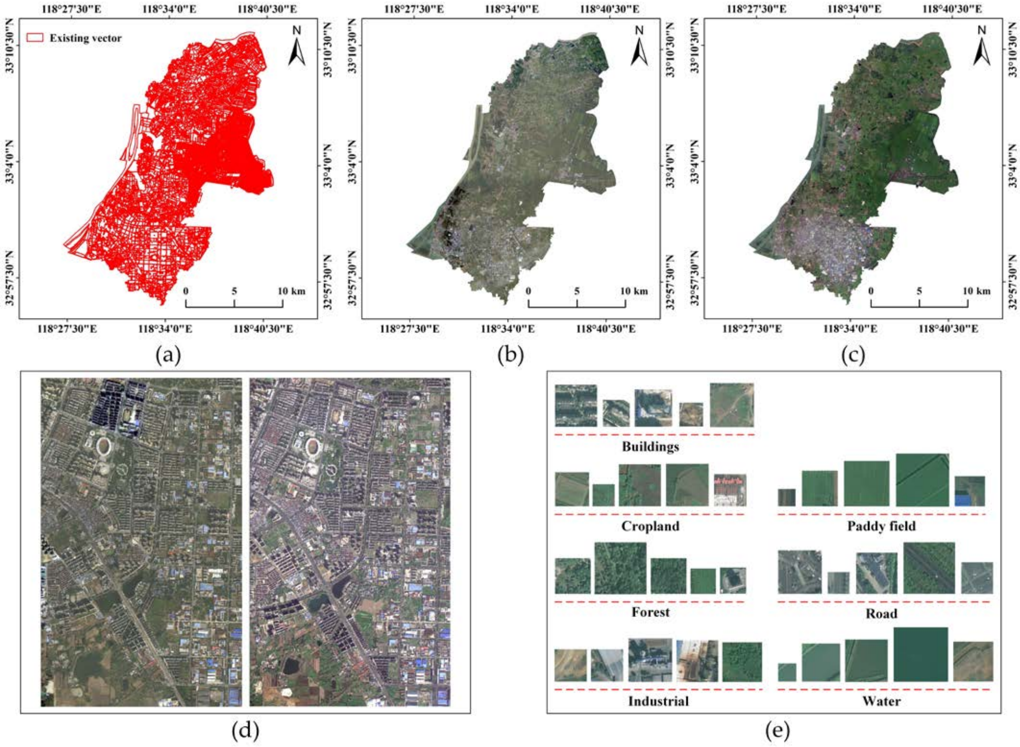


Figure 15. Guantan dataset obtained from Guantan, Xuyi, Huai'an, China. (**a**) Existing vector updated in November 2021. (**b**) Image acquired from GF-2 in October 2021. (**c**) Image acquired from GF-2 in October 2022. (**d**) Detailed presentation of two phases of images. (**e**) Examples of generated samples in the Guantan dataset.

Table 1. Sample Information of the Nantong dataset and the Guantan dataset.

Number	Land Use Types	Nantong Dataset	Guantan Dataset
C1	buildings	5243	4849
C2	cropland	2647	1744
C3	forest	3201	3768
C4	industrial	5844	4160
C5	paddy field	4848	5115
C6	road	2738	1240
C7	water	5690	5139

The experiments were conducted on a computer equipped with 64GB of RAM and an Intel Xeon CPU E-2186M @ 2.9GHz processor. The training process utilized the NVIDIA GeForce GTX 1080 Ti with 11GB of memory. The source code for the proposed ConvTransformer was implemented using PyTorch 1.5.1+cu92 and Python 3.7. During training, the stochastic gradient descent algorithm was employed, incorporating a momentum value of 0.9 and a weight decay penalty coefficient of 10^{-5}. The initial learning rate was set to 0.001 and decayed following a cosine annealing schedule. A batch size of 64 was utilized.

3.2. Results

3.2.1. Change Detection and Post-Processing

There exists a significant semantic disparity between RS images and existing vector polygons, resulting in the potential inclusion of non-subject objects within the subject object, thereby leading to impure land classification within vector polygons. To achieve reliable change detection results in accordance with task requirements, we have developed a "change decisionmaker" (see Figure 16) that effectively mitigates the influence of non-subject objects through the implementation of distinct thresholds. As depicted in Figure 16, vector polygons are categorized into two types, namely 'sensitive' and 'non-sensitive'. The 'sensitive' type [see Figure 16a] represents the most stringent change scenario, involving the presence of new buildings in croplands, which is prohibited by law in certain countries. On the other hand, other scenarios can be classified as the 'non-sensitive' type [see Figure 16b], which do not violate legal regulations or relevant rules and receive comparatively less attention than the 'sensitive' type.

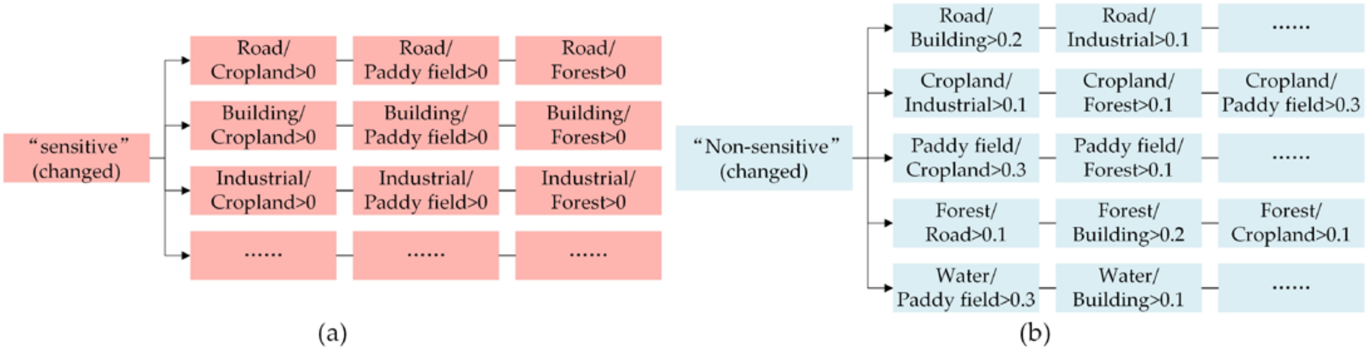


Figure 16. Detailed structure of change decisionmaker. (**a**) Conditions for changed "sensitive" type. (**b**) Conditions for changed "non-sensitive" type. Specifically, when C: A/B > threshold, it is regarded as the change of C, where A is the number of congener data in C whose predicted value is inconsistent with the initial category, B is the total number of data in C, and C is a vector polygon (as long as one of the conditions is met, C is determined to be changed).

3.2.2. Evaluation Metrics

To compare the performance of our method with that of traditional models, we employed overall accuracy (OA), Precision, Recall, F1 score, and specificity as the primary quantitative metrics. A higher value for each metric indicates superior model performance.

The F1 score considers both precision and recall of the classification model on a scale from 0 to 1. Furthermore, tests with high specificity indicate a lower Class I error rate [57].

The definitions of these metrics are outlined as follows:

$$\text{OA} = \frac{TP + TN}{TP + FP + TN + FN} \tag{25}$$

$$Precision = \frac{TP}{TP + FP} \tag{26}$$

$$Recall = \frac{TP}{TP + FN} \tag{27}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{28}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{29}$$

where *TP* denotes true positives, *FP* denotes false positives, *TN* denotes true negatives, and *FN* denotes false negatives.

3.2.3. Vector Polygons Change Detection Results

The accuracy of subsequent vector polygons' change detection is directly influenced by the classification accuracy of the proposed method [58], which adopts a post-classification comparison approach. To assess the classification performance of the proposed AMCT-Net model, we conducted comparisons with several other state-of-the-art models, including ViT [59], MTC-Net [53], and HCTM [25]. Specifically, the ViT model utilizes the encoder module of the Visual Change Transformer (VcT), MTC-Net combines the advantages of the multi-scale Transformer with the convolutional block attention module (CBAM), and HCTM represents a hybrid CNN-Transformer model. Before training the model, 80% of samples from each dataset are allocated for training purposes, while the remaining 20% are reserved for accuracy validation.

To validate the effectiveness and robustness of the proposed method in denoising samples, we conducted comparative experiments using consistent training and test sets. All models were evaluated under the same baseline conditions, and the classification results on the two datasets are presented in Table 2.

Table 2. Comparison of the classification metrics of different models for the two datasets.

Model	Metrics	Nantong Dataset	Guantan Dataset
ViT	Accuracy	0.9068	0.9165
	Precision	0.8938	0.8914
	Recall	0.8903	0.9072
	Specificity	0.9816	0.9811
	F1 Score	0.8921	0.8992
MTC-Net	Accuracy	0.9117	0.9164
	Precision	0.9103	0.8969
	Recall	0.8904	0.9203
	Specificity	0.9824	0.9811
	F1 Score	0.8985	0.8996
HCTM	Accuracy	0.9135	0.9292
	Precision	0.8975	0.9230
	Recall	0.9109	0.9100
	Specificity	0.9817	0.9859
	F1 Score	0.9182	0.9215
AMCT-Net (ours)	Accuracy	0.9134	0.9351
	Precision	0.9179	0.9228
	Recall	0.9134	0.9292
	Specificity	0.9839	0.9898
	F1 Score	0.9201	0.9306

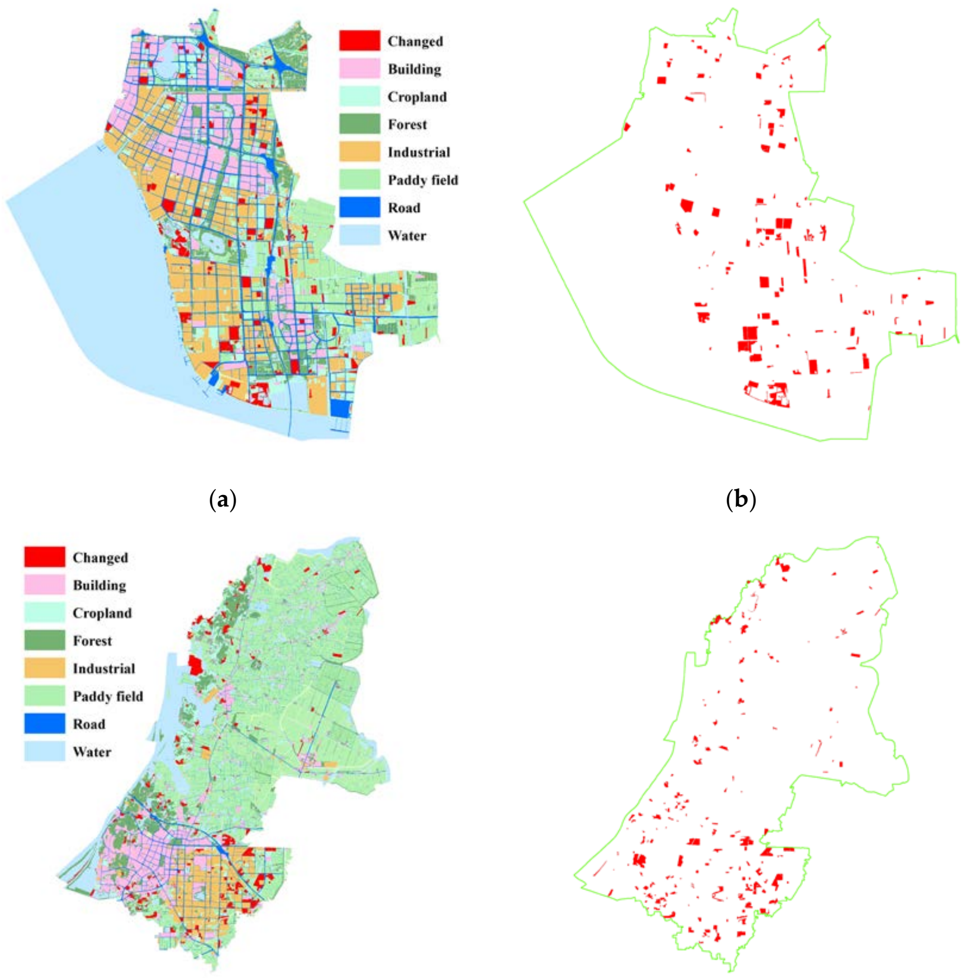
Color convention: best in red; second best in blue.

We can see the following from Table 2:

- Across both datasets, the baseline model (ViT) exhibits unsatisfactory performance across the five evaluation metrics, while the enhanced model incorporating attention mechanisms and a multi-scale convolution module demonstrates a notable improvement in accuracy. Notably, the AMCT-Net model outperforms other architectures in terms of Recall, specificity, and F1 score. Specifically, on the Nantong dataset, AMCT-Net achieves a Recall of 0.9134, specificity of 0.9839, and F1 score of 0.9201, representing a 0.25% increase in Recall compared to the sub-optimal HCTM model (Recall = 0.9109). On the Guantan dataset, AMCT-Net's Recall reaches 0.9292, specificity stands at 0.9898, and F1 score is 0.9306, with a significant 1.92% increase in Recall compared to the sub-optimal HCTM model (Recall = 0.9100). This underscores the substantial advancement in classification accuracy achieved by AMCT-Net.
- It is noteworthy that the performance of the model differs between the two datasets. For instance, the proposed AMCT-Net model only marginally improves accuracy by 0.66% compared to the baseline model on the Nantong dataset. Conversely, the model exhibits a much more substantial improvement in accuracy on the Guantan dataset, with an increase of 1.86% compared to the baseline model. The variation in performance may be attributed to the urban development context of the Nantong dataset, where change types are inherently more complex compared to the Guantan dataset. However, as AMCT-Net integrates the local feature extraction capabilities of CNN structures with the global information processing characteristics of Transformer architecture, supplemented by the introduction of a multi-scale module, these enhancements prove particularly advantageous for processing the multi-scale sample set in this study, underscoring its adaptability to diverse dataset features.

Following image classification, the detection of change vector polygons is achieved through a change decisionmaker and post-processing rules. Figure 17a,c illustrate the change detection outcomes for the two datasets, while the confusion matrix generated by the proposed AMCT-Net model in this study is presented in Figure 18. Apart from confusion between roads and buildings, there are instances of misclassification in industrial areas. This is due to the spatial distribution, color form, and density similarities between industrial areas, buildings, and roads, posing a challenge in scene classification. Moreover, croplands share local semantic characteristics with paddy fields, leading to potential misclassifications. Taking Figure 18 as an example, only croplands exhibit a classification accuracy below 90%. Nonetheless, the experimental model effectively identifies other scene categories with high recognition accuracy. Notably, some forest test samples contain mixed category information, such as croplands and paddy fields surrounding forest farms, resulting in a lower classification accuracy for this category. Additionally, for the Guantan dataset, the classification accuracy of water bodies is not 100% due to aquaculture activities in the research area, including the phenomenon of digging breeding pits in some croplands, leading to confusion between paddy fields and breeding pits. Overall, the AMCT-Net in this study achieves a high classification accuracy, demonstrating satisfactory performance.

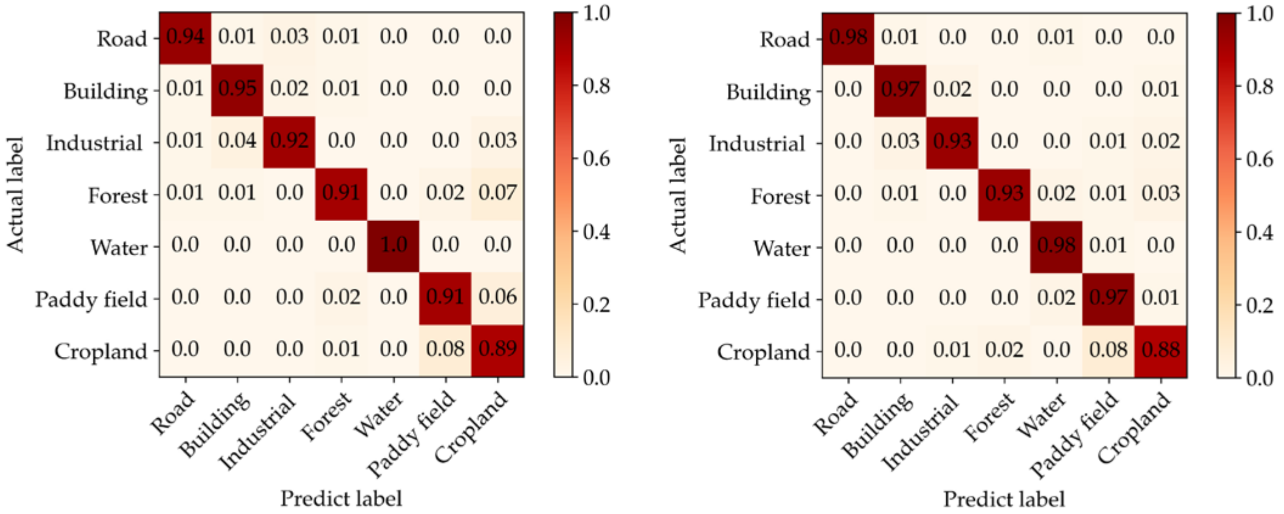
The visual interpretation results of the proposed method are illustrated in Figure 19, showcasing typical examples. Part (a) demonstrates the conversion of cropland into buildings, part (b) exhibits the construction of a road within the cropland, part (c) displays the transformation of part of the cropland into an industrial and mining area, and part (d) reveals traces of earthwork in the region. These aforementioned examples demonstrate that our proposed method is capable of effectively detecting changes in vector polygons with support from high-resolution RS images.



(**a**) (**b**)

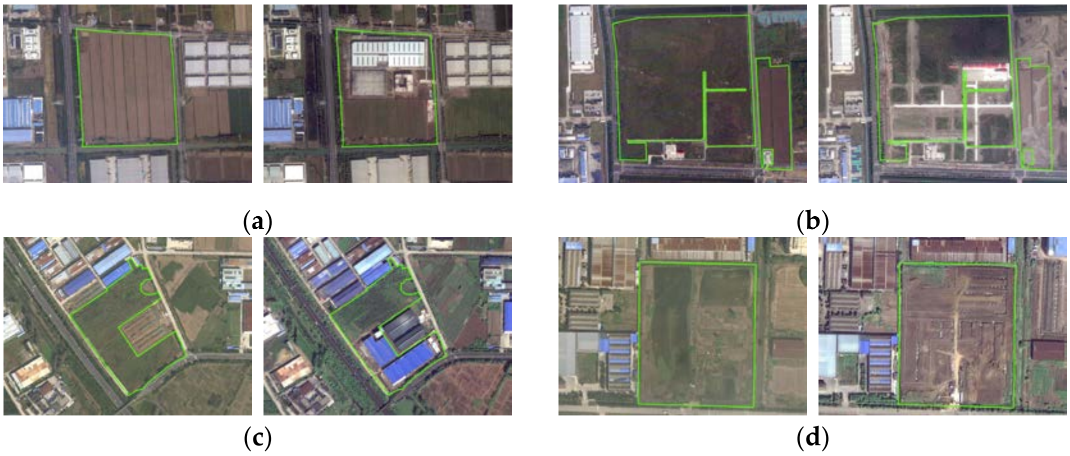
(**c**) (**d**)

Figure 17. Change detection results of the proposed model. Parts (**a**,**c**): results of the Nantong dataset and Guantan dataset. Parts (**b**,**d**): ground truth of the Nantong dataset and Guantan dataset.



(**a**) (**b**)

Figure 18. Confusion matrix of the proposed AMCT-Net on the Nantong dataset (**a**) and the Guantan dataset (**b**).



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

Figure 19. Typical examples of visual interpretation results. Parts (**a**,**c**): results of the Nantong dataset. Parts (**b**,**d**): results of the Guantan dataset.

4. Analysis and Discussion

4.1. Analysis of the DP Algorithm Parameters Selections

The influence of the parameters p and λ on the performance of the proposed method is analyzed in this section. To facilitate this analysis, a simple accuracy evaluation index called denoising accuracy (DA) was devised in this article.

The DA is calculated as follows:

$$\mathrm{DA} = \frac{N_{DP}}{N_{total} + N_{normal}} \tag{30}$$

where N_{DP} is the selected correct noise samples, N_{total} represents all noise samples in a category, and N_{normal} is the selected erroneous noise samples. The range of DA is between 0 and 1, and the larger its value, the better the denoising performance of the proposed method. Experimental results obtained from the two datasets are presented in Figure 20, using cropland as an illustrative example. The values of p and λ are selected from the intervals 10∼30 and 0.03∼0.07, respectively. Based on the experimental results presented in Figure 20, it can be found that p is actually not related to the number of samples in the training set N. For instance, despite different sample sizes for the Nantong dataset and Guantan dataset, a fixed $P = 20$ consistently achieves optimal denoising performance (see Figure 20). Moreover, λ emerges as a crucial parameter influencing denoising performance. Taking Figure 20a as an example, DA exhibits significant variations in response to changes in parameter λ. It is noted that $\lambda = 0.05$ and $P = 20$ consistently yield relatively optimal denoising accuracies. Furthermore, the value of λ is associated with the intensity of changes within the study area contextually considered at this time node (2021–2022), where China experienced a COVID-19 epidemic, which lead to restricted human activities; hence, there is no need for excessively large values of λ under such circumstances. Therefore, given a new dataset, $\lambda = 0.05$ and $P = 20$ are suggested to be used as the default parameters in the proposed method.

4.2. Influence of Sample Set Denoising

In contrast to conventional sample denoising methods that heavily rely on manual intervention, the proposed EViTCC-DP approach significantly enhances automation. To evaluate the impact of mislabeled samples on model training, we compared the accuracy of the Two-Classifier Cross-Validation (TCCV) [35] and EViTCC-DP before and after denoising the training samples. As shown in Tables 3 and 4, the OA of the model improved by 2.80% and 2.56% on the Nantong dataset and Guantan dataset, respectively, after denoising with EViTCC-DP.

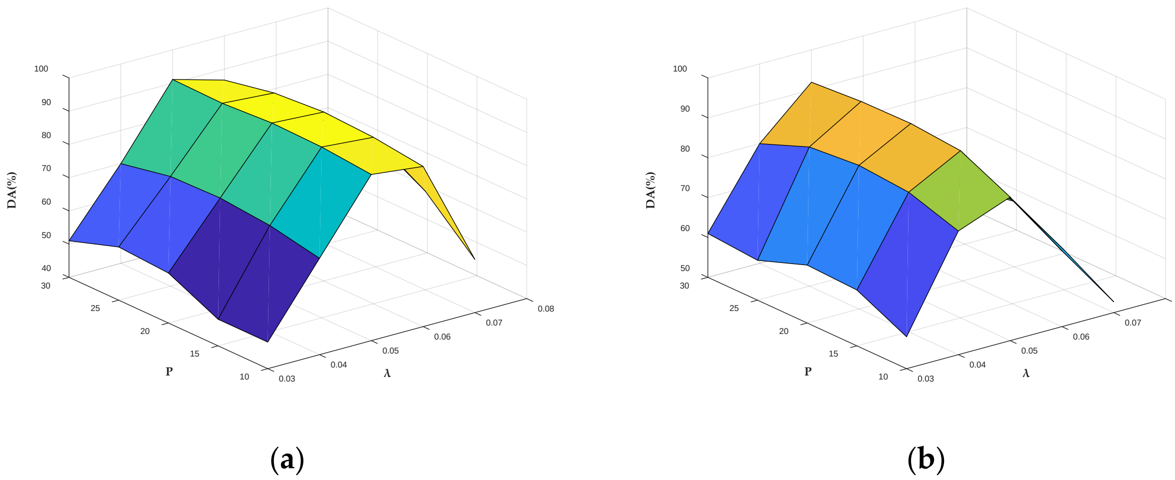


Figure 20. Influence of the parameters p and λ on the performance of the proposed method. Part (**a**): results of cropland in the Nantong dataset. Part (**b**): results of cropland in the Guantan dataset. A similar behavior occurred to the one that used other classes in the datasets.

Table 3. Comparison of the classification accuracy of the Nantong dataset before and after sample denoising (no improvement observed after epoch 82).

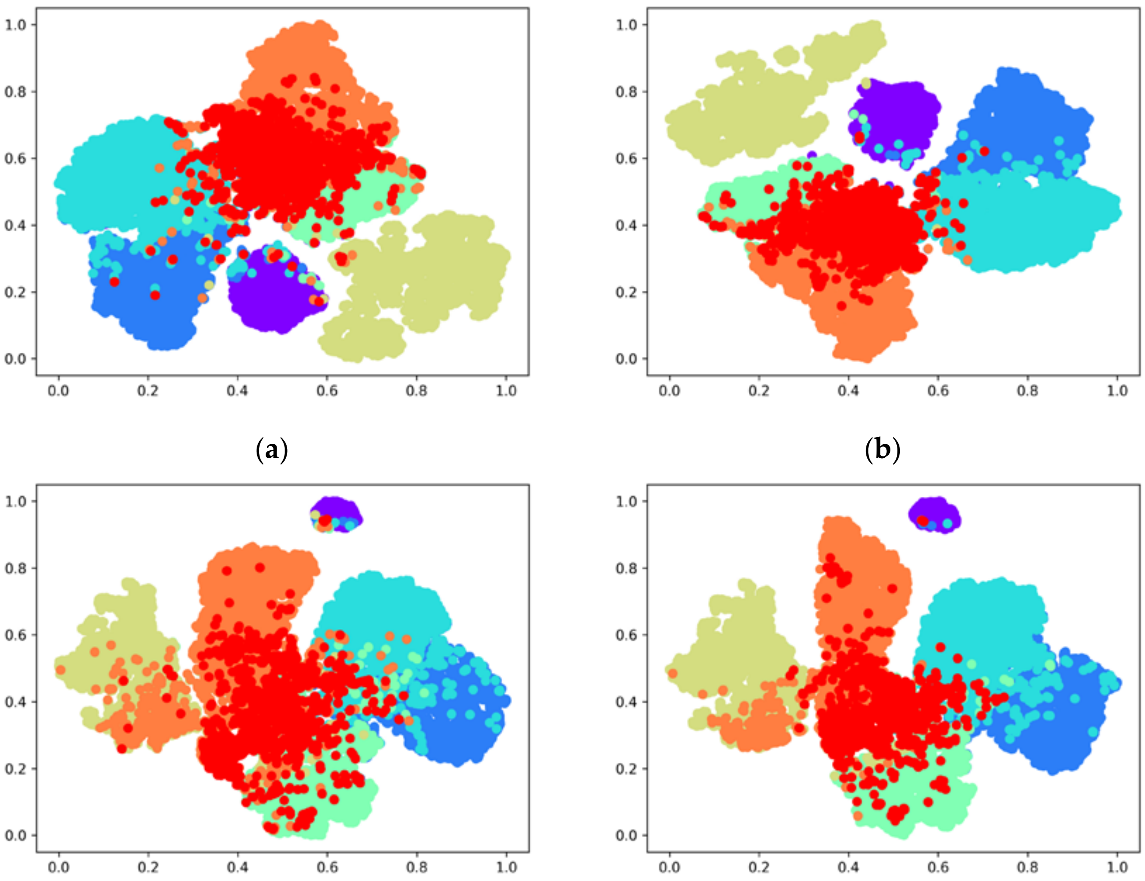
Training Set		**Epoch**					
		10	**20**	**30**	**40**	**50**	**82**
Initial	OA	0.8153	0.8459	0.8644	0.8798	0.8984	0.9068
Denoised by TCCV	OA	0.8342	0.8709	0.8804	0.8979	0.9079	0.9288
Denoised by EViTCC-DP	OA	0.8420	0.8727	0.8875	0.9052	0.9129	0.9348

Table 4. Comparison of the classification accuracy of the Guantan dataset before and after sample denoising (no improvement observed after epoch 84).

Training Set		**Epoch**					
		10	**20**	**30**	**40**	**60**	**84**
Initial	OA	0.8063	0.8379	0.8595	0.8752	0.8906	0.9165
Denoised by TCCV	OA	0.8389	0.8711	0.8888	0.9010	0.9093	0.9333
Denoised by EViTCC-DP	OA	0.8441	0.8771	0.8953	0.9089	0.9197	0.9421

The comparison results indicate that EViTCC-DP achieves a higher accuracy with fewer iterations, highlighting its advantage in mitigating noise interference. Although EViTCC-DP only slightly outperforms TCCV in accuracy (by 0.6% for the Nantong dataset and 0.88% for the Guantan dataset), TCCV removes RTS, which is crucial for enhancing model performance. This emphasizes the superior ability of our method in identifying relevant instances within the dataset and in effectively removing noise.

In order to enhance the visualization of the denoising effect, t-SNE [60] analysis was employed to assess the discriminability of high-dimensional sample features before and after denoising, as depicted in Figure 21. Each dot represents a distinct class, with increased clustering among dots of the same class indicating reduced noise levels. As illustrated in Figure 21, samples exhibit a more compact distribution within each class after denoising, leading to improved homogeneity and facilitating model training.



(**a**) (**b**)

(**c**) (**d**)

Figure 21. Visualization of sample feature distribution before and after denoising. Parts (**a**,**c**): results before denoising for the Nantong and Guantan datasets. Parts (**b**,**d**): results after denoising for the Nantong and Guantan datasets. (Parameter P = 20; denoising ratio set to 5%). Different colors represent different classes of the scene samples.

4.3. Introducing Representative Training Samples

The crucial role of RTS has been widely acknowledged by researchers [61–65]. The proposed transformation procedure of RTS is illustrated in Figure 22. Firstly, the detected mislabeled samples were fed into the ViT model to extract depth features. Subsequently, using the k-means clustering algorithm, these noisy samples were supervised and transformed into RTS.

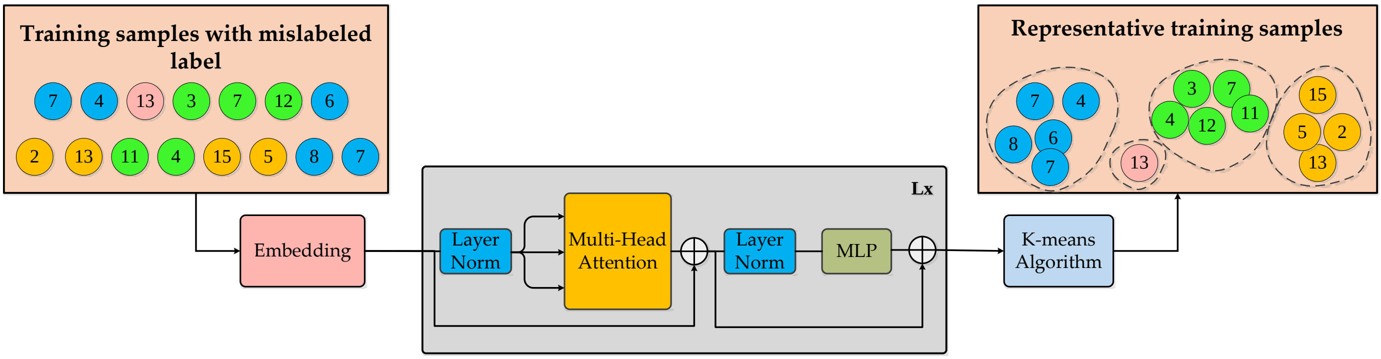


Figure 22. The proposed transformation procedure of representative training samples.

We utilized artificially generated reference data [see Figure 17b,d] to further visually assess the model's detection accuracy. Table 5 presents the detection outcomes of the model on the two datasets. Excluding RTS, the model achieved precision and recall rates of 88.22% and 91.41%, respectively, in change detection for the Nantong dataset, as well as achieving precision and recall rates of 89.97% and 92.13%, respectively, for the Guantan dataset. Interestingly, including RTS improved precision by 2.11% and 1.09% on the Nantong dataset and Guantan dataset, respectively, while maintaining consistent recall due to its high-value information, which is conducive to change detection.

Table 5. Comparison of detection accuracy results with and without the inclusion of representative training samples.

Training Set	%	Nantong Dataset	Guantan Dataset
Denoised (excluding RTS)	Precision Recall	88.22 91.26	89.97 92.13
Denoised (including RTS)	Precision Recall	90.33 91.41	91.06 92.38

5. Conclusions

The integration of RS images and deep learning in vector polygon change detection presents a promising solution to streamline data collection and processing, thereby mitigating challenges associated with spatial data updating. Despite notable progress in leveraging vector data for RS image change detection driven by advancements in artificial intelligence, geospatial big data, and RS interpretation, persistent challenges remain. This paper introduces a novel change detection method for land cover vector polygons using high-resolution RS images and deep learning techniques. The method is tailored to update vector polygons using up-to-date RS images and deep learning, distinguishing itself from previous approaches by incorporating vector polygons as prior knowledge into the detection process, thus automating sample construction. The change detection process involves four phases: segmentation, denoising, classification, and detection. First, RS images are combined with land cover vector polygons to automatically generate an initial sample set using boundary constraints. Subsequently, the class-constrained DP clustering algorithm denoises the initial set while converting filtered samples into RTS using the k-mean algorithm to construct a high-quality multi-scale sample set. Finally, the improved AMCT-Net model classifies this sample set, which is followed by detecting changed vector polygons by combining change rules. The effectiveness of our proposed method was validated and analyzed using real data from two typical regions in Jiangsu Province.

The main conclusions of this paper are as follows:

- The boundary constraint segmentation method utilized in this study accurately segments the boundaries of ground objects, while the adaptive cropping strategy facilitates comprehensive sampling within vector polygons, minimizing confusion among ground objects in the generated samples. The proposed sample denoising method, EViTCC-DP, significantly enhances model accuracy, leading to a 2.80% and 2.56% improvement in OA on the Nantong and Guantan datasets, respectively.
- To enhance classification performance, we introduced multi-scale modules and attention mechanisms to construct a novel model, AMCT-Net. This network combines the advantages of CNNs and Transformers, enabling the extraction of more discriminative features. Experimental results on the two datasets demonstrate the effectiveness of the proposed method, with the accuracy of the AMCT-Net model reaching 91.34% and 93.51%, respectively, surpassing that of other advanced models.
- Visual interpretation results demonstrate the significance of RTS in enhancing detection accuracy. The introduction of RTS yields a 2.11% and 1.09% increase in change detection accuracy for the Nantong and Guantan datasets, respectively. Our approach enables the swift construction of a high-quality multi-scale scene sample set incor-

porating RTS, requiring minimal manual intervention. Furthermore, in conjunction with designed change decision rules featuring adjustable parameters and improved applicability, the change detection method outlined in this paper effectively identifies changed vector polygons, offering clear advantages over traditional manual vector polygons updating methods.

Our future research will focus on incorporating prompt learning into our innovative change detection method to develop a comprehensive methodology for detecting unauthorized land encroachment, evaluating spatial database quality, and monitoring urban development. Moreover, we will carry out an additional investigation that will explore the incorporation of sample transfer learning into our proposed method to meet the increased demands for high-resolution RS image change detection across various sensors, time periods, and resolutions.

Author Contributions: All of the authors made significant contributions to this work. S.W. and Y.Z. conceived and designed the experiments; S.W. and N.Z. performed the experiments; S.W., H.Z. and X.Z. analyzed the data; W.L. and Y.L. contributed reagents/materials/analysis tools; S.W. wrote this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the State Key Program of National Natural Science Foundation of China under grant U22A20569 and in part by the Research Project of Jiangsu Provincial Geological Bureau under grant 2022KY15.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Sefrin, O.; Riese, F.M.; Keller, S. Deep learning for land cover change detection. *Remote Sens.* **2020**, *13*, 78. [CrossRef]
2. Zhang, C.; Yue, P.; Tapete, D.; Jiang, L.; Shangguan, B.; Huang, L.; Liu, G. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 183–200. [CrossRef]
3. Jiang, M.; Zhang, X.; Sun, Y.; Feng, W.; Gan, Q.; Ruan, Y. AFSNet: Attention-guided full-scale feature aggregation network for high-resolution remote sensing image change detection. *GISci. Remote Sens.* **2022**, *59*, 1882–1900. [CrossRef]
4. Ning, X.; Zhang, H.; Zhang, R.; Huang, X. Multi-stage progressive change detection on high resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2024**, *207*, 231–244. [CrossRef]
5. Dong, S.; Wang, L.; Du, B.; Meng, X. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 53–69. [CrossRef]
6. Deng, X.; Huang, J.; Rozelle, S.; Zhang, J.; Li, Z. Impact of urbanization on cultivated land changes in China. *Land Use Policy* **2015**, *45*, 1–7. [CrossRef]
7. Lv, Z.; Huang, H.; Sun, W.; Jia, M.; Benediktsson, J.A.; Chen, F. Iterative Training Sample Augmentation for Enhancing Land Cover Change Detection Performance with Deep Learning Neural Network. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–14. [CrossRef] [PubMed]
8. Wu, C.; Du, B.; Zhang, L. Fully Convolutional Change Detection Framework with Generative Adversarial Network for Unsupervised, Weakly Supervised and Regional Supervised Change Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 9774–9788. [CrossRef]
9. Kulinan, A.S.; Cho, Y.; Park, M.; Park, S. Rapid wildfire damage estimation using integrated object-based classification with auto-generated training samples from Sentinel-2 imagery on Google Earth Engine. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *126*, 103628. [CrossRef]
10. Sun, B.; Zhang, Y.; Zhou, Q.; Zhang, X. Effectiveness of Semi-Supervised Learning and Multi-Source Data in Detailed Urban Landuse Mapping with a Few Labeled Samples. *Remote Sens.* **2022**, *14*, 648. [CrossRef]
11. Zhao, Y.; Zhang, X.; Feng, W.; Xu, J. Deep Learning Classification by ResNet-18 Based on the Real Spectral Dataset from Multispectral Remote Sensing Images. *Remote Sens.* **2022**, *14*, 4883. [CrossRef]
12. Cui, Y.; Yang, G.; Zhou, Y.; Zhao, C.; Pan, Y.; Sun, Q.; Gu, X. AGTML: A novel approach to land cover classification by integrating automatic generation of training samples and machine learning algorithms on Google Earth Engine. *Ecol. Indic.* **2023**, *154*, 110904. [CrossRef]
13. Cao, Y.; Huang, X. A full-level fused cross-task transfer learning method for building change detection using noise-robust pretrained networks on crowdsourced labels. *Remote Sens. Environ.* **2023**, *284*, 113371. [CrossRef]
14. Li, J.; Huang, X.; Chang, X. A label-noise robust active learning sample collection method for multi-temporal urban land-cover classification and change analysis. *ISPRS J. Photogramm. Remote Sens.* **2020**, *163*, 1–17. [CrossRef]

15. Xuan, F.; Dong, Y.; Li, J.; Li, X.; Su, W.; Huang, X.; Huang, J.; Xie, Z.; Li, Z.; Liu, H.; et al. Mapping crop type in Northeast China during 2013–2021 using automatic sampling and tile-based image classification. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *117*, 103178. [CrossRef]
16. Zhang, L.; Hu, X.; Zhang, M.; Shu, Z.; Zhou, H. Object-level change detection with a dual correlation attention-guided detector. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 147–160. [CrossRef]
17. Gu, F.; Xiao, P.; Zhang, X.; Li, Z.; Muhtar, D. FDFF-Net: A Full-Scale Difference Feature Fusion Network for Change Detection in High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 2161–2172. [CrossRef]
18. Peng, Y.; He, J.; Yuan, Q.; Wang, S.; Chu, X.; Zhang, L. Automated glacier extraction using a Transformer based deep learning approach from multi-sensor remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 303–313. [CrossRef]
19. Jiang, M.; Su, Y.; Gao, L.; Plaza, A.; Zhao, X.-L.; Sun, X.; Liu, G. GraphGST: Graph Generative Structure-Aware Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5504016. [CrossRef]
20. Chen, K.; Zou, Z.; Shi, Z. Building Extraction from Remote Sensing Images with Sparse Token Transformers. *Remote Sens.* **2021**, *13*, 4441. [CrossRef]
21. Noman, M.; Fiaz, M.; Cholakkal, H.; Narayan, S.; Anwer, R.M.; Khan, S.; Khan, F.S. Remote Sensing Change Detection with Transformers Trained from Scratch. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]
22. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection With Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5607514. [CrossRef]
23. Roy, S.K.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; Chanussot, J. Multimodal Fusion Transformer for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 6826. [CrossRef]
24. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A. Vision Transformers for Remote Sensing Image Classification. *Remote Sens.* **2021**, *13*, 516. [CrossRef]
25. Jiang, M.; Chen, Y.; Dong, Z.; Liu, X.; Zhang, X.; Zhang, H. Multiscale Fusion CNN-Transformer Network for High-Resolution Remote Sensing Image Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 5280–5293. [CrossRef]
26. Wang, G.; Li, B.; Zhang, T.; Zhang, S. A Network Combining a Transformer and a Convolutional Neural Network for Remote Sensing Image Change Detection. *Remote Sens.* **2022**, *14*, 2228. [CrossRef]
27. Liu, W.; Lin, Y.; Liu, W.; Yu, Y.; Li, J. An attention-based multiscale transformer network for remote sensing image change detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 599–609. [CrossRef]
28. Song, F.; Zhang, S.; Lei, T.; Song, Y.; Peng, Z. MSTDSNet-CD: Multiscale Swin Transformer and Deeply Supervised Network for Change Detection of the Fast-Growing Urban Regions. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6508505. [CrossRef]
29. Shao, M.; Li, K.; Wen, Y.; Xie, X. Large-scale Foundation Model enhanced Few-shot Learning for Open-pit Minefield Extraction. *IEEE Geosci. Remote Sens. Lett.* **2024**, 1–1. [CrossRef]
30. 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.* **2023**, *61*, 5612822. [CrossRef]
31. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z. RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4701117. [CrossRef]
32. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4015–4026.
33. Zhang, C.; Wu, R.; Li, G.; Cui, W.; Jiang, Y. Change detection method based on vector data and isolation forest algorithm. *J. Appl. Remote Sens.* **2020**, *14*, 024516. [CrossRef]
34. Wei, D.; Hou, D.; Zhou, X.; Chen, J. Change Detection Using a Texture Feature Space Outlier Index from Mono-Temporal Remote Sensing Images and Vector Data. *Remote Sens.* **2021**, *13*, 3857. [CrossRef]
35. Shi, J.; Liu, W.; Zhu, Y.; Wang, S.; Hao, S.; Zhu, C.; Shan, H.; Li, E.; Li, X.; Zhang, L. Fine Object Change Detection Based on Vector Boundary and Deep Learning with High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4094–4103. [CrossRef]
36. Guo, Z.; Liu, W.; Xu, J.; Li, E.; Li, X.; Zhang, L.; Zhang, J. Land type authenticity check of vector patches using a self-trained deep learning model. *Int. J. Remote Sens.* **2022**, *43*, 1226–1252. [CrossRef]
37. Zhang, H.; Liu, W.; Niu, H.; Yin, P.; Dong, S.; Wu, J.; Li, E.; Zhang, L.; Zhu, C. Land Cover Change Detection Based on Vector Polygons and Deep Learning with High Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 4402218. [CrossRef]
38. Fang, H.; Guo, S.; Lin, C.; Zhang, P.; Zhang, W.; Du, P. Scene-level change detection by integrating VHR images and POI data using a multiple-branch fusion network. *Remote Sens. Lett.* **2023**, *14*, 808–820. [CrossRef]
39. Tu, B.; Zhang, X.; Kang, X.; Zhang, G.; Li, S. Density Peak-Based Noisy Label Detection for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 1573–1584. [CrossRef]
40. Tu, B.; Zhang, X.; Kang, X.; Wang, J.; Benediktsson, J.A. Spatial Density Peak Clustering for Hyperspectral Image Classification with Noisy Labels. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5085–5097. [CrossRef]
41. Algan, G.; Ulusoy, I. Image classification with deep learning in the presence of noisy labels: A survey. *Knowl.-Based Syst.* **2021**, *215*, 106771. [CrossRef]

42. Liu, S.; Zheng, Y.; Du, Q.; Bruzzone, L.; Samat, A.; Tong, X.; Jin, Y.; Wang, C. A Shallow-to-Deep Feature Fusion Network for VHR Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5410213. [CrossRef]
43. Li, G.; Ning, X.; Zhang, H.; Wang, H.; Hao, M. Remote sensing monitoring for the non-agriculturalization of cultivated land guided by the third national land survey results data. *Sci. Surv. Mapp.* **2022**, *47*, 149–159. [CrossRef]
44. Kang, X.; Duan, P.; Xiang, X.; Li, S.; Benediktsson, J.A. Detection and Correction of Mislabeled Training Samples for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 5673–5686. [CrossRef]
45. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]
46. Chen, Y.; Ma, S.; Chen, X.; Ghamisi, P. Hyperspectral data clustering based on density analysis ensemble. *Remote Sens. Lett.* **2017**, *8*, 194–203. [CrossRef]
47. Krishna, K.; Narasimha Murty, M. Genetic K-means algorithm. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **1999**, *29*, 433–439. [CrossRef]
48. Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492–1496. [CrossRef]
49. Li, Z.; Li, E.; Samat, A.; Xu, T.; Liu, W.; Zhu, Y. An Object-Oriented CNN Model Based on Improved Superpixel Segmentation for High-Resolution Remote Sensing Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4782–4796. [CrossRef]
50. Yang, G.; Yu, W.; Yao, X.; Zheng, H.; Cao, Q.; Zhu, Y.; Cao, W.; Cheng, T. AGTOC: A novel approach to winter wheat mapping by automatic generation of training samples and one-class classification on Google Earth Engine. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102446. [CrossRef]
51. Zhang, C.; Dong, J.; Xie, Y.; Zhang, X.; Ge, Q. Mapping irrigated croplands in China using a synergetic training sample generating method, machine learning classifier, and Google Earth Engine. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102888. [CrossRef]
52. Zhao, Z.; Luo, Z.; Li, J.; Wang, K.; Shi, B. Large-scale fine-grained bird recognition based on a triplet network and bilinear model. *Appl. Sci.* **2018**, *8*, 1906. [CrossRef]
53. Wang, W.; Tan, X.; Zhang, P.; Wang, X. A CBAM Based Multiscale Transformer Fusion Approach for Remote Sensing Image Change Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 6817–6825. [CrossRef]
54. Shi, J.; Liu, W.; Shan, H.; Li, E.; Li, X.; Zhang, L. Remote Sensing Scene Classification Based on Multibranch Fusion Attention Network. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 3001505. [CrossRef]
55. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
56. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722.
57. Li, J.; Lin, D.; Wang, Y.; Xu, G.; Zhang, Y.; Ding, C.; Zhou, Y. Deep Discriminative Representation Learning with Attention Map for Scene Classification. *Remote Sens.* **2020**, *12*, 1366. [CrossRef]
58. Wu, C.; Du, B.; Cui, X.; Zhang, L. A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion. *Remote Sens. Environ.* **2017**, *199*, 241–255. [CrossRef]
59. Jiang, B.; Wang, Z.; Wang, X.; Zhang, Z.; Chen, L.; Wang, X.; Luo, B. VcT: Visual change Transformer for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 2005214. [CrossRef]
60. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
61. Wen, Y.; Li, X.; Mu, H.; Zhong, L.; Chen, H.; Zeng, Y.; Miao, S.; Su, W.; Gong, P.; Li, B.; et al. Mapping corn dynamics using limited but representative samples with adaptive strategies. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 252–266. [CrossRef]
62. Jia, W.; Pang, Y.; Tortini, R. The influence of BRDF effects and representativeness of training data on tree species classification using multi-flightline airborne hyperspectral imagery. *ISPRS J. Photogramm. Remote Sens.* **2024**, *207*, 245–263. [CrossRef]
63. An, Y.; Yang, L.; Zhu, A.X.; Qin, C.; Shi, J. Identification of representative samples from existing samples for digital soil mapping. *Geoderma* **2018**, *311*, 109–119. [CrossRef]
64. Shrivastava, S.; Zhang, X.; Nagesh, S.; Parchami, A. DatasetEquity: Are All Samples Created Equal? In the Quest for Equity within Datasets. *arXiv* **2023**, arXiv:2308.09878.
65. Du, J.; Zhou, Y.; Liu, P.; Vong, C.M.; Wang, T. Parameter-Free Loss for Class-Imbalanced Deep Learning in Image Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 3234–3240. [CrossRef]

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