applied sciences

MDPI

Article

Research on the Applicability of Transformer Model in Remote-Sensing Image Segmentation

Minmin Yu [1,2] and Fen Qin [1,2,3,4,*]

[1] The College of Geography and Environment Science, Henan University, Kaifeng 475004, China
[2] Key Laboratory of Geospatial Technology for Middle and Lower Yellow River Regions, Ministry of Education, Kaifeng 475004, China
[3] Henan Technology Innovation Center of Spatio-Temporal Big Data, Henan University, Zhengzhou 450046, China
[4] Henan Industrial Technology Academy of Spatio-Temporal Big Data, Henan University, Zhengzhou 450046, China
* Correspondence: qinfen@henu.edu.cn

Abstract: Transformer models have achieved great results in the field of computer vision over the past 2 years, drawing attention from within the field of remote sensing. However, there are still relatively few studies on this model in the field of remote sensing. Which method is more suitable for remote-sensing segmentation? In particular, how do different transformer models perform in the face of high-spatial resolution and the multispectral resolution of remote-sensing images? To explore these questions, this paper presents a comprehensive comparative analysis of three mainstream transformer models, including the segmentation transformer (SETRnet), SwinUnet, and TransUnet, by evaluating three aspects: a visual analysis of feature-segmentation results, accuracy, and training time. The experimental results show that the transformer structure has obvious advantages for the feature-extraction ability of large-scale remote-sensing data sets and ground objects, but the segmentation performance of different transfer structures in different scales of remote-sensing data sets is also very different. SwinUnet exhibits better global semantic interaction and pixel-level segmentation prediction on the large-scale Potsdam data set, and the SwinUnet model has the highest accuracy metrics for KAPPA, MIoU, and OA in the Potsdam data set, at 76.47%, 63.62%, and 85.01%, respectively. TransUnet has better segmentation results in the small-scale Vaihingen data set, and the three accuracy metrics of KAPPA, MIoU, and OA are the highest, at 80.54%, 56.25%, and 85.55%, respectively. TransUnet is better able to handle the edges and details of feature segmentation thanks to the network structure together built by its transformer and convolutional neural networks (CNNs). Therefore, TransUnet segmentation accuracy is higher when using a small-scale Vaihingen data set. Compared with SwinUnet and TransUnet, the segmentation performance of SETRnet in different scales of remote-sensing data sets is not ideal, so SETRnet is not suitable for the research task of remote-sensing image segmentation. In addition, this paper discusses the reasons for the performance differences between transformer models and discusses the differences between transformer models and CNN. This study further promotes the application of transformer models in remote-sensing image segmentation, improves the understanding of transformer models, and helps relevant researchers to select a more appropriate transformer model or model improvement method for remote-sensing image segmentation.

Keywords: transformer; multihead attention; remote-sensing image segmentation; deep learning; SwinUner; TransUnet; SETRnet; visual classification

Citation: Yu, M.; Qin, F. Research on the Applicability of Transformer Model in Remote-Sensing Image Segmentation. *Appl. Sci.* **2023**, *13*, 2261. https://doi.org/10.3390/app13042261

Academic Editor: Atsushi Mase

Received: 2 January 2023
Revised: 31 January 2023
Accepted: 8 February 2023
Published: 9 February 2023

1. Introduction

The semantic segmentation of images using computers has a wide range of application scenarios in remote sensing, medicine [1], agriculture [2,3], and other fields. The semantic segmentation of remote-sensing images is an important part of processing and analyzing

remote-sensing data and is one of the most widely used areas of remote-sensing applications [4–6]. The accurate and rapid acquisition of remote-sensing image classification is important for urban management, resource investigation, environmental monitoring, natural disaster assessment, and military reconnaissance. This is because it provides managers with a source of information for more-robust decision-making [7,8]. For example, Misbah et al., used remote sensing to detect nitrogen, phosphorus, and potassium elements in widely grown crops in Africa for the purpose of protecting the environment while increasing food production [9]. Sataer et al., analyzed the remotely sensed images of Miami Park cliffs at the edge of East Lake Michigan to study the factors of park cliff deformation and prevent disasters from cliff landslides [10]. However, the rich and complex features in remote-sensing images have been a challenge for segmentation.

Early semantic segmentation methods rely mainly on manual visual interpretation to classify remote-sensing images, which not only is time-consuming and laborious but also relies heavily on experience for classification accuracy. As the resolution of remote-sensing images continues to develop, image-element-based and object-oriented semantic segmentation methods are becoming widely used [11]. The image-element-based method makes full use of mainly the spectral reflection information of remotely sensed features for classification, but it lacks a consideration of the relationship between adjacent image elements [12]. Object-oriented methods are used for classification in the framework of object-based image analysis [13], and the problems with such methods are that they are prone to noise, and the classification scale of different features is difficult to determine [14]. Owing to the improvement of computer technology, machine learning has also been applied to the research of remote-sensing image segmentation [15]. Different from the previous methods, machine learning extracts a large number of remote-sensing image features by different classifiers, further reducing the problem of using a human interference. The main methods include decision trees, support vector machines (SVMs), and random forests (RF) [16]. For example, Ujjwal et al., used a large number of different advanced support vector machines to fully learn previously unlabelable data with a view to providing guidance for the ensuing research on remote-sensing applications [17]. Du et al., used two effective methods, random forest and rotation forest, for classification in order to fully learn the texture features of polarimetric synthetic aperture radar remote-sensing images [18]. Although machine learning has significantly improved in efficiency and accuracy, this improvement applies only to its ability to extract the shallow-feature information of remote-sensing images. The efficiency and the accuracy of classification are still low when faced with complex remote-sensing images [19,20].

Deep-learning methods have powerful and fast modeling capabilities that can improve segmentation by using the spectral information and texture features of remote-sensing images [21]. In remote-sensing image semantic segmentation, using convolutional neural networks (CNNs) is a popular deep-learning method that has significantly better image semantic segmentation capability than previous methods and has been widely used in both academic and industrial fields [22]. Thanks to its excellent ability to express high-level semantic features, CNN and its derivatives have shown potential in many image semantic segmentation tasks. For example, the feature pyramid module and the attention-feature-aggregation module are combined to improve the feature-learning capability of a CNN and to accomplish the task of the semantic segmentation of high-resolution remote-sensing images [23]. The segmentation of building data from high-resolution imagery and LiDAR data use gated residual refinement networks [24], build a multichannel deep convolutional neural network to learn remote-sensing information in different bands, and further improve the segmentation of urban land-use features [25]. It can be seen that CNN often performs semantic segmentation in remote-sensing images with high spatial and spectral resolution, and it has achieved remarkable results. However, CNNs tend to perform generally in the face of different scales of feature learning, and many improved methods have been proposed, such as the spatial pyramid pooling model, the jump link structure, and the multiscale feature fusion model [26–28]. However, the segmentation results still have issues,

which are due to the inability of CNNs to fully learn contextual semantic information and retain more spatial features [29].

Recently, the transformer model has achieved excellence in the field of semantic segmentation. Compared with CNNs, this module has a more outstanding ability to learn global semantic information [30,31]. The transformer model was originally used, and achieved remarkable results, in the field of natural language processing [32]. Researchers began to apply the transformer model to the study of image semantic segmentation. The vision transformer is the first example of the transformer model applied to image classification. Although the researchers found that the classification accuracy of the transformer method was significantly better than that of the CNN method, it ultimately did not complete the image segmentation task. The segmentation transformer (SETR) is an improved model based on the vision transformer, and it has been applied to, and has performed well in, segmentation tasks [33]. Although SETR proves that the transformer is competent for the image semantic segmentation task, it comes at an expensive cost. Because of this, the Swin Transformer uses the hierarchical transformer model to obtain multiscale features and effectively reduce computational effort [34]. TransUnet learns the features of the input image with CNNs and then inputs the feature-learning results into the transformer model, effectively combining the advantages of both the CNNs and transformer models [35].

Transformers have been widely used in image semantic segmentation tasks. However, their use in the field of remote-sensing images can be improved. Key questions include the following: (1) Which method is more suitable for remote-sensing segmentation? (2) How well does the transformer perform in remote-sensing images at different scales? (3) How do different transformer models perform in the face of high-spatial resolution and the multispectral resolution of remote-sensing images? To address these issues, three-channel and four-channel remote-sensing images are used as the data set in this study, in which the NIR band data are added. By comparing different transformer models from the perspective of segmentation accuracy and training time, the results of this study are beneficial to the selection and understanding of transformer models and provide a reference for future researchers with a view to promoting the development of fine segmentation tasks for remote-sensing images.

2. Methods

2.1. Transformer Model

The transformer model, a network architecture first proposed by Vaswani et al., in 2017 [36], is a network that eschews recursion and convolution and is based entirely on attention mechanisms, as shown in Figure 1.

Specifically, the transformer model is still a network structure from encoder to decoder, but with the abandonment of recursion and convolution, a multihead self-attention mechanism is added to the encoder and decoder modules, respectively. The multihead self-attention mechanism is a key component of the transformer model in that the multihead attention mechanism is able to capture remote dependencies between elements and to encode interactions between sequential tokens. As shown in Figure 2b, the multihead self-attention mechanism allows the model to jointly attend to information from different subspaces at different locations. The objective of the multihead attention mechanism is to simultaneously perform multiple parallel attention functions. A single attention (Figure 2a) function can be represented as a function consisting of a query and a set of keys and values corresponding to an output, where the query, key, value, and output are all represented as vectors. The input is composed of a query and key with dimension d_k and a value with dimension d_v. The dot product of the query is calculated by using all the keys divided by $\sqrt{d_k}$, and the weights of the values are obtained by the function. During the calculation, the attention function of a set of queries is simultaneously computed, which is defined

as a matrix Q. The keys and values are defined as matrices K and V, respectively, and the attention function can be expressed as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

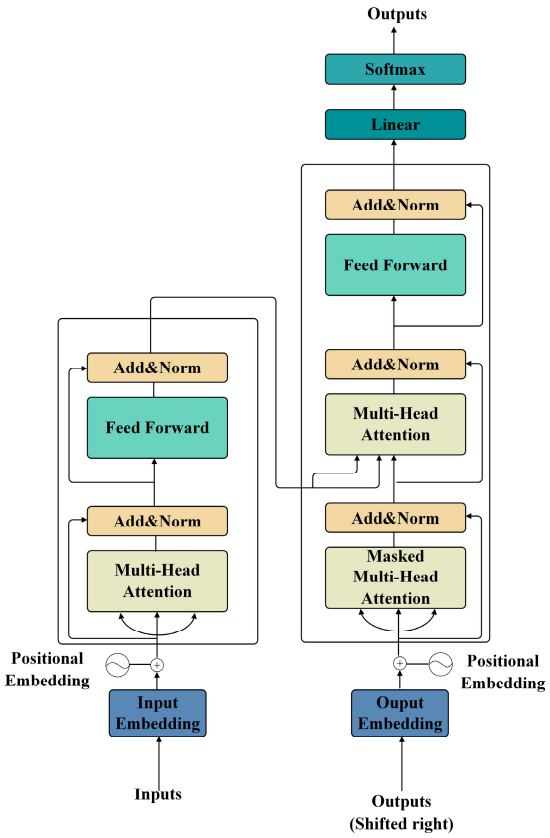

Figure 1. Transformer model structure.

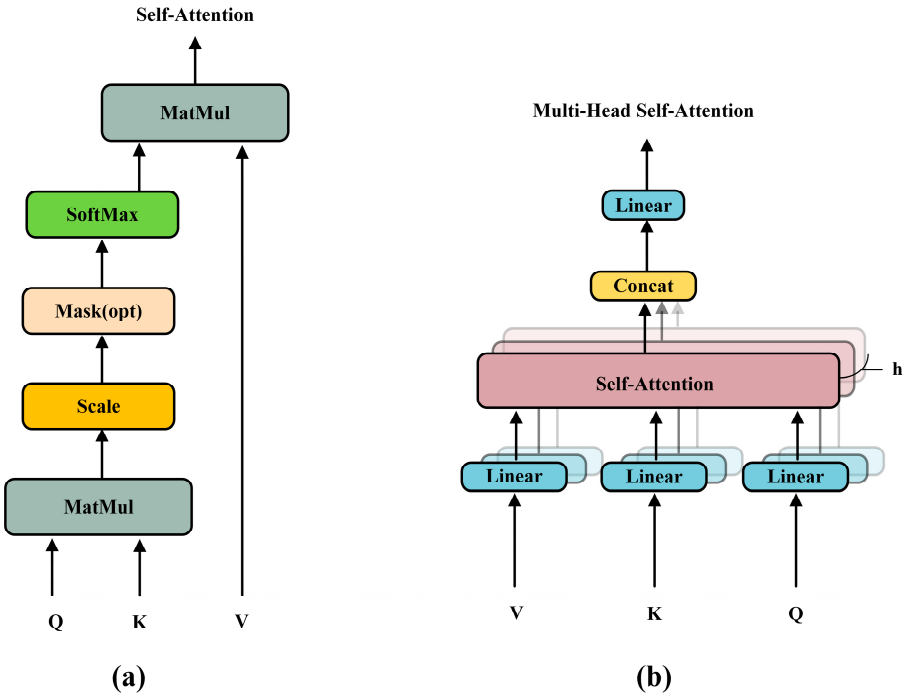

Figure 2. (**a**) Self-attention structure, (**b**) multihead attention structure.

The multihead self-attention module can acquire information from different representation subspaces at different locations. This cannot be done with single-head attention. The output of the multihead attention module can be expressed as follows:

$$MultiHead(Q, K, V) = Concat(\ head_1, \ldots,\ head_h)W^O \tag{2}$$

$$wherehead_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{3}$$

The multihead self-attention mechanism plays a crucial role in the transformer model that can not only improve the efficiency of remote-sensing image classification but also more accurately acquire the global and local features of remote-sensing images. However, if the network does not contain an attention layer, the model-based network cannot be implemented unless the network is changed, which defeats the original purpose of experimenting with high efficiency and accuracy.

2.2. SETRnet (Segmentation Transformer)

The segmentation transformer SETRnet is the first representative model of the vision-transformer-based semantic segmentation proposed by Zheng et al., in 2021 [33]. The structure of SETRnet is shown in Figure 3.

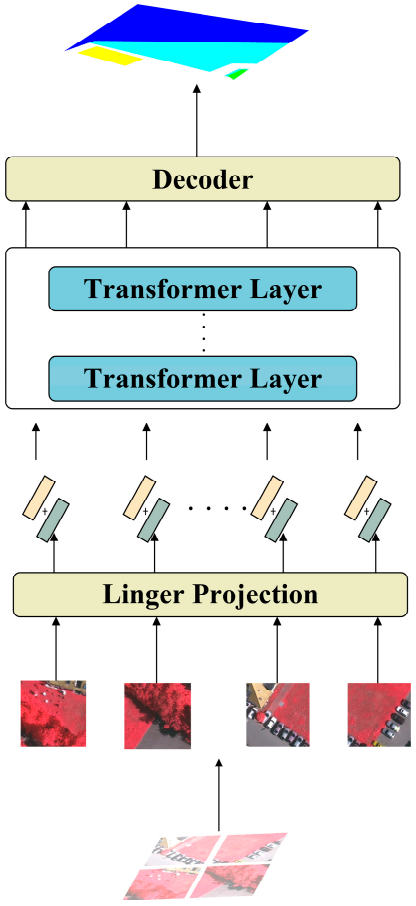

Figure 3. SETRnet structure.

SETRnet abandons the stacked convolutional feature-extraction method in the encoded layer and instead uses the transformer-only feature-extraction method. In the model, the images are first sliced, and then all the two-dimensional image slices are considered as a one-dimensional sequence and fed into the network as a whole. The input one-dimensional series will become a one-dimensional feature-embedding series. In each layer, the input of attention consists of a query, key, and value triad computed by $Z^{l-1} \in \mathbb{R}^{L \times C}$ (where L is

the sequence length and C is the hidden channel size). The computed query, key, and value triad can be expressed as follows:

$$query = Z^{l-1}w_Q,\ key = Z^{l-1}w_K,\ value = Z^{l-1}w_V \tag{4}$$

where W_Q, W_K, $W_V \in \mathbb{R}^{C\times d}$ are the learnable parameters of the 3 linear projection layers and d is the dimensionality of (query, key, value). Then the attention function of SETRnet can be expressed as follows:

$$SA\left(Z^{l-1}\right) = Z^{l-1} + softmax\left(\frac{Z^{l-1}W_Q(ZW_K)^\top}{\sqrt{d}}\right)\left(Z^{l-1}W_V\right) \tag{5}$$

The output of SETRnet's multihead self-attention (*MSA*) module is converted by an multilayer perceptron (*MLP*) module with residual jumps; the structure is shown in Figure 4.

$$Z^l = MSA\left(Z^{l-1}\right) + MLP\left(MSA\left(Z^{l-1}\right)\right) \in \mathbb{R}^{L\times C} \tag{6}$$

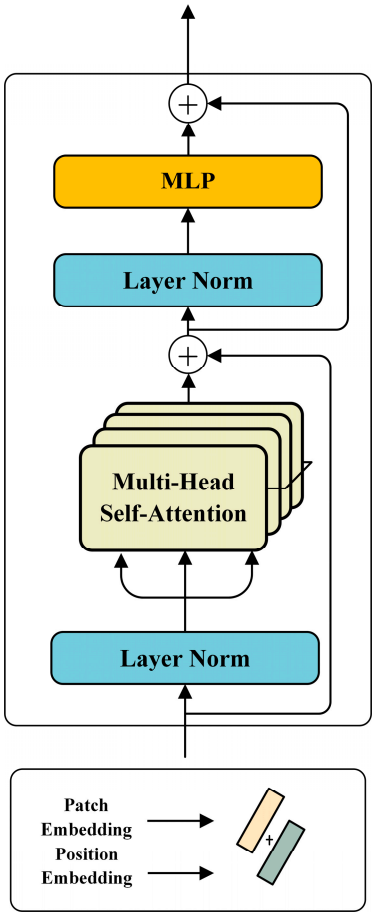

Figure 4. SETRnet's transformer layer structure.

SETRnet's decoder is designed with three methods, namely plain upsampling (naïve), progressive upsampling (PUP), and multilevel feature fusion (MLA). SETRnet constructs a new semantic segmentation model from a new perspective. Compared with the traditional semantic segmentation model, SETRnet models the global context in each layer of the encoder by using the transformer as the encoder, which effectively obtains the global context information and removes the semantic segmentation network's dependence on convolution.

2.3. SwinUnet

SwinUnet is a transformer-based network proposed by Cao et al., in 2021 [37]. SwinUnet consists of an encoder, a channel, a decoder, and a jump connection, as shown in Figure 5.

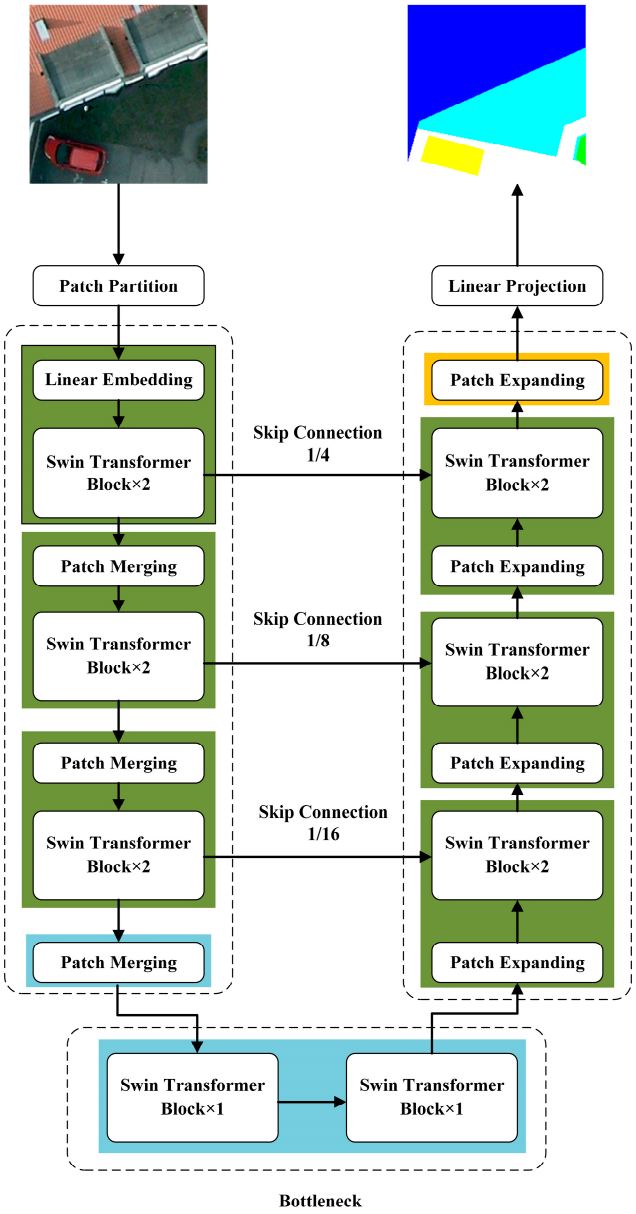

Figure 5. SwinUnet structure.

Inspired by the Unet network, SwinUnet is designed with a symmetric decoder with a patch extension layer based on the Swin Transformer. Unlike the traditional multihead self-attention (*MSA*), the Swin Transformer module is built based on a shifted window. Figure 6 shows the structure of the Swin Transformer module, which consists of LayerNorm (*LN*) layers, a multihead self-attention module, a residual connection, and a two-layer MLP with Gaussian error linear units (GELUs) nonlinearity. The window-based multihead self-attention module (*W-MSA*) and the translational window-based multihead self-attention (*SW-MSA*) module are used in two consecutive Swin Transformer modules, respectively. For such a modular composition, the continuous Swin Transformer can be expressed as follows:

$$\hat{z}^l = W - MSA\left(LN\left(z^{l-1}\right)\right) + z^{l-1} \quad (7)$$

$$z^l = MLP\left(LN\left(\hat{z}^l\right)\right) + \hat{z}^l \tag{8}$$

$$\hat{z}^{l+1} = SW - MSA\left(LN\left(z^l\right)\right) + z^l \tag{9}$$

$$z^{l+1} = MLP\left(LN\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \tag{10}$$

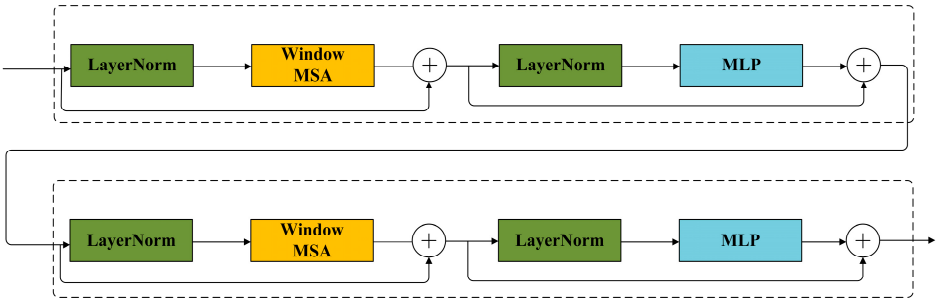

Figure 6. Swin transformer block.

$\hat{z}^l$ and z^l denote the output of the (S)W-MSA module and the MLP module of the lth block, respectively. Then the attention function can be expressed as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{11}$$

where Q, K, and $V \in \mathbb{R}^{M^2 \times d}$ denote the matrix of queries, keys, and values; M^2 and d denote the number of patches in the window and the dimension of the query or key, respectively; and B comes from the bias matrix $\hat{B} \in \mathbb{R}^{(2M-1)\times(2M+1)}$.

SwinUnet successfully puts the transformer module into the encoder and decoder, and together with the jump connection of the Unet network, the SwinUnet network can more quickly and comprehensively acquire the global and local feature information of images, avoiding the limitation of the CNN model, which cannot acquire global and long-range feature information.

2.4. TransUnet

TransUnet is a network proposed by Chen et al., in February 2021 [35]; unlike the SwinUnet network structure based entirely on the transformer, the encoder in the TransUnet network does not use a pure transformer but instead uses a hybrid CNN-transformer model. The network structure is shown in Figure 7. TransUnet is a network proposed by Chen et al., in February 2021 [35]; unlike the SwinUnet network structure based entirely on the transformer, the encoder in the TransUnet network does not use a pure transformer but instead uses a hybrid CNN-transformer model. The network structure is shown in Figure 7.

TransUnet first uses CNN as a feature extractor to generate the input feature maps. This is when the patch-embedding module extracts 1 × 1 patches from the feature map generated by the CNN, rather than from the original map. The transformer encoder in the TransUnet network consists of the L-layer MSA and MLP modules (as shown in Figure 8), so the output of the lth layer can be expressed as follows:

$$z'_\ell = MSA(LN(z_{\ell-1})) + z_{\ell-1} \tag{12}$$

$$z^\ell = MLP\left(LN\left(z'_\ell\right)\right) + z'_\ell \tag{13}$$

where $LN(\cdot)$ denotes the layer normalization operator and z^ℓ indicates the encoded image.

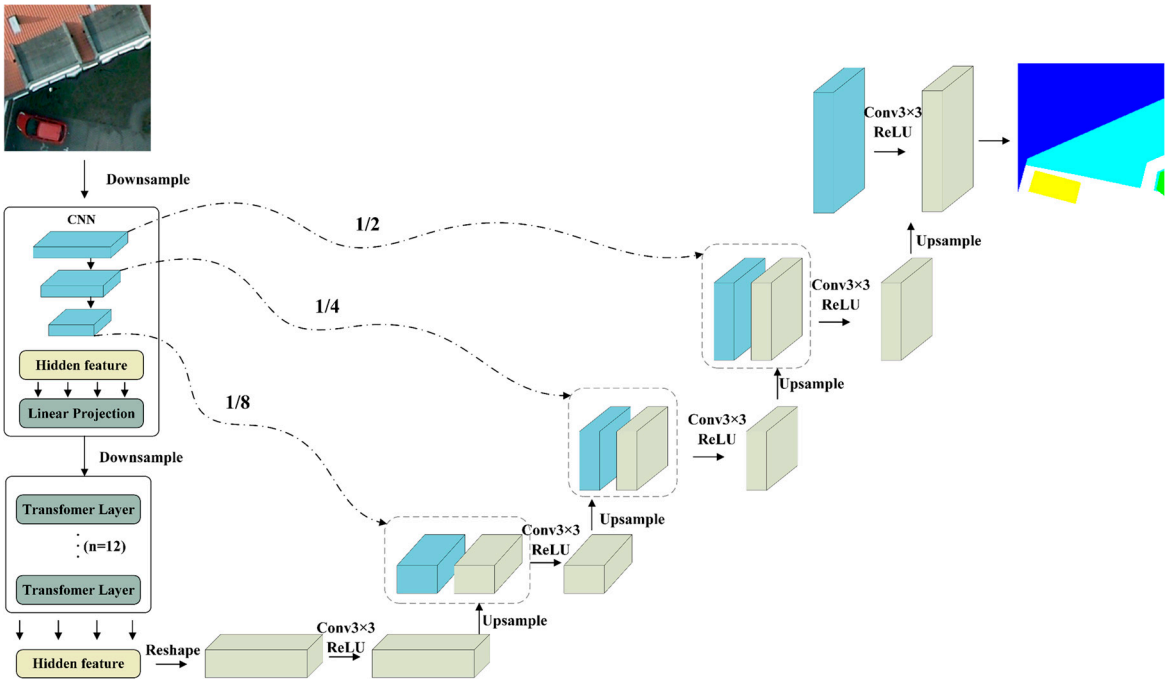

Figure 7. TransUnet structure.

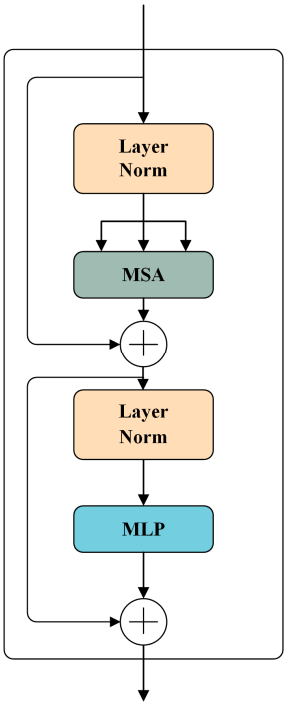

Figure 8. TransUnet's transformer layer structure.

By combining the CNN and transformer into the encoder, TransUnet not only avoids the limitations of the CNN method in acquiring remote relational features, thanks to convolutional operations, but also avoids the problem of coarse segmentation results, thanks to the transformer's excessive focus on modeling between global contexts.

2.5. Accuracy Comparison of Models

In order to verify which transformer model is more suitable for remote-sensing image segmentation, three aspects of the segmentation results, the segmentation accuracy, and

the training time of three transformer models in two data sets, namely the Vaihingen data set and Potsdam data set, were compared.

2.6. Training Time Comparison of Models

For the consumption of training time for the three transformer models, this study will compare the analysis on the basis of the training time of the three models in two data sets, namely the Vaihingen data set and Potsdam data set (Tables 1 and 2), with time in seconds.

3. Experiment and Results

3.1. Experimental Setup

In this study, the network is constructed using the Pytorch framework and Python language, on an Intel(R) Xeon(R) Gold 5218 CPU with a GeForce RTX 2080Ti GPU. The Adam optimizer is used to optimize the training process; the learning rate of each model is the same 3×10^{-4}; and the number of iterations is set to 40. The learning rate of each model is the same 3×10^{-4}; the number of epochs is set to 40; and the image size is set to 224×224.

3.1.1. Data Set

In order to better compare the segmentation effect of transformers on remote-sensing data sets at different scales, the experimental data were obtained from the state-of-the-art airborne image data sets Vaihingen data set and Potsdam data set, provided by ISPRS. The Vaihingen data set and Potsdam data set have the same type of features and different scales.

Vaihingen data set: Vaihingen is a relatively small village located in Germany with a number of detached buildings and small multistory buildings. The Vaihingen data set contains 33 remote-sensing images of different sizes covering an area of 1.38 km^2 in Vaihingen. The spatial resolution of the top image and DSM is 9 cm, and each image has four bands: near-infrared, red, green, and blue. We cropped 33 remote-sensing images of different sizes into small images with a size of 224×224, and the cropped images were divided into a training set, a validation set, and a test set according to the ratio of 8:1:1. The percentage of each type of feature in the data set is shown in Figure 9.

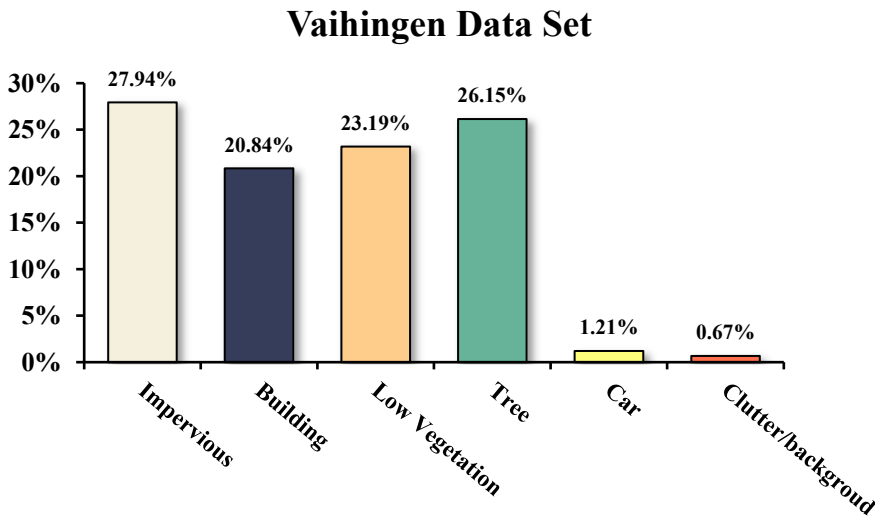

Figure 9. Proportion of each type of feature in the Vaihingen data set.

Potsdam data set: Potsdam is a city with a long history of large buildings, narrow streets, and a dense settlement structure. The Potsdam data set contains 38 remotely sensed images of the same size (6000×6000) covering an area of 3.42 km^2 in Potsdam. The spatial resolution of the top image and DSM is 5 cm, and each image has four bands: near-infrared, red, green, and blue. Similarly, we cropped 38 remote-sensing images of the same size into small images with a size of 224×224, and the cropped images were divided into a training

set, a validation set, and a test set according to the ratio of 8:1:1. The percentage of each type of feature in the data set is shown in Figure 10.

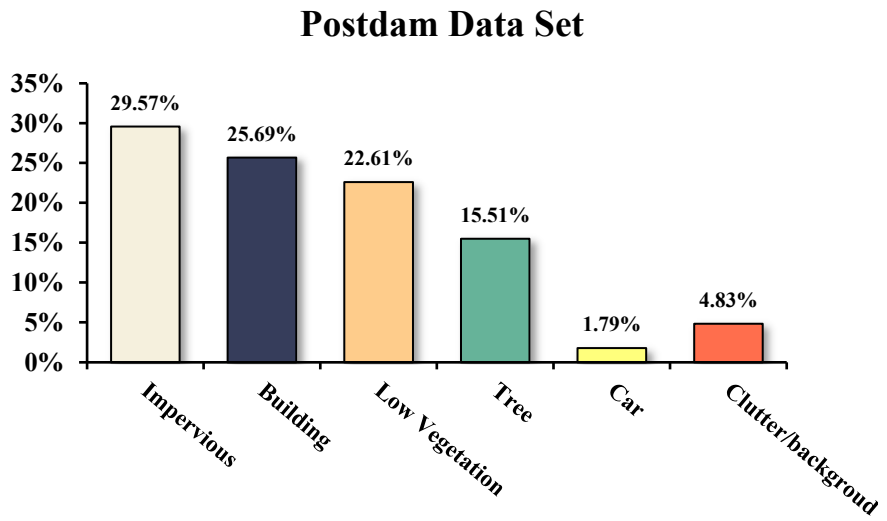

Figure 10. Proportion of each feature type in the Potsdam data set.

3.1.2. Metrics

The segmentation-extraction results of different transformer models can be evaluated on the basis of both subjective and objective aspects [38]. The subjective aspects include whether the segmentation-extraction results of the remote-sensing image features are complete and whether the segmentation edges of the features are clear and consistent. The objective aspect can be quantitatively calculated on the basis of the evaluation criteria, and afterward, the model classification accuracy can be assessed. The following criteria were used in this study mainly to evaluate the classification performance of the three transformer models for the data in the training data set.

F1 score: the F1 score is a metric to evaluate the model proposed on the basis of precision and recall, which explains the extent to which the true value overlaps with the predicted outcome pixels. The F1 score also serves as a summed average of precision and recall as a whole, as defined below:

$$F1score = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

where TP is a positive sample predicted by the model as a positive class, FP is the negative sample predicted by the model as a positive class, and FP is the negative sample predicted by the model as a positive class.

Overall classification accuracy (OA): the ratio of the number of correctly classified samples to the number of all samples can be defined as follows.

$$OA = \frac{TP + TN}{TP + FN + FP + TN} \tag{17}$$

TN is the negative sample predicted by the model as the negative class.

Mean intersection and merge ratio (*MIoU*): the MIoU is the calculation of the ratio of the intersection of two sets of true values to the merged set of predicted values, and it is a global evaluation of the image classification results, as defined below.

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN+FP+TP} \tag{18}$$

Kappa coefficient: the Kappa coefficient can be used to measure classification accuracy and also to test consistency. In practical classification problems, the Kappa coefficient is often used as an indicator to evaluate the "bias" of the model if the consistency between samples is poor. If P_o is the overall classification accuracy and if the number of real samples in each category is $a_1, a_2, \ldots, a_c$, the predicted number of samples in each category is b_1, $b_2, \ldots, b_c$, and the total number of samples is n, then P_e is the consistency error.

$$p_e = \frac{a_1 \times b_1 + a_2 \times b_2 \ldots a_c + b_c}{n \times n} \tag{19}$$

$$kappa = \frac{p_o - p_e}{1 - p_e} \tag{20}$$

3.2. Result

3.2.1. Metrics Visual Analysis of Classification Results

In order to better analyze the gap between the three transformer models in terms of segmentation details, partially cropped images in the two data sets were selected for the analysis of segmentation results in this study. The segmentation results of the Vaihingen and Potsdam data sets are shown in Figures 11 and 12, respectively.

In the first set of results from the Vaihingen data set, the SETRnet, SwinUnet, and TransUnet networks all showed missegmentation at the junction of impervious surfaces and trees. At the junction of buildings and low vegetation, SwinUnet showed fragmentation.

In the second set of results, all three networks showed errors at the junction of trees and impervious surfaces in the upper right corner, but SETRnet and SwinUnet showed more classification errors. The classification results of TransUnet were more complete overall, and the contours were more clearly continuous among the three networks.

In the third set of results, SETRnet better segmented the area compared with SwinUnet and TransUnet but misclassified at the junction of low vegetation and trees. TransUnet was the only network that correctly distinguished low vegetation from trees, while SwinUnet showed edge jaggedness.

In the fourth set of results, SETRnet and SwinUnet could not well identify the vehicles, resulting in broken classification results and distorted contours for both networks for vehicles. TransUnet well identified the vehicle contours, but there were also misclassifications at the junction of vehicles and low vegetation.

The four sets of segmentation results show that when features exist in proximity to each other, especially when such relationships exist between large-area features and small-area features, the transformer model cannot segment well to deal with the relationships between such features, and it often misclassifies small-area features into large-area features. At the same time, we can see that there are great differences between the three models for vehicle segmentation. TransUnet is the best for vehicle segmentation in terms of both vehicle profile and number, SwinUnet is the second best, and SETRnet is the worst.

The first set of results in the Potsdam data set show severe fragmentation and profile breakage in SETRnet. TransUnet misclassified within the low vegetation area, while SwinUnet was successful in identifying the low vegetation and separating the low vegetation from the impervious surface.

The second set of results indicated a clear error for TransUnet at the junction of road and background, while SETRnet and SwinUnet were classified as complete. SETRnet

showed a wavy profile in the low vegetation area with the background area, and SwinUnet showed a jagged profile.

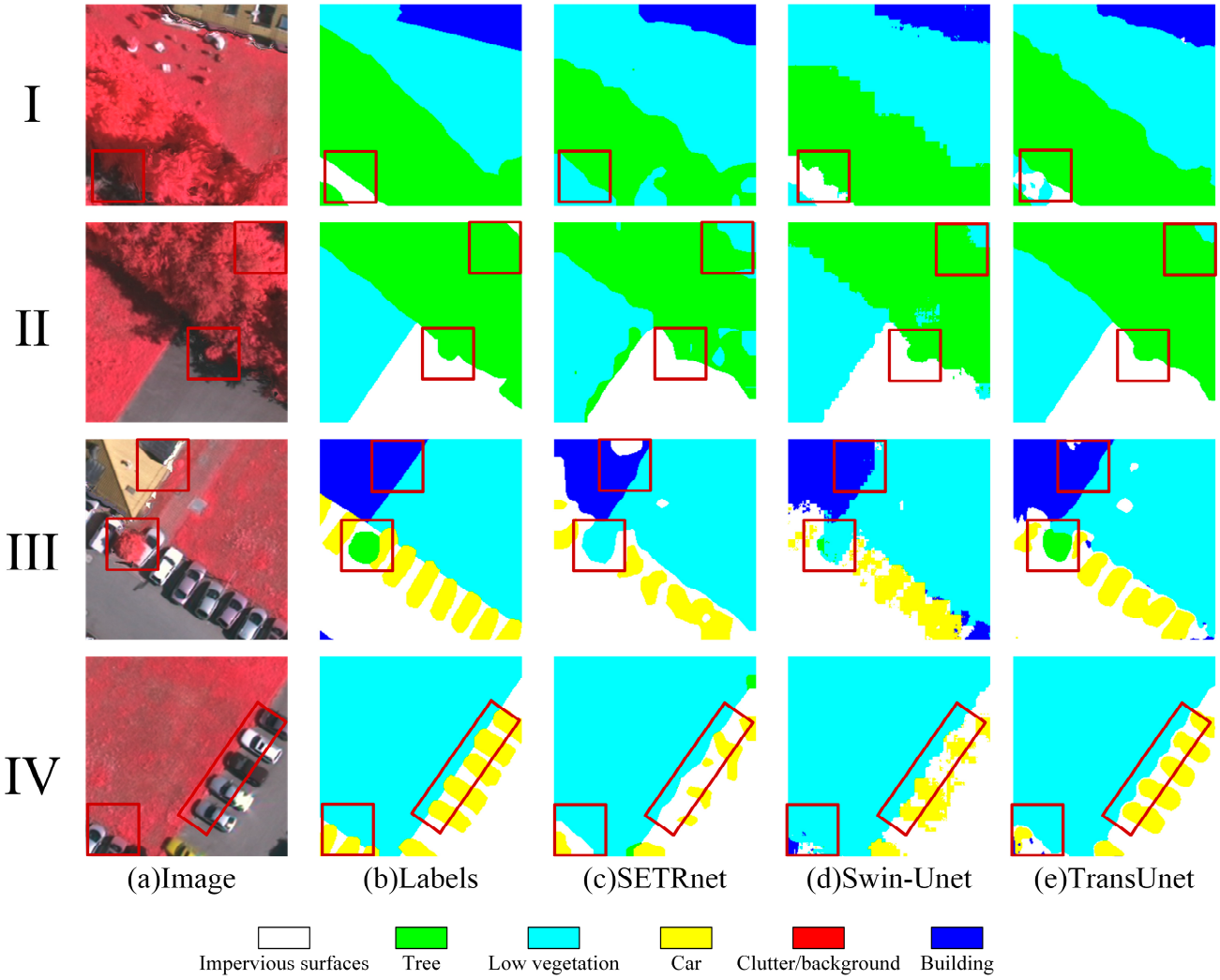

Figure 11. Comparison of transformer's segmentation results in the Vaihingen data set. The red boxes are where the differences between the three network segmentations are large.

In the third set of results, SwinUnet was more complete and accurate in its overall classification than SETRnet and TransUnet were. TransUnet did not identify vehicles well, and SETRnet misclassified impervious surface areas and low vegetation areas.

The segmentation results in the figure show that SwinUnet has the best segmentation in the large-scale Potsdam data set, followed by SETRnet and then by TransUnet. TransUnet exhibits a very different segmentation result from the Vaihingen data set, with large feature confusion and more segmentation fragmentation. It instead indicates that the method combining a CNN with a transformer is not suitable for large-scale remote-sensing data sets. In the face of large-scale data sets, a CNN combined with the transformer method face the explosion of feature information when extracting and processing feature information, owing to its strong local feature-extraction ability, which leads to a poor segmentation effect. The better segmentation effect of SwinUnet compared with SETRnet also indicates that the method of the local attention enhancement of a transformer in the face of large-scale data can effectively improve the ability of a transformer to extract global and local feature information. Among the four sets of segmentation results, different degrees of shadows exist on different features within the original image, and among the three models, only SwinUnet can better reduce the shadow effect and segment the features, while the other

two models produced more incorrect segmentation for the shadow part, but SETRnet produced less of this situation compared with TransUnet. This situation may be due to the transformer's insufficient learning ability for shadows. At the same time, we can see a significant improvement in the segmentation effect of SwinUnet and SETRnet for vehicles in the Potsdam data set. It shows that with the expansion of the data set scale, the sample size of the ground objects increases, and the segmentation accuracy of the transformer for small-scale ground objects also increases.

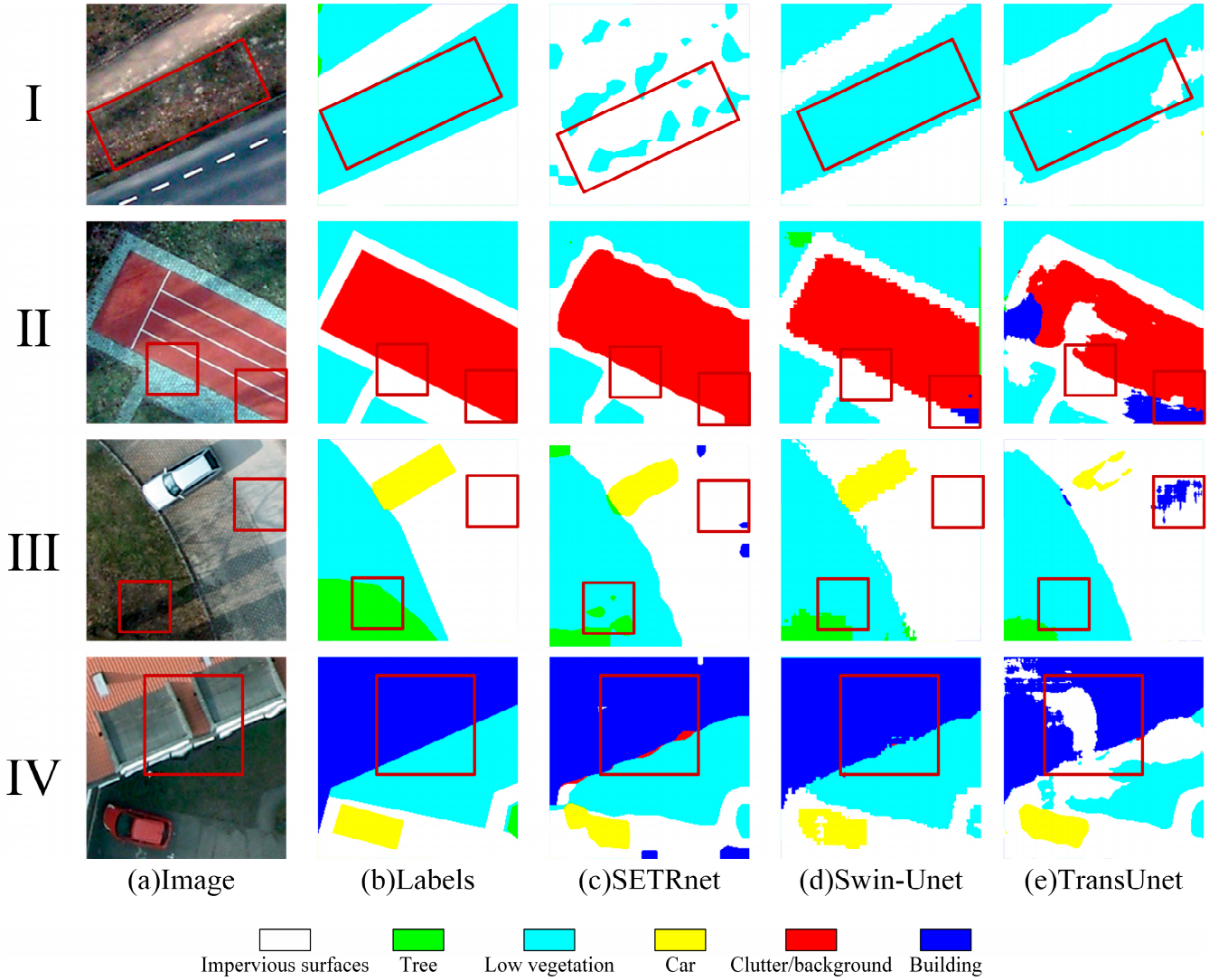

Figure 12. Comparison of transformer's segmentation results in the Potsdam data set. The red boxes are where the differences between the three network segmentations are large.

3.2.2. Training Time Comparison

Table 1 shows the training time of the three models in the Vaihingen data set. From the table, it can be seen that SwinUnet has the shortest training time, SETRnet follows, and TransUnet has the longest training time. In each epoch, SwinUnet's time is 36.79 s and 79.75 s faster compared with SETRnet and TransUnet, respectively.

Table 1. Training time of the Vaihingen data set.

Method	SETRnet	SwinUnet	TransUnet
Time(s)	4401.1	2929.7	6119.4
Average time(s)	110.03	73.24	152.99

Table 2 shows the training time of the three models for the Potsdam data set. From the table, we can see that the respective training times of SwinUnet and TransUnet are basically equal, and both are faster than SETRnet by more than 10,000 s, which is faster than 300 s per epoch, on average.

Table 2. Training time of the Potsdam data set.

Method	SETRnet	SwinUnet	TransUnet
Time(s)	38,877.1	26,405.7	27,399.3
Average time(s)	971.93	660.14	684.98

3.2.3. Accuracy Comparison of Results

The classification results in the Vaihingen data set test were collated and statistically compared with the accuracy evaluation results of the three models, as shown in Table 3. In the comparison, TransUnet had the highest precision in all categories except for the tree category, where SETRnet had the highest precision. In the comparison of recall, SETRnet and SwinUnet were the highest in the low vegetation and impervious surfaces categories, respectively. TransUnet had the highest recall in the rest of the categories. In the F1 score comparison, TransUnet had the highest F1 value in all the categories. In the Kappa comparison, SETRnet was 73.80%, SwinUnet was 77.50%, and TransUnet was the highest, at 80.54%. TransUnet in MIoU was the highest, at 56.25%, an improvement of 8.57% and 4.77% relative to SETRnet and SwinUnet, respectively. In the OA comparison, TransUnet remained the highest, at 85.55%, with SETRnet and SwinUnet at 80.50% and 83.29%, respectively.

Table 3. Evaluation table of classification results in the Vaihingen data set. The bolded numbers are the models with the best performance in terms of accuracy.

Method	Category	Precision	Recall	F1 Score	Kappa	MIoU	OA
SETRnet	Impervious surfaces	79.30%	85.46%	82.26%	73.80%	47.68%	80.50%
	Building	87.94%	83.53%	85.68%			
	Low vegetation	61.40%	**78.88%**	69.05%			
	Tree	**89.74%**	76.53%	82.61%			
	Car	70.35%	19.67%	30.75%			
	Clutter/background	0.00%	0.00%	0.00%			
SwinUnet	Impervious surfaces	83.60%	**85.58%**	84.58%	77.50%	51.48%	83.29%
	Building	88.53%	87.87%	88.20%			
	Low vegetation	67.52%	74.01%	70.62%			
	Tree	87.94%	84.68%	86.28%			
	Car	68.66%	29.83%	41.59%			
	Clutter/background	0.00%	0.00%	0.00%			
TransUnet	Impervious surfaces	**84.83%**	85.33%	**85.01%**	**80.54%**	**56.25%**	**85.55%**
	Building	**89.19%**	**89.77%**	**89.48%**			
	Low vegetation	**73.65%**	76.93%	**75.26%**			
	Tree	89.67%	**89.04%**	**89.35%**			
	Car	**84.27%**	**44.90%**	**58.59%**			
	Clutter/background	0.00%	0.00%	0.00%			

The classification results in the Potsdam data set test were collated and statistically compared with the accuracy evaluation results of the three models, as shown in Table 4. In the comparison of precision, SETRnet had the highest accuracy in the impervious surfaces category, SwinUnet had the highest accuracy in both car and clutter/background categories, and TransUnet had the highest accuracy in the building, low vegetation and tree item categories. In the recall comparison, SETRnet's accuracy was highest in the building and clutter/background categories, SwinUnet's accuracy was highest in the low vegetation and tree categories, and TransUnet's accuracy was highest in the impervious surfaces and car categories. In the F1 score comparison, SETRnet obtained the highest accuracy in the building category, SwinUnet in the impervious surfaces, low vegetation, and tree categories, and TransUnet in the car category. In the Kappa comparison, SwinUnet's accuracy was the highest, at 76.47–4.4% and 8.29% higher than SETRnet and TransUnet, respectively. SwinUnet's accuracy was still the highest in MIoU, at 63.62%, while SETRnet's and TransUnet's accuracies were 59.97% and 58%, respectively. In the comparison of OA, SwinUnet remained the highest, at 85.01%, with an improvement of 6.5% and 9.25% relative to the OA values of SETRnet and TransUnet, respectively.

Table 4. Evaluation table of classification results in the Potsdam data set. The bolded numbers are the models with the best performance in terms of accuracy.

Method	Category	Precision	Recall	F1-Score	Kappa	MIoU	OA
SETRnet	Impervious surfaces	**79.79%**	78.52%	79.15%	72.07%	59.97%	78.57%
	Building	82.60%	**91.60%**	**86.87%**			
	Low vegetation	75.70%	75.81%	75.76%			
	Tree	78.69%	75.79%	77.22%			
	Car	81.73%	70.11%	75.48%			
	Clutter/background	57.67%	**43.70%**	**49.72%**			
SwinUnet	Impervious surfaces	77.50%	90.51%	**83.50%**	**76.47%**	**63.62%**	**85.01%**
	Building	89.98%	82.71%	86.19%			
	Low vegetation	79.34%	**82.38%**	**80.83%**			
	Tree	85.91%	**78.41%**	**81.99%**			
	Car	**86.52%**	72.43%	78.85%			
	Clutter/background	**73.67%**	35.99%	48.36%			
TransUnet	Impervious surfaces	61.12%	**91.52%**	73.36%	68.18%	58.00%	75.76%
	Building	**90.13%**	68.86%	78.07%			
	Low vegetation	**83.90%**	72.67%	77.88%			
	Tree	**86.33%**	71.98%	78.50%			
	Car	83.33%	**75.43%**	**79.19%**			
	Clutter/background	57.52%	42.10%	48.62%			

3.2.4. Kappa Coefficient Effect Size Test

To demonstrate that the experimental results in the two data sets are not chance events, the Kappa coefficients of the models in the two data sets are therefore tested for effect sizes. The results are shown in Table 5. Cohen's d reflects the degree of difference between two aggregates after they are affected by something, and the larger the effect size, the greater the degree of difference. Generally, $0.2 \leq d < 0.5$ is called a small effect, $0.5 \leq d < 0.8$ is called a medium effect, and $d \geq 0.8$ is called a large effect. As can be seen in Table 3, the Cohen's d value for the Kappa coefficient in both data sets is 0.795, which is a medium effect and very close to the large effect. This also indicates that the experimental results in both data sets are not chance events and are statistically significant.

Table 5. Kappa coefficient effect scale.

Data Set	Kappa	Average Difference	Standard Value of Difference	Cohen's d
Vaihingen data set	SETRnet 73.80% SwinUnet 77.50% TransUnet 80.54%	0.05	0.063	0.794
Potsdam data set	SETRnet 72.07% SwinUnet 76.47% TransUnet 68.18%			

3.2.5. Training Process of Different Transformers

The whole training process of SwinUnet was relatively stable, the accuracy value shows a steady increase, and the loss value shows a steady decrease (Figure 13). There are fluctuations in both SETRnet and TransUnet; however, all the models converged after about the 25th epoch. Among them, fluctuations occurred at the beginning of the training for SETRnet. The final accuracy and loss values of SETRnet were the worst. TransUnet showed optimal results directly after the wave. This shows that SwinUnet is more robust and less difficult to train. The training result of TransUnet is better than the SwinUet and SETRnet models, but the training difficulty is relatively high. It is necessary to set the appropriate learning rate and simultaneously adjust the training strategy, including the epoch, learning rate decay strategy, etc.

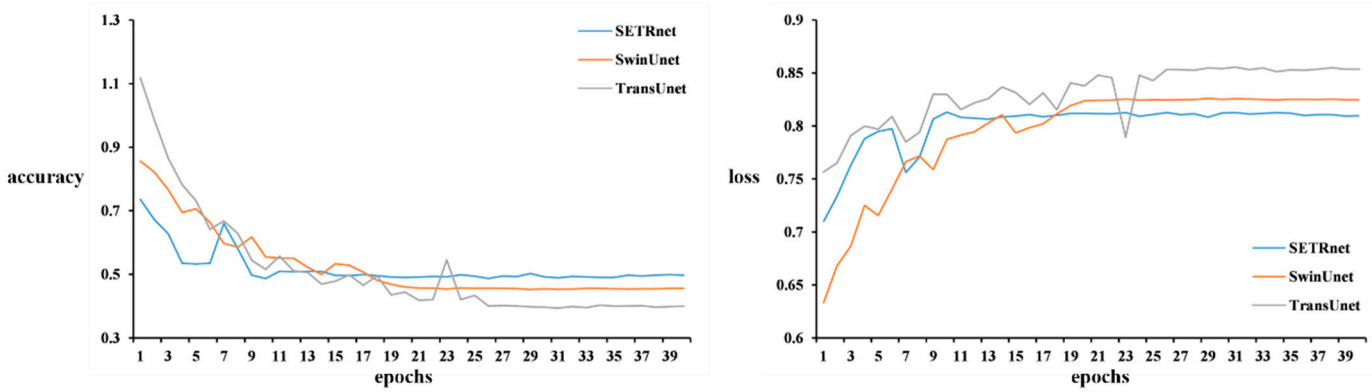

Figure 13. Training process of transformers on the Vaihingen data set.

3.2.6. Comparison of CNNs

Figure 14 shows the comparison results of transformers and Unet [27], DeepLab V3+ [39], and MAnet (multiscale attention net) [40] for the Vaihingen data set segmentation results. SwinUnet and SETRnet were significantly better than CNN for large-scale feature segmentation, which further proves that a transformer is beneficial to improve the large-scale feature-learning ability [41]. Regarding the case of feature confusion, SwinUnet had fewer such occurrences, and the problem of feature misclassification confusion is commonly found in CNNs, mainly because the feature-extraction ability of convolution is not as good as that of transformers. However, the segmented edges in both transformer networks appear to be less fine than those of convolutional networks, and even the results of TransUnet containing a CNN are better than those of the other two transformers. This indicates that the transformer still needs improvement in edge-extraction capability, and it is necessary to improve the spatial feature-information-learning capability.

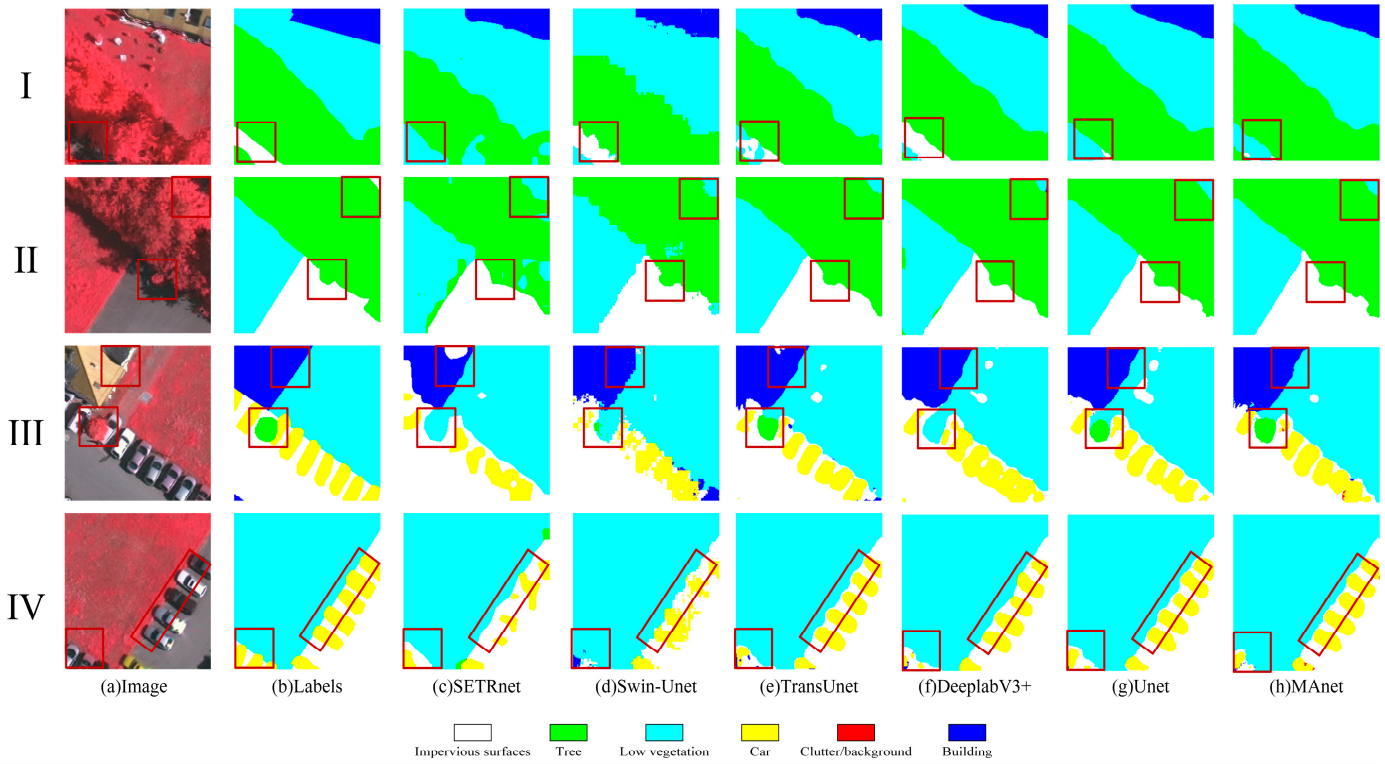

Figure 14. Classification results of CNN and transformers on the Vaihingen data set. The red boxes are where the differences between the five network segmentations are large.

4. Discussion

In the Potsdam data set experiment, SwinUnet was more suitable for the feature-segmentation extraction of large-scale remote-sensing images, while TransUnet performed relatively poorly. The overall accuracy performance of TransUnet was inferior to the other models. This is because the whole network structure of SwinUnet is built using a transformer, which can improve the interaction ability of the global semantic information of the model. Moreover, the fusion of features at different scales by jump linking enhances the model's ability to segment predictions at the pixel level [42]. Using the Vaihigen data set experiment, TransUnet not only had the best overall accuracy performance but also had the highest accuracy for different features. This is because the encoder constructed by the transformer of SwinUnet is still inferior to CNN for small-scale image feature extraction in the small-scale case. However, TransUnet is a combination of CNN and a transformer that enhances the transformer and accelerates its convergence by using appropriate convolutional bias to obtain more local feature information. Thus, TransUnet has better segmentation results in the Vaihingen data set and better segmentation details and contours for features [43]. Therefore, before selecting a transformer, it should be considered according to the remote-sensing image scale. SwinUnet was preferential for large-scale images and TransUnet for smaller-scale images, while SETRnet was not suitable as a remote-sensing image segmentation network. Meanwhile, the comparison experiment between a transformer and a CNN proves that a transformer is inferior to a CNN for the segmentation of edge features. However, it is significantly better than a CNN for large-scale feature segmentation. This situation may be related to the fact that a transformer itself focuses too much on the global features, resulting in ignoring some edge features.

5. Conclusions

In this study, we investigated which transformer model is more suitable for remote-sensing image feature segmentation by evaluating the performance of different transformer models. In this study, first, three transformer models were briefly described, and the

network structure of the transformer model was separately constructed. In this study, experiments were conducted on two data sets, namely the Vaihingen and Potsdam data sets, and the SETRnet, SwinUnet, and TransUnet models were compared by conducting a visual analysis of feature-segmentation results and by assessing their accuracy and training time. The three models were further discussed and analyzed with CNNs. This research will aid in the understanding of different transformer models and the selection of more-suitable transformer models for remote-sensing image feature segmentation in future experiments. The results indicated that SwinUnet performed better on the large-scale Potsdam data set thanks to its excellent global semantic interaction and pixel-level segmentation prediction ability. TransUnet benefits from its network structure jointly constructed by a transformer and a CNN, and it has the highest accuracy on the small-scale Vaihingen data set. Compared with SwinUnet and TransUnet, SETRnet is not suitable for the segmentation extraction of remote-sensing image features. At the same time, the experimental results also show that a transformer has obvious advantages for the segmentation of large-scale objects, but the pure transformer structure is not suitable for remote-sensing image segmentation. For different scales of remote-sensing data, researchers need to choose appropriate transformer models and improve methods.

In the future, we should pay more attention to the following two areas. First, the transformer model's ability to extract the edges of features is insufficient. We should address the issue of the transformer model's overly focusing on the semantic relationship between using the global details and ignoring the edge details. Second, we should invest in expanding the application of different transformer models in the segmentation and extraction of remote-sensing image features and further verify their effectiveness.

Author Contributions: Conceptualization, M.Y.; methodology, F.Q. and M.Y; formal analysis, M.Y; investigation, M.Y.; data curation, M.Y.; writing—original draft preparation, M.Y.; writing—review and editing, F.Q.; visualization, M.Y.; supervision, F.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the High-Resolution Satellite Project of the State Administration of Science, Technology, and Industry for the National Defense of PRC, grant number 80-Y50G19-9001-22/23; the National Science and Technology Platform Construction, grant number 2005DKA32300; the Key Projects of the National Regional Innovation Joint Fund, grant number U21A2014; the Ministry of Education, grant number 16JJD770019; and the Open Program of Collaborative Innovation Center of Geo-Information Technology for Smart Central Plains Henan Province, grant number G202006.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Srivastava, A.; Jha, D.; Chanda, S.; Pal, U.; Johansen, H.D.; Johansen, D.; Riegler, M.A.; Ali, S.; Halvorsen, P. MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 2252–2263. [CrossRef]
2. Beyaz, A.; Martínez Gila, D.M.; Gómez Ortega, J.; Gámez García, J. Olive Fly Sting Detection Based on Computer Vision. *Postharvest Biol. Technol.* **2019**, *150*, 129–136. [CrossRef]
3. Beyaz, A.; Gerdan, D. Meta-Learning Based Prediction of Different Corn Cultivars from Colour Feature Extraction with Image Processing Technique. *Tarım Bilim. Derg.* **2021**, *27*, 32–41. [CrossRef]
4. Yuan, J.; Wang, D.; Li, R. Remote Sensing Image Segmentation by Combining Spectral and Texture Features. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 16–24. [CrossRef]
5. Kotaridis, I.; Lazaridou, M. Remote Sensing Image Segmentation Advances: A Meta-Analysis. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 309–322. [CrossRef]
6. Ibrahim, A.; El-kenawy, E.-S.M. Image Segmentation Methods Based on Superpixel Techniques: A Survey. *J. Comput. Sci. Inf. Syst.* **2020**, *15*, 1–11.

7. Xiong, D.; He, C.; Liu, X.; Liao, M. An End-To-End Bayesian Segmentation Network Based on a Generative Adversarial Network for Remote Sensing Images. *Remote Sens.* **2020**, *12*, 216. [CrossRef]
8. Zheng, K.; Wang, H.; Qin, F.; Han, Z. A Land Use Classification Model Based on Conditional Random Fields and Attention Mechanism Convolutional Networks. *Remote Sens.* **2022**, *14*, 2688. [CrossRef]
9. Misbah, K.; Laamrani, A.; Khechba, K.; Dhiba, D.; Chehbouni, A. Multi-Sensors Remote Sensing Applications for Assessing, Monitoring, and Mapping NPK Content in Soil and Crops in African Agricultural Land. *Remote Sens.* **2021**, *14*, 81. [CrossRef]
10. Sataer, G.; Sultan, M.; Emil, M.K.; Yellich, J.A.; Palaseanu-Lovejoy, M.; Becker, R.; Gebremichael, E.; Abdelmohsen, K. Remote Sensing Application for Landslide Detection, Monitoring along Eastern Lake Michigan (Miami Park, MI). *Remote Sens.* **2022**, *14*, 3474. [CrossRef]
11. Zhang, C.; Sargent, I.; Pan, X. Joint Deep Learning for Land Cover and Land Use Classification. *Remote Sens Env.* **2019**, *221*, 173–187. [CrossRef]
12. Verburg, P.H.; Neumann, K.; Nol, L. Challenges in Using Land Use and Land Cover Data for Global Change Studies. *Glob. Change Biol.* **2011**, *17*, 974–989. [CrossRef]
13. Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Queiroz Feitosa, R.; van der Meer, F.; van der Werff, H.; van Coillie, F.; et al. Geographic Object-Based Image Analysis—Towards a New Paradigm. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 180–191. [CrossRef]
14. Ming, D.; Li, J.; Wang, J.; Zhang, M. Scale Parameter Selection by Spatial Statistics for GeOBIA: Using Mean-Shift Based Multi-Scale Segmentation as an Example. *ISPRS J. Photogramm. Remote Sens.* **2015**, *106*, 28–41. [CrossRef]
15. Talukdar, S.; Singha, P.; Mahato, S.; Shahfahad; Pal, S.; Liou, Y.-A.; Rahman, A. Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. *Remote Sens.* **2020**, *12*, 1135. [CrossRef]
16. Sheykhmousa, M.; Mahdianpari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; Homayouni, S. Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6308–6325. [CrossRef]
17. Maulik, U.; Chakraborty, D. Remote Sensing Image Classification: A Survey of Support-Vector-Machine-Based Advanced Techniques. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 33–52. [CrossRef]
18. Du, P.; Samat, A.; Waske, B.; Liu, S.; Li, Z. Random Forest and Rotation Forest for Fully Polarized SAR Image Classification Using Polarimetric and Spatial Features. *ISPRS J. Photogramm. Remote Sens.* **2015**, *105*, 38–53. [CrossRef]
19. Bouguettaya, A.; Zarzour, H.; Kechida, A.; Taberkit, A.M. Vehicle Detection from UAV Imagery with Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 1–21. [CrossRef]
20. Seo, H.; Badiei Khuzani, M.; Vasudevan, V.; Huang, C.; Ren, H.; Xiao, R.; Jia, X.; Xing, L. Machine Learning Techniques for Biomedical Image Segmentation: An Overview of Technical Aspects and Introduction to State-of-art Applications. *Med. Phys.* **2020**, *47*, e148–e167. [CrossRef]
21. Alem, A.; Kumar, S. Deep Learning Methods for Land Cover and Land Use Classification in Remote Sensing: A Review. In Proceedings of the 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 4–5 June 2020; pp. 903–908.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Curran Associates, Inc., Lake Tahoe, NV, USA, 3–6 December 2012; Volume 25.
23. Li, R.; Wang, L.; Zhang, C.; Duan, C.; Zheng, S. A2-FPN for Semantic Segmentation of Fine-Resolution Remotely Sensed Images. *Int. J. Remote Sens.* **2022**, *43*, 1131–1155. [CrossRef]
24. Huang, J.; Zhang, X.; Xin, Q.; Sun, Y.; Zhang, P. Automatic Building Extraction from High-Resolution Aerial Images and LiDAR Data Using Gated Residual Refinement Network. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 91–105. [CrossRef]
25. Huang, B.; Zhao, B.; Song, Y. Urban Land-Use Mapping Using a Deep Convolutional Neural Network with High Spatial Resolution Multispectral Remote Sensing Imagery. *Remote Sens. Environ.* **2018**, *214*, 73–86. [CrossRef]
26. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *arXiv* **2018**, arXiv:180202611.
27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI, Munich, Germany, 5–9 October 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
28. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid Attention Network for Semantic Segmentation. *arXiv* **2018**, arXiv:1805.10180. [CrossRef]
29. Wang, W.; Zhou, T.; Yu, F.; Dai, J.; Konukoglu, E.; Van Gool, L. Exploring Cross-Image Pixel Contrast for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
30. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2022**, arXiv:2010.11929.
31. Aleissaee, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.-S.; Khan, F.S. Transformers in Remote Sensing: A Survey. *arXiv* **2022**, arXiv:2209.01206.
32. Chen, N.; Watanabe, S.; Villalba, J.; Żelasko, P.; Dehak, N. Non-Autoregressive Transformer for Speech Recognition. *IEEE Signal Process. Lett.* **2021**, *28*, 121–125. [CrossRef]

33. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6877–6886.
34. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
35. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306.
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
37. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. *arXiv* **2021**, arXiv:2105.05537.
38. Zhang, H.; Fritts, J.E.; Goldman, S.A. Image Segmentation Evaluation: A Survey of Unsupervised Methods. *Comput. Vis. Image Underst.* **2008**, *110*, 260–280. [CrossRef]
39. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
40. Chattopadhyay, S.; Basak, H. Multi-Scale Attention U-Net (MsAUNet): A Modified U-Net Architecture for Scene Segmentation. *arXiv* **2020**, arXiv:2009.06911.
41. Sun, Z.; Zhou, W.; Ding, C.; Xia, M. Multi-Resolution Transformer Network for Building and Road Segmentation of Remote Sensing Image. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 165. [CrossRef]
42. Yao, J.; Jin, S. Multi-Category Segmentation of Sentinel-2 Images Based on the Swin UNet Method. *Remote Sens.* **2022**, *14*, 3382. [CrossRef]
43. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]