STransFuse: Fusing Swin Transformer and Convolutional Neural Network for Remote Sensing Image Semantic Segmentation

Liang Gao,Hui Liu,Minhang Yang,Long Chen,Yaling Wan,Yurong Qian,Zhengqing Xiao

***Abstract*—The convolutional neural network (CNN) bolstered the application study of remote sensing images. However, the fixed size of the receptive field limits CNN's ability in semantic modelling. Although the Transformer model based on self-attention may model global semantic information, it remains a challenge to work effectively in the remote sensing field with limited data. Leveraging the benefits of both Transformer and CNN, a new model is proposed in this article as a semantic segmentation method for remote sensing images which is constructed by fusing Swin Transformer and CNN and is thus named as STransFuse model. On the one hand, the STransFuse model makes use of the capability of the transformer network to model global semantic information in remote sensing images. On the other hand, it utilized the capability of the CNN with pre-training weights to solve the problem of low performance of the Transformer network when the amount of remote sensing image data is small and to provide the contextual location information of the image. An Adaptive Fusion Module (AFM) is created to adaptively fuse the feature maps produced by the Transformer network and CNN to improve the model's feature representation capacity. The OA (Overall Accuracy) of the STransFuse model is 1.36% higher than the baseline on the Vaihingen dataset, and 1.27% higher than baseline on the Potsdam dataset. When compared with other state-of-the-art models, the STransFuse model performed competitively.**

***Index Terms*—Transformer, remote sensing, self-attention, semantic segmentation.**

I. Introduction

CNN has performed admirably in the realm of computer vision. Fully convolutional networks (FCNs) [1] based on CNN have become the most popular image segmentation architecture. However, the scale of features contained in remote sensing images varies greatly, and large features of huge ground objects (e.g., buildings) in an image will occupy a large proportion of the image, and FCNs cannot acquire the contextual information of the image well due to the limitation of the fixed perceptual field of the convolution kernel. To address this issue, a CNN-based model will employ the pooling method to reduce the resolution of the feature map to obtain a global representation of feature information. However, the global pooling method will cause the model to lose information about small targets of the image.

Some researchers [2]–[4] have attempted to tackle the above challenges by fusing multi-scale contextual data. By merging atrous convolutions to gather multi-scale contextual information, Chen et al. [3] improved the Atrous Spatial Pyramid Pooling (ASPP) module. To improve the expression of feature map information, U-Net [5] uses an encoder-decoder structure to retrieve feature map information at different levels via skip connections. To mine the ability of global contextual information, the Pyramid Scene Parsing Network (PSPNet) [4] gathers contextual information based on different regions through the pyramid pooling module.

Furthermore, several researchers [6], [7] attempted to tackle the problem of a lack of network perceptual fields by utilizing self-attention [8]. To model semantic significance from spatial and channel dimensions, Fu et al. [6] proposed Compact Position Attention Module and Compact Channel Attention Module based on the self-attentive process. In the field of computer vision, the Transformer technique [9]–[13] based on sequence-to-sequence prediction has demonstrated exceptional performance. The transformer's structure foregoes the convolution operation in favor of a pure attention mechanism. Unlike CNN to obtain features, Transformer can obtain global context information through self-attention. An experiment [9] demonstrated that the Transformer network can achieve high performance in image tasks like image classification, image recognition, and semantic segmentation when a large-scale pre-training was conducted. In this article, we explored the application potential of Transformer for semantic segmentation in the context of remote sensing images. We tried various Transformer networks that had produced excellent results on public datasets by using remote sensing images and discovered surprisingly that they couldn't deliver sufficient results. This is because when an image is supplied to the Transformer network, the image patch gets compressed into a 1D sequence. Through self-attention, the Transformer network concentrates on the image's semantic global contextual information, and the network lost the image's spatial contextual information

This work was supported by the National Natural Science Foundation of China (61966035); the National Science Foundation of China under Grant (U1803261); the Xinjiang Uygur Autonomous Region Innovation Team (XJEDU2017T002), and the Autonomous Region Graduate Innovation Project (XJ2020G074).*(Corresponding author: Yurong Qian.)*

L.Gao, M.Yang, L.Chen, Y.Qian, and Y.Wan are with College of Software, Xinjiang University, Urumqi 830008, China; Key Laboratory of signal detection and processing in Xinjiang Uygur Autonomous Region,, and also with the Key Laboratory of Software Engineering, Urumqi 830008, China (e-mail: gaoliang@stu.xju.edu.cn; yangminhang@stu.xju.edu.cn; ry19chenlong@stu.xju.edu.cn; wyl@stu.xju.edu.cn; qyr@xju.edu.cn).

H.Liu is with College of Information Science and Engineering, Xinjiang University, Urumqi 830014, China; Key Laboratory of signal detection and processing in Xinjiang Uygur Autonomous Region, , Urumqi 830014, China, and also with the Key Laboratory of Software Engineering, Urumqi 830008, China (e-mail: 903123414@qq.com).

Z.Xiao is with College of Mathematics and Systems Science, Xinjiang University, Urumqi 830014, China (e-mail: xiaozq@xju.edu.cn).

(location information) during computation. The spatial context information cannot be retrieved well by up-sampling in the encoder stage of the Transformer network, resulting in poor picture segmentation. We merged the feature maps of distinct stages to acquire the semantic context information and spatial context information of the image, inspired by the Unet [5].

To this end, we propose a new model for semantic segmentation of remote sensing images, named STransFuse, which is constructed by combining the Swin Transformer's architecture with CNN. Swin Transformer acquires features in the form of shifted windows to establish self-attention and uses CNN to acquire spatial context information. Transformer's success is predicated on extensive data training. However, the image dataset obtained in the field of remote sensing is limited, which severely limits Transformer's application in this field. Resnet34 with training weights is employed as the network backbone of the CNN branch and paired with Swin Transformer to acquire rich feature information of remote sensing images, as inspired by the article [9]. Our main contributions are as follows:

1) A model combining Swin Transformer and Resnet34 is intended to incorporate global semantic information retrieved by the Transformer network with the spatial contextual information extracted via Resnet34 from the images. The problem of degraded Transformer performance owing to tiny remote sensing datasets can be solved by using Resnet34 with pre-trained weights.
2) An AFM is created to adaptively fuse feature maps using the self-attention mechanism to improve the model's ability in expressing features.
3) On the Vaihingen and Potsdam datasets, the proposed STransFuse model performs better.

The remainder of this article is structured in the following manner. The related work on semantic segmentation of remote sensing images is discussed in Section II and some Transformer researches is also reviewed in this Section. In Section III, the specifics of the STransFuse framework as well as the AFM design is explored. The datasets used in the studies, as well as the experimental parameters, are described in Section IV. Section V presents a complete ablation research and experimental comparison between the STransFuse model and some state-of-the-art models to validate the proposed module. The conclusion is given in Section VI.

II. Related Works

A. Semantic Segmentation of Remote Sensing Images

Remote sensing images are widely used in many application fields, including crop yield estimation [14], military reconnaissance and natural disaster monitoring [15]. The accuracy of these applications is largely determined by the segmentation accuracy of remote sensing images. Traditional remote sensing image semantic segmentation relies on the texture information and spectral information of images, which requires a lot of manpower and material resources. The introduction of deep learning into remote sensing image segmentation has increased the accuracy, resulting in a significant increase in image segmentation efficiency. Lin et al. [16] created a scale-aware module to let the network distinguish different features using weighted feature maps. Chong et al. [17] proposed the Context Union Edge Network for semantic segmentation of remote sensing images, and a context-based feature augmentation module to improve CNN's capacity to differentiate small targets as well as a dual-stream network to refine small target edge information. Xiang et al. [18] created an adaptive feature selection module that learns the weight contribution of each feature block at different scales to improve the network's performance. Li et al. [19] introduced a semantic boundary aware network to collect correct boundary information for land cover categorization. The network is adaptable to obtain image boundary information using a bottom-up method and can reduce noise information from low-level characteristics. AFNet [20] employed the scale-feature attention module and scale-layer attention module to better tackle the difference between intra-class and inter-class in remote sensing images, and conducted adaptive feature improvement for targets of various sizes. Pan et al. [21] introduced a conditional generative adversarial network that actively generates new sample images while extracting advanced spatial information from previous training images. The network achieved greater classification accuracy using this strategy. To overcome the problem of cloud segmentation in remote sensing images and increase the network's feature extraction ability, Yao et al. [22] presented a multi-scale feature extraction and content-aware recombination network. The spatial relation module and the channel relation module proposed by Mou et al. [23] could learn and infer the relationship between any two geographical locations or feature maps to produce effective contextual spatial relation modeling.

B. Contextual Information

To increase the accuracy of image semantic segmentation, it's critical to understand how to properly extract the image's contextual information. FCNs [1] first widened the receptive field by pooling to capture the image's context information, but multiple downsampling processes resulted in the feature map losing certain details. Unet [5] created a network framework with an encoder-decoder structure that allows detailed information from low-level feature maps to be merged into high-level feature maps by skipping network layers. The Feature Pyramid Transformer (FPT) [11] is a completely active feature interaction that extends the receptive field through the specified Transformer. Chen et al. [24] developed a tensor generation module to capture contextual data and offered a new way to modeling 3D context representations. The spatial relation module and the channel relation module were introduced by Mou et al. [25] to learn and infer the global link between any two spatial positions or feature maps, and then build a feature representation with improved relationship. In this article [26], an axial-attention model was presented to widen the receptive field in the model and alleviate the problem of losing remote context information in convolution. For remote sensing images, context information indicates the relationship between features. The difficulty of obtaining context information is due to the high resolution and imbalanced proportion

of various characteristics reflecting various ground objects if compared with ordinary images. Generally it is difficult to analyze remote sensing images directly and handling remote sensing images frequently requires preprocessing (cropped, normalized). Some methods based on self-attention mechanisms generate excessive waste of computer resources in getting the context relationship when the processed images patch only contains one category of ground object.

When the common model executes the convolution operation, the proportion of big scale features in the patch can be substantially higher than that of small scale features, causing small scale features to be heavily influenced by large scale features. A hot issue in the research of remote sensing image processing is how to efficiently resolve intra-class and inter-class disparities in remote sensing images while balancing the accuracy and efficiency of remote sensing image processing.

C. Transformer

The Transformer was originally used in the realm of Natural Language Processing (NLP) [8]. It is a deep neural network model that extracts intrinsic properties via the self-attention approach. The good experimental performance Transformer achieved in the field of NLP suggested that it may be applied to the field of image processing. The first Transformer model based on pure self-attention for image recognition, Vision Transformer (ViT) [9], has achieved outstanding results in image processing, but the model requires a large number of datasets for training, and the results obtained by applying the model directly to small or medium-sized datasets were not promising. A great number of researchers tried many ways to make the Transformer more successful in the field of computer vision, inspired by the construction of the visual Transformer model [27]–[32]. Semantic SEgmentation Transformer (SETR) [33] is a model for semantic segmentation that used Transformer as an encoder. A sophisticated segmentation model can be created by combining the pure Transformer encoder with some simple decoders. DEtection Transformer (DETR) [10] is a Transformer that was developed by Facebook AI researchers and applied to a vision model. It's the first target detection framework to successfully incorporate Transformer as a pipeline's core building block. In the areas of target identification and panorama segmentation, the DETR model performed well. The Transformer-in-Transformer (TNT) model [34] makes use of an inner Transformer block to extract the images patch's internal structure information, allowing the model to extract both global and local properties. The model performed well on the ImagesNet benchmark dataset and in various downstream tasks. The Shifted Windows Transformer (Swin Transformer) [35] is a hierarchical Transformer that, like CNN, is capable of increasing the perceptual field of nodes as the network layers deepens. The use of shifted windows allows self-attention to be computed in non-overlapping local windows, reducing the computational complexity that is quadratic of an increase in image size, and thus potentially lowering the hardware requirements for the studies which require dense pixel-level prediction (e.g., semantic segmentation). On datasets including ImageNet-1K andADE20K, the Swin Transformer achieved good results. The experimental results, however, are unsatisfactory when the model is applied to the field of remote sensing. It is because the dataset of remote sensing images is small, and the features of remote sensing images are quite different from those of ordinary images. As inspired by the ViT model [9], we combined the pre-trained Resnet34 as the CNN backbone with the Swin Transformer model to create a two-branch network model that can perform well on remote sensing images.

III. PROPOSED METHODS

A. Overview

The input remote sensing image $x \in R^{H\times W\times C}$, where H represents the height of the image, W represents the width of the image, and C represents the number of channels of the image. We use Swin Transformer and Resnet34 to handle the images, fuse the feature maps at different stages, and finally restore the feature maps to their original size. In Paragraph B, we will introduce the overall structure of STransFuse. Then, the details of Swin Transformer given in Paragraph C. Finally, the AFM is described in Paragraph D.

B. STransFuse Overall Architecture

As shown in Fig. 1(a), the image x is input into the Swin Transformer network and the Resnet34 network respectively. There are 4 stages in Swin Transformer network to get x_{s1},x_{s2},x_{s3},x_{s4} feature maps respectively, and each stage contains Patch Merging and Swin Transformer. Patch Merging works in a similar way to CNN's pooling layer in that it downsamples the image. By shifting the input image's window, this module separates the image into non-overlapping patches. Each patch is considered as a "token". We initially fixed the patch size to 4×4. Then, the eigenvalues in the feature map are projected to the C dimension through a linear embedding layer. Finally, Swin Transformer block is applied to these patch tokens ($\frac{H}{4} \times \frac{W}{4}$). These steps above are collectively referred to as "Stage 1". In the following "Stage 2", Patch Merging concatenates the features of each group of 2 × 2 neighboring patches, and applies linear embedding layer to change the output dimension to 2C, and applies Swin Transformer for feature transformation. In "Stage2", The resolution of the patch is maintained at $\frac{H}{8} \times \frac{W}{8}$. "Stage 3" and "Stage 4" are similar to "Stage 2", and the output patch resolutions are $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, respectively.

The images are input to the Resnet34 network to get the feature maps, which are output by layer1 to layer4 as feature maps x_{c1}, x_{c2}, x_{c3} and x_{c4} respectively, and the sizes of these feature maps are $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$ and $\frac{H}{32} \times \frac{W}{32}$, respectively. The feature maps generated by Resnet34 are merged with those generated by different stages of Swin Transformer to make use of Swin Transformer's capacity in collecting global semantic contextual information of features. Finally, the fused feature map is upsampled twice more, and the feature map is returned to its original size.

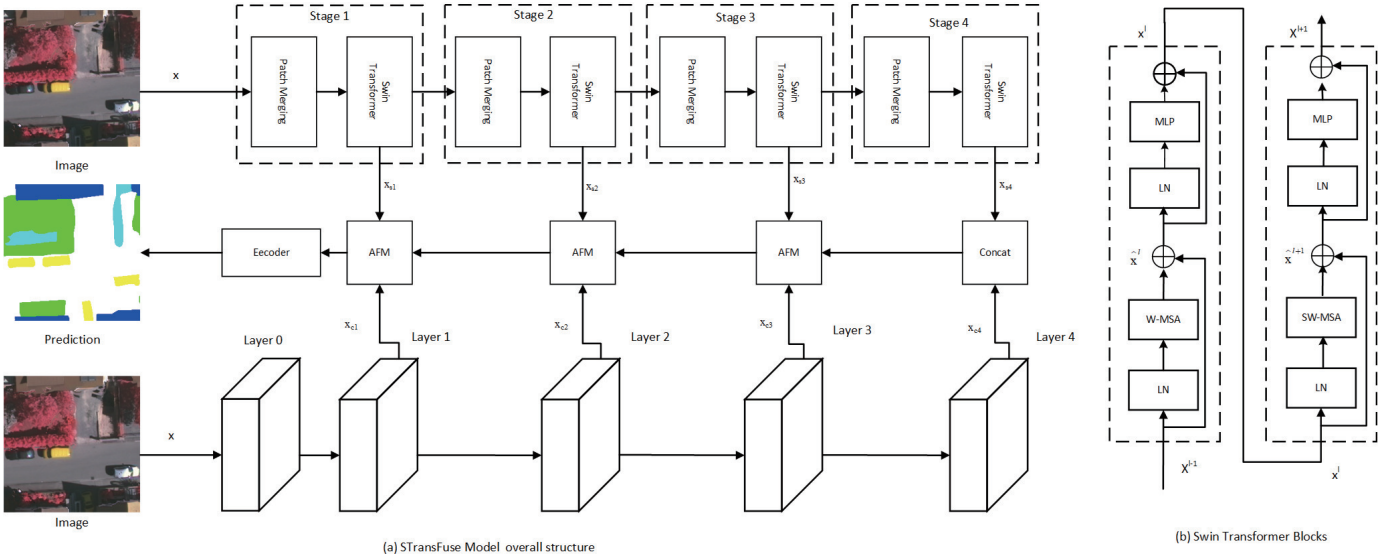

Fig. 1. (a) The overall structure of STransfuse Model; (b) the detail of Swin Transformer Blocks

C. Swin Transformer Block

Swin Transformer uses the feature map enters the Window Multi-head Self Attention (W-MSA) to replace the Multi-head Self Attention (MSA) in the Transformer module. As shown in Fig. 1(b), Swin Transformer inputs the feature map processed by Patching Merging into the Swin Transformer block. Then, the feature map enters the W-MSA module through the LayerNorm layer, and there is a residual connection between each module and another LayerNorm layer.

The self-attention used in the standard Transformer block is calculated by relating one of the tokens to all other tokens. This calculation makes the computation workload of the network grow quadratically with respect to the resolution size of the image, and for some intensive prediction tasks (e.g., semantic segmentation), the model will require high-end computing devices. The Swin Transformer will perform the self-attentive computation in a local window. Each window will be evenly split into M×M patches in a non-overlapping manner. In this case, the computational complexity of MSA is shown in (1), and the computational complexity of W-MSA is shown in (2).

$$\Omega\text{MSA}=4hwC^2+2(hw)^2C \tag{1}$$

$$\Omega\text{W-MSA}=4hwC^2+2M^2hwC \tag{2}$$

Where h and w are the height and width of the image, respectively. In (1), the computational complexity of MSA is quadratic to the production of h and w. In (2), when M is a fixed size (set to 7 by default), the computational complexity of W-MSA is linearly related to the production of h and w. Compared with W-MSA, Shifted Window Multi-head Self Attention (SW-MSA) shifts the window. Because of the sliding window segmentation operation performed by W-MSA, the cropped patches do not overlap, and there is no correlation between the windows which limits the performance of Swin Transformer. Therefore, in order to realize the cross-window connection of the model, article [32] introduced SW-MSA in the model.

In general, the calculation process of the feature map in Swin Transformer block is shown in (3)-(6):

$$\widehat{x}^{l}=\text{W-MSA}(LN(x^{l-1}))+x^{l-1} \tag{3}$$

$$\mathrm{x}^{l}=\text{MLP}(LN(\widehat{x}^{l}))+\widehat{x}^{l} \tag{4}$$

$$\widehat{x}^{l+1}=\text{SW-MSA}(LN(x^{l}))+x^{l} \tag{5}$$

$$\widehat{x}^{l+1}=\text{MLP}(LN(\widehat{x}^{l+1}))+x^{l+1} \tag{6}$$

Where $\hat{x}^{l}$ denote the output characteristics of the W-MSA module of l block, $\hat{x}^{l+1}$ denote the output characteristics of the SW-MSA module of l+1 block, and x^{l} denote the MLP module of l block. W-MSA denote based multi-head self-attention using regular window partitioning configurations. LN denote Layer Normalization. MLP denote Multi Layer Perceptron. SW-MSA denote based multi-head self-attention using shift window partitioning configurations.

D. AFM

To efficiently fuse the encoded features from CNN and Swin Transformer, we designed an AFM based on the self-attentive mechanism, whose structure is shown in Fig. 2. We will perform the fusion of features with the following :

$$\mathrm{x}_{cs,i}=\text{Re}\,LU(Conv(Interpolate(concat(x_{s,i},x_{c,i})))) \tag{7}$$

$$\mathrm{x}_{BN,i-1}=\text{Re}\,LU(BN(Conv(Concat(x_{cs,i},x_{s,i-1})))) \tag{8}$$

$$\mathrm{x}_{q}=Soft\max(Conv(x_{\text{BN},i-1})) \tag{9}$$

$$x_k = Linear(Concat(\text{AdaptiveAvgPool2d}(x_{BN,i-1}))) \tag{10}$$

$$x_v = Linear(Concat(\text{AdaptiveAvgPool2d}(x_{BN,i-1}))) \tag{11}$$

$$x_{s,i-1} = (x_q \otimes x_v) \otimes x_v \oplus Concat(x_{cs,i}, x_{s,i-1}) \tag{12}$$

Among them, $x_{s,i}$ represents the feature matrix output by the i-th stage of Swin Transformer, and $x_{c,i}$ represents the feature matrix output by the i-th layer of CNN. x_q is the query in self-attention calculation, x_k represents the key in self-attention calculation, x_v represents the value in self-attention calculation, $x_q \bigotimes x_v$ gets the self-attention weight matrix, $(x_q \bigotimes x_v) \bigotimes x_v$ obtains the weighted feature matrix, and add the weighted feature matrix and the fusion feature matrix to obtain $x_{s,i-1}$.

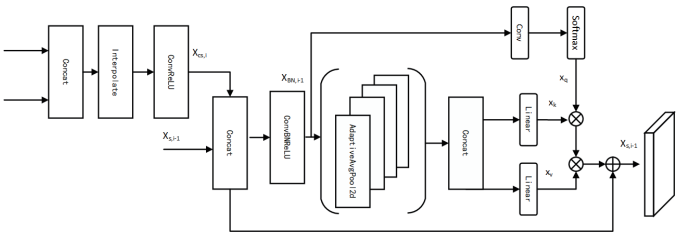

Fig. 2. Detail display of AFM

IV. DATASET DESCRIPTION AND DESIGN OF EXPERIMENTS

A. Dataset

1) Vaihingen: There are 33 patches in the Vaihingen dataset. Each patch is made up of genuine orthoimages that were recovered from a larger mosaic. The ground sampling distance (GSD) is 9 cm, and each image has a resolution of roughly 2500×2500 pixels. The image contains three wavebands, namely near-infrared (NIR), red (R) and green (G). We did not use normalized digital surface model (nDSM) data, and DSM data. We used the ground truth whose boundaries of objects have not been eroded by 3-pixel radius for testing. According to the official division principle, 17 patches were used as the test set (image id: 1, 3, 5, 7, 11, 13, 15, 17, 20, 21, 23, 26, 28, 30, 32, 34, 37), and the other 16 as the training set (image id: 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, 38). For the large image, we cut it into 256×256 slices. In the data augment strategy, we adopted random horizontal and vertical flip operations.

2) Potsdam: The Potsdam-2D semantic annotation collection consists of 38 patches, each with a GSD of 5 cm and a resolution of 6000×6000 pixels. We used IRRG images as the dataset instead of nDSM or DSM data. We followed the official division principle and used 13 of them as the test set (including the image ids of 02_13, 02_14, 03_14, 04_13, 04_14, 04_15, 05_13, 05_14, 05_15, 06_13, 06_14, 06_15, 07_13), and the other 24 as the training set (with image ids of 2_10, 2_11, 2_12, 3_10, 3_11, 3_12, 4_10, 4_11, 4_12, 5_10, 5_11, 5_12, 6_7, 6_8, 6_9, 6_10, 6_11, 6_12, 7_7, 7_8, 7_9, 7_10, 7_11 and 7_12). We also used the ground truth that has not been eroded for testing, and used the same data enhancement method as the Vaihingen dataset.

B. Evaluation Metric

We employed the data publisher's evaluation approach [36], which was also used in the articles [19], [21], [27], [37]. We used Intersection over Union (IoU) for each category, F1-score for each category, mean Intersection over Union (mIoU), mean F1-score, overall accuracy (OA) as our evaluation indicators. Because many indicators are based on confusion matrix for calculation, before introducing the specific formula of each indicator, the meaning of some symbols of the confusion matrix is defined as follows: True positive (TP), True negative (TN), False positive (FP) and False negative (FN). Therefore, the precision rate is calculated by using (13), and the recall rate is calculated using (14):

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{13}$$

$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{14}$$

The definition of OA is shown in Equation (15):

$$\text{OA}=\frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}} \tag{15}$$

The F1-score formula for each category is defined as shown in Equation (16):

$$\text{F1}=2\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}} \tag{16}$$

The mean F1-score is obtained by averaging the F1-score of each category. The higher the value of F1-score is, the better the experimental result is. The definition of IOU is shown in the following formula :

$$\text{IoU}=\frac{\mathbb{N}_p \cap \mathbb{N}_{gt}}{\mathbb{N}_p \cup \mathbb{N}_{gt}} \tag{17}$$

where, $\mathbb{N}_p$ represents the prediction set, and $\mathbb{N}_{gt}$ represents the ground truth images. mIoU is generally calculated based on class. With the calculated IoU of each class, a global evaluation is obtained by using the average of the IoUs.

C. Training Configuration

All the experiments were implemented using PyTorch 1.4.0, Python3.7, CUDA 10.1 and CuDNN 7.6.5. The networks use the Adam optimizer, and the weight decay is 0.0002. We adopted "ploy" learning rate policy with a power of 0.9. The cross entropy loss with weight was defined as shown in (18):

$$x_{s,i-1} = (x_q \otimes x_v) \otimes x_v \oplus Concat(x_{cs,i}, x_{s,i-1}) \tag{18}$$

For all datasets, we set the size of batch size to 16 for all models, except for the TNT model and the Transunet model. Because the TNT model and the Transunet model are computationally expensive, in order to cater to our GPU memory size, we set the batch size of these two models to 12. All experiments were measured on a single 2080Ti with a memory size of 11G.

V. EXPERIMENTAL

We tested the effectiveness of the proposed module through an ablation studies. Then, we compared the proposed STransFuse with some state-of-the-art methods and discussed the experimental results.

A. Ablation Studies

We conducted a series of ablation studies in the Vaihingen and Potsdam datasets to demonstrate the validity of the proposed model. We choose FCN [1] (Resnet34) as the baseline network for comparison.

Table I summarized the Ablation results with different configuration of the network blocks. Among them, Swin_xs4 represents that only Swin Transformer is used as the feature extractor, and only the feature map output by stage4 is input into the decoder. Swin uses concat for feature fusion of the feature maps output by all stages of Swin Transformer, and inputs the fused feature maps into the decoder. Res34+Swin represents that we use Resnet34 to process the input map, and then input the extracted feature map into Swin Transformer. Swin+Res34 is a two-branch network model built by fusing Swin Transformer network and Resnet34. This fusion model uses concat to fuse the feature maps of different phases of Swin Transformer network and Resnet34. STransFuse also uses a dual-branch structure as a feature extractor. At different stages of the feature map, we used our own AFM instead of the concat module to fuse the feature map.

It is seen from the Table I that the STransFuse model produced the best results. We first examined the impacts of the single-branch network model and the double-branch network model. Table I shows that single-branch network models (FCN, Swin_xs4, Swin, Res34+Swin) perform worse for semantic segmentation of images than two-branch network models (Swin+FCN, STransFuse). At the same time, we compared the experimental results of using only the features (xs4) output from the final stage of SwinTransformer and fusing features of different stages (xs1, xs2, xs3, xs4). The testing results demonstrate that the Swin model, which combines the feature maps of several phases of the Swin Transformer, can enhance metric OA by 0.49 %. Then, we compared the experimental effects of connecting the Resnet34 network with pre-trained weights to the Transformer model in series (Res34+Swin) and in parallel (Swin+Res34, STransFuse). The parallel network model performs better in the experiments, as seen in Table I. Swin+Res34 is 1.19% better than Res34+Swin in terms of OA. Finally, we compared the AFM and concat modules' performance. As shown in Table I, the model employing our proposed AFM for feature map fusion performs 0.24 % better in OA than the model using concat for fusion. By comparing the Swin Transformer's trial findings, it is observed that our model can solve the problem of Swin Transformer's inability to distinguish small targets. This is because when the Transformer network computes the picture, it stretches the patch into a one-dimensional token. Under the influence of the surrounding large target pixels, the same pixel values of tiny targets will be separated into locations far apart, and the features of the pixels of small targets will appear less visible. The STransFuse model can learn features from both semantic and spatial context information, which helps to tackle the problem of Transformer's inability to learn small target features.

It can be seen clearly from Fig. 3 that the STransFuse model segmented better than the baseline network FCNs, and that the STransFuse model did not misclassify the features with shading effects in row (b). It is demonstrated that combining the feature maps of several stages of the swin Transformer is more effective than utilizing simply single stage feature maps when comparing Swin_xs4 and Swin, and the two-branch network model has superior segmentation performance for buildings than that by the single-branch network model when comparing the visualization effect maps of Res34+Swin and Swin+Res34 in row (b).

We compared the recognition capabilities of the benchmark model FCNs and STransFuse models for different categories of the ground objects. We visualized the last Score layer in the FCNs model and STransFuse model. The highlighted area (red) in the figure depicts the network focusing on the area, whereas the dark (dark blue) area reflects the model not focusing on the area, as seen in Fig. 4. It is shown that the STransFuse model can better detect different sorts of targets in the Vaihingen dataset by comparing the CAM of FCNs and STransFuse. In the building column, our STransFuse model is able to have a more accurate classification of building. Because the features in the photographs are obtained from an above view, the height information of the features in the images is absent, resulting in the texture representation of the tops of building and impervious surface being comparable. Therefore, the FCNs model appeared the phenomenon of "car flying on the roof" in the recognition image. However, due to the use of self attention, STransFuse model modelled the long-range semantic correlation and determined the category information of similar semantics. Therefore, the STransFuse model can recognize semantic information better. In the column where the category car is located, the FCNs did not identify all car features and were not accurate enough in the already identified car boundary information, compared to the STransFuse model which is also good at identifying features with small targets like car. In the column where impervious surface is located, it is shown that FCNs recognized some car's semantic information as impervious surface. This is because car occupies a smaller proportion of the image compared to impervious surface, and impervious surfaces enclose the car. There is no correlation between car and car. This is a common inter-class imbalance in remote sensing photos, which occurs because remote sensing photographs often span a wide range of locations, and larger objects can fill a larger proportion of the image, whereas smaller-scale elements can

TABLE I
ABLATION RESULTS OF DIFFERENT BLOCKS COMBINED SWIN TRANSFORMER, RESNET34 AND STRANSFUSE FRAMEWORK USING VAIHINGEN AND POTSDAM DATASETS. THE VALUE IN BOLD IS THE BEST. ALL VALUES ARE EXPRESSED AS PERCENTAGES

Method	xs1	xs2	xs3	xs4	concat	AFM	Vaihingen mF1	Vaihingen mIoU	Vaihingen OA	Potsdam mF1	Potsdam mIoU	Potsdam OA
FCNs [1]							76.57	63.85	84.71	80.36	69.19	85.44
Swin_xs4				√			71.77	58.87	83.4	76.26	63.93	82.46
Swin [35]	√	√	√	√	√		73.04	60.18	83.79	76.82	64.82	82.95
Res34+Swin	√	√	√	√	√		77.17	64.76	84.91	80.16	68.89	85.28
Swin+Res34	√	√	√	√	√		77.95	66	85.94	81.61	70.92	86.47
STransFuse	√	√	√	√		√	**78.67**	**66.66**	**86.07**	**82.08**	**71.46**	**86.71**

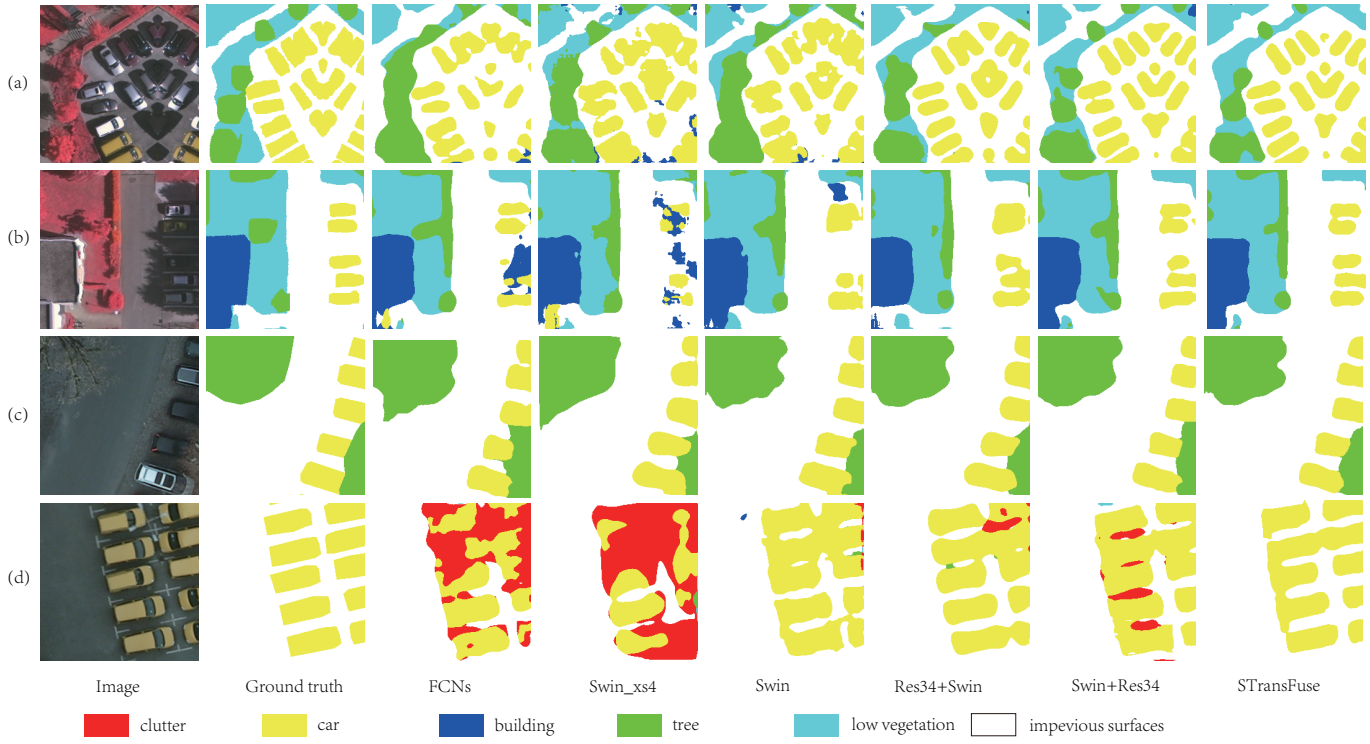

Fig. 3. The result figures of ablation studies visualized in different datasets. (a) and (b) represent the results in the Vaihingen dataset while (c) and (d) represent the results in the Potsdam dataset

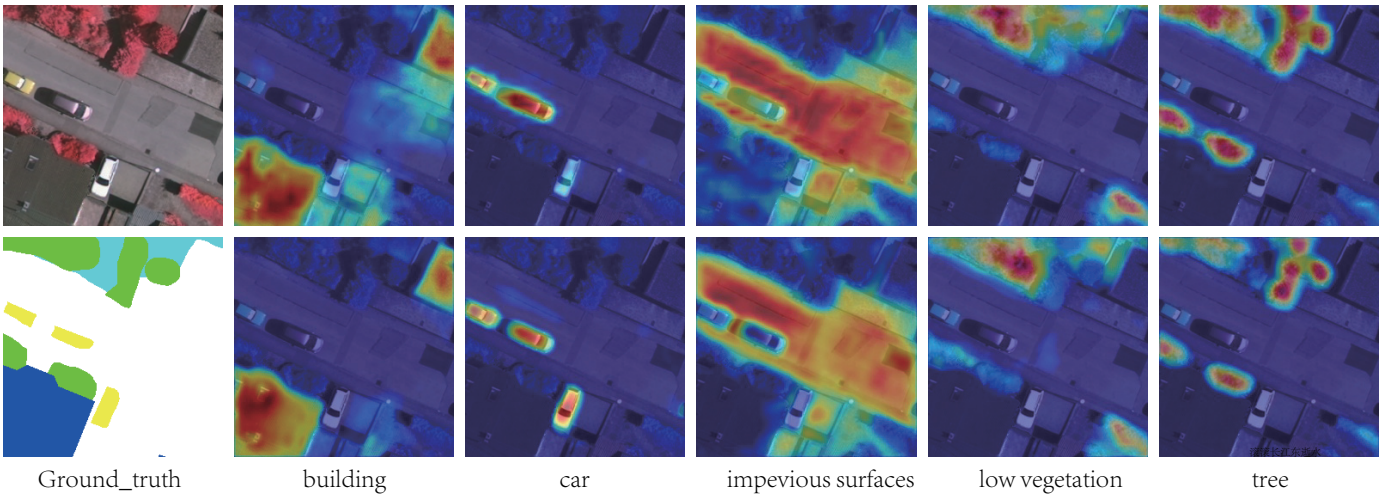

Fig. 4. The class activation mappings (CAM) of different categories of features. The image of the first row is generated by FCNs and the second row is for STransFuse

only occupy a smaller number of pixels. FCNs rely on a fixed-size convolution kernel to obtain features. Therefore, when extracting such small-scale features, they are easily affected by the surrounding feature categories. The Transformer branch we use can effectively solve this type of problem. Low vegetation and tree can be found that the two features have similar feature information through images. In the absence of image height information, it is easy to cause misclassification. Compared to FCNs, the STransFuse model has a better distinction between two different features.

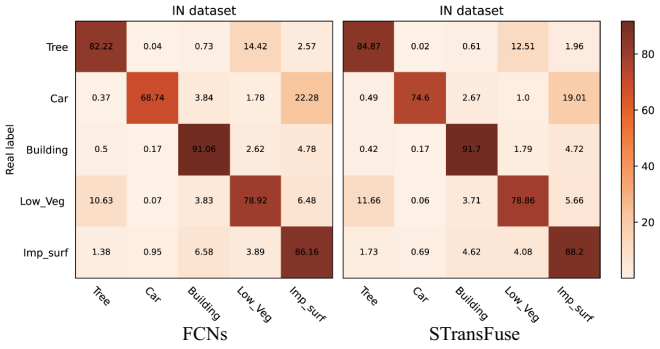

Fig. 5. Confusion matrixes of a sample of Vaihingen dataset with FCNs and STransFuse

Fig. 5 shows the confusion matrix generated after the completion of the test on the Vaihingen dataset. The proportion of accurately predicted categories of the images to total predicted categories is represented by the values in the image blocks at the main diagonal places of the confusion matrix. The darker the image block is, the higher the model's classification accuracy would be. Low vegetation and tree are prone to be misclassified, as seen in Fig. 5, and small-scale car are easily labeled as large-scale impervious surface. To some extent, the STransFuse model solves this problem.

B. Evaluation and Comparisons on the Vaihingen Dataset

Table II shows the results of the comparative experiments. All convolutional network models use Resnet34's pre-training weights. It can be seen from the table that the STransFuse model can achieve the best results. Although Deeplabv3+ [38] produced impressive results, the network uses a lot of GPU memory during training due to the ASPP of the model architecture, and Deeplabv3+ has the longest training time of all the comparable experimental models, as seen in the Fig. 10. The Deeplabv3+ model's overall efficiency is low. Scale-Aware Network (SANet) [16] is a network model that uses the same dataset. This model designed a re-sampling module that implicitly introduced spatial attention through re-sampling feature maps. Through experimental comparison, the model we designed is better than SANet. The mean F1-score was increased by 5.32%. BoTNet [39] replaced the last three bottleneck blocks in Resnet with a global attention module, and was implicitly regarded as multi-head attention through the author's model design. On the Vaihingen dataset, this model performed reasonably well. Due to the limited amount of remote sensing image data, the TNT model [34] has poor experimental results. The Transformer was used as an encoder in the Transunet model [40] to present modeled remote dependencies and to add low-level detail information to the feature maps in the decoder via skip connections. However, due to the design of the encoder and the skip connection, Transunet model has higher requirements for hardware equipment. Comparing the experimental results in testing CNN-based improved models (FCN, Deeplabv3+, Unet, SANet, PSPNet) and Transformer-based improved models (BoTNet, SETR_PUP , TNT, Transunet), the STransFuse model achieved better performance.

On the Vaihingen dataset, the qualitative comparison results are displayed in Fig. 6. As shown in Fig. 6, the STransFuse is capable of recognizing a variety of target categories. It benefits from the Transformer network's capabilities, such as improved global context modeling efficiency without sacrificing low-level detailed localization capability, and the ability to more precisely recognize small-scale objects (car). The car is anticipated as a fuzzy area in comparison to other networks, and the car boundary cannot be reliably determined. Furthermore, we discovered that the model with a pure Transformer encoder (SETR_PUP, TNT) correctly recognizes a building as an impervious surface, but they mistakenly recognized Building as Impermeable Surface. The reason for this phenomenon may be that the two categories of features, building and permeable surface, have similar characteristics. Transformer has similar building and impermeable surface feature values when stretching the patch into a 1D token. When calculating the similarity, the self-attention judges the two types of features as the same type. When the characteristics are very different, the Transformer can distinguish them.

The results of testing the original image on the Vaihingen dataset are shown in Fig. 7. The STransFuse model can better identify large-scale buildings and accurately identify the boundaries of various features.

C. Evaluation and Comparisons on the Potsdam Dataset

Table III shows that the STransFuse model is able to get the highest overall accuracy score on the Potsdam Dataset. About the indicators mean Intersection over Union and mean F1-score, the STransFuse achieved the second highest score.

The results of the qualitative comparison of the models are displayed in Fig. 8, where it can be shown that the STransFuse model performed well for various sizes of features. The STransFuse model determine more precisely the borders of features of small-scale car. It discriminated trees from low vegetation better than previous models. It also reliably determined the boundaries of buildings with huge dimensions. As a result, the STransFuse model is able to recognize multi-scale remote sensing images with high accuracy.

Fig. 9 shows the results of testing the original images from the Potsdam dataset. The STransFuse model showed stronger capability in feature identification at various scales and also in feature boundary identification.

D. Comparison of the Efficiency of State-of-the-art Models in Different Datasets

Fig. 10 (a) shows a comparative plot of the efficiency of the different models for the Vaihingen data.It can be seen that

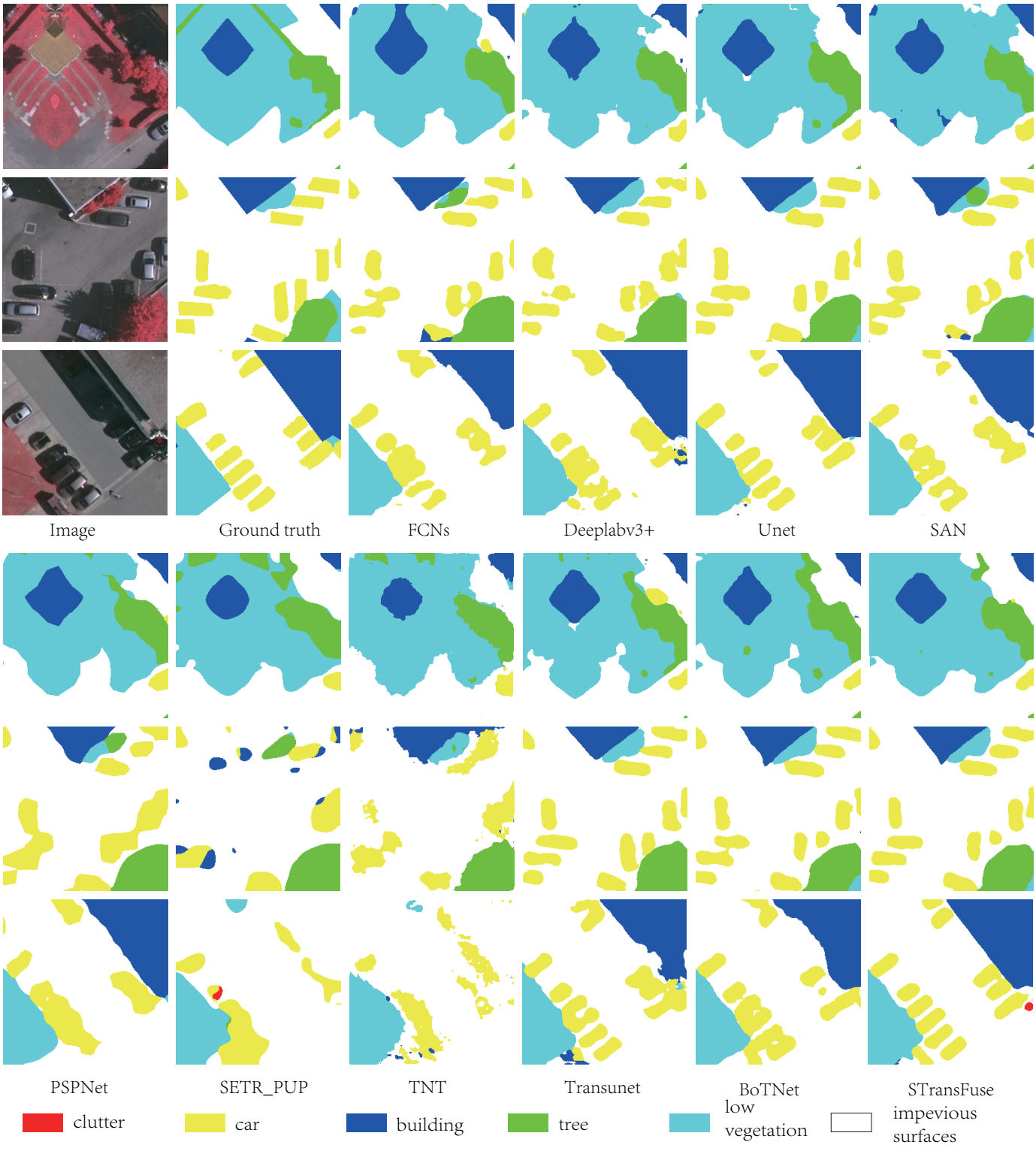

Fig. 6. Visualization results on Vaihingen Dataset

TABLE II
THE COMPARISON OF STRANSFUSE WITH SOME STATE-OF-THE-ART MODELS USING VAIHINGEN DATASET. THE VALUE IN BOLD IS THE BEST, AND THE UNDERLINED VALUE IS THE SECOND BEST. ALL VALUES ARE EXPRESSED AS PERCENTAGES

Method	Trees		Cars		Buildings		Low Veg		Imp surf		mF1	mIoU	OA
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU			
Deeplabv3+ [38]	84.85	73.69	69.99	53.83	90.13	82.04	77.14	62.79	87	76.98	75.39	62.82	84.69
Unet [5]	85.49	74.65	**77.49**	**63.25**	90.86	83.25	77.88	63.77	87.83	78.31	77.43	65.38	85.53
SANet [16]	84.67	73.42	71.78	55.98	90.43	82.53	77.21	62.88	86.93	76.89	73.35	63.74	84.79
PSPNet [4]	84.45	73.08	67.26	50.67	90.53	82.69	76.49	61.93	86.47	76.16	75.21	62.41	84.44
BoTNet [39]	85.07	74.03	75.33	60.42	91.2	83.82	78.26	64.28	88.04	78.64	77.71	65.51	85.65
SETR_PUP [33]	81.67	69.03	44.5	28.62	83.23	71.27	70.93	54.96	81.06	68.15	65.19	51.58	78.9
TNT [34]	81.68	69.04	41.89	26.49	84.58	73.28	70.82	54.83	81.76	69.15	64.76	51.49	79.39
Transunet [40]	84.93	73.8	70.44	54.37	88.26	78.99	77.34	63.05	85.86	75.23	72.78	60.49	83.93
STransFuse	**85.51**	**74.69**	77.14	62.79	**91.46**	**84.27**	**79.04**	**65.35**	**88.25**	**78.97**	**78.67**	**66.66**	**86.07**

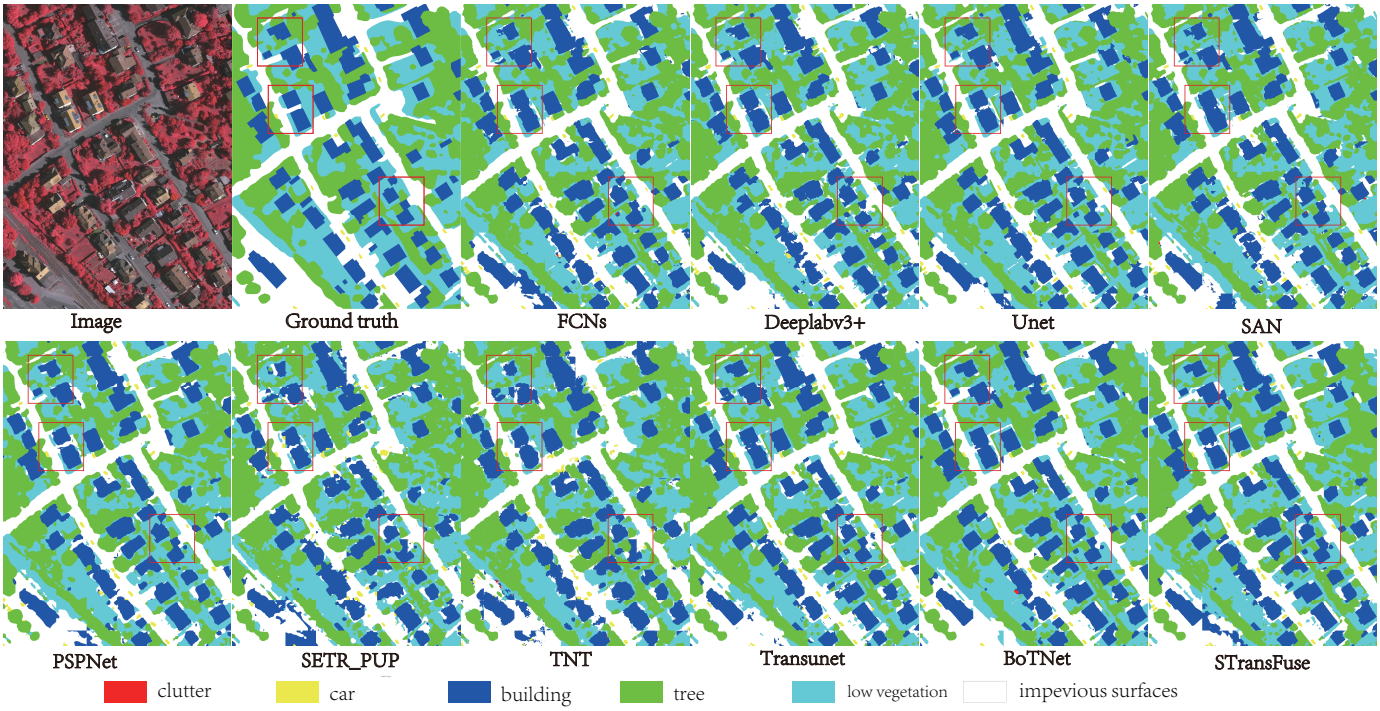

Fig. 7. Visualization results on Vaihingen Dataset original image

TABLE III
THE COMPARISON OF STRANSFUSE WITH SOME STATE-OF-THE-ART MODELS USING POTSDAM DATASET. THE VALUE IN BOLD IS THE BEST, AND THE UNDERLINED VALUE IS THE SECOND BEST. ALL VALUES ARE EXPRESSED AS PERCENTAGES

Method	Trees		Cars		Buildings		Low Veg		Imp surf		mF1	mIoU	OA
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU			
Deeplabv3+ [38]	82.14	69.69	87.2	77.31	92.35	85.79	81.35	68.57	88.21	78.9	80.12	68.85	85.14
Unet [5]	**84.17**	**72.67**	**89.68**	**81.28**	93.22	87.3	**82.96**	**70.89**	89.46	80.94	**82.14**	**71.58**	86.5
SANet [16]	82.25	69.85	86.83	76.72	92.61	86.24	81.8	69.2	88.47	79.32	80.51	69.27	85.39
PSPNet [4]	82.44	70.12	84.99	73.9	92.54	86.11	81.46	68.72	88.06	78.66	80.31	68.83	85.21
BoTNet [39]	81.66	69	87.68	78.06	93.17	87.21	81.64	68.98	88.95	80.1	80.17	69.14	85.39
SETR_PUP [33]	64.58	47.69	73.2	57.73	83.98	72.38	71.18	55.25	80.44	67.27	67.56	53.23	74.68
TNT [34]	68.08	51.61	70.29	54.19	84.9	73.77	72.45	56.8	79.29	65.69	67.44	53.24	75.1
Transunet [40]	80.27	67.04	87.58	77.91	89.97	81.77	80.44	67.29	86.64	76.42	78.33	66.59	83.36
STransFuse	83.61	71.84	88.51	79.39	**93.92**	**88.53**	82.91	70.81	**89.75**	**81.41**	82.08	71.46	**86.71**

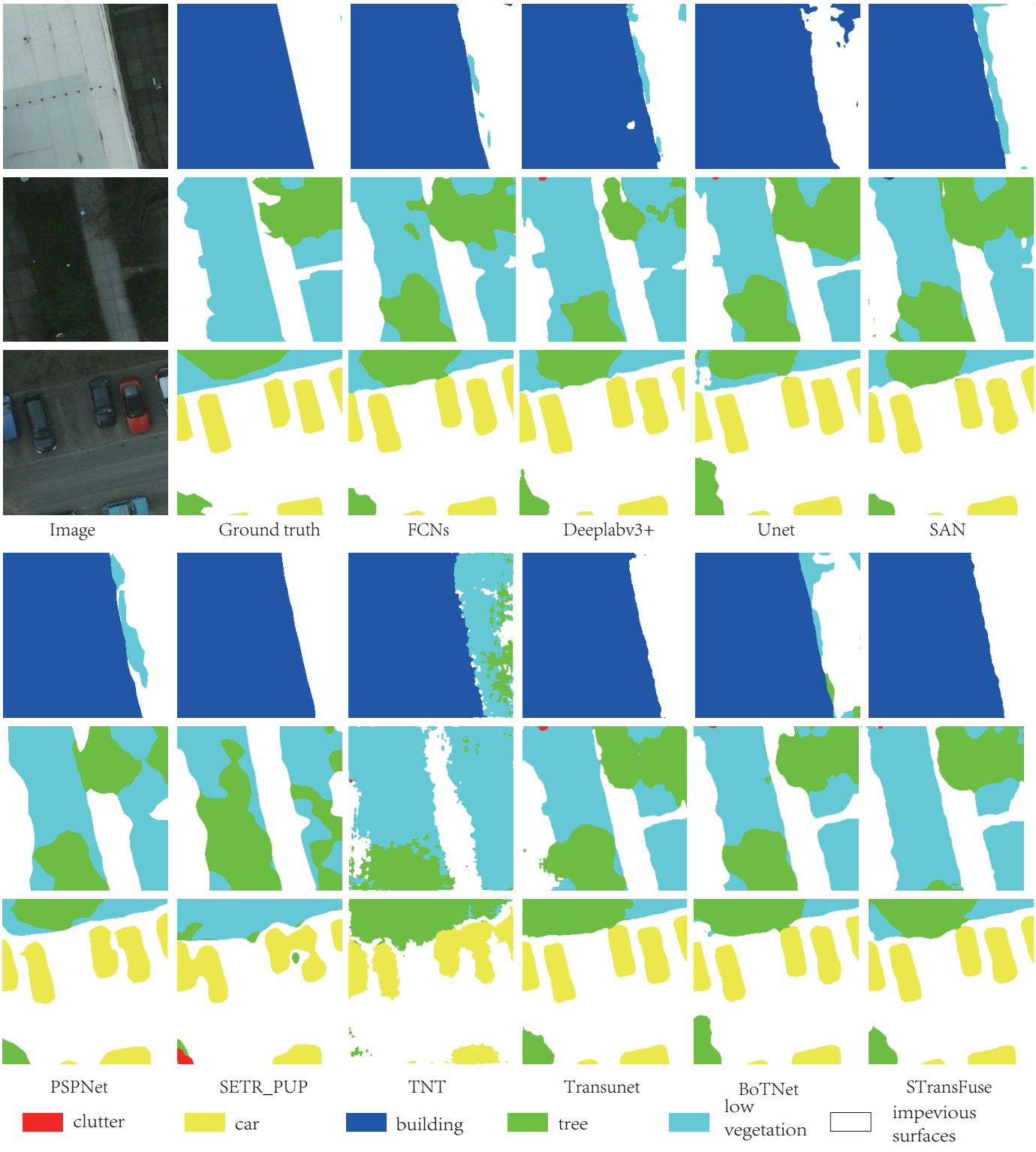

Fig. 8. Visualization results on Potsdam Dataset

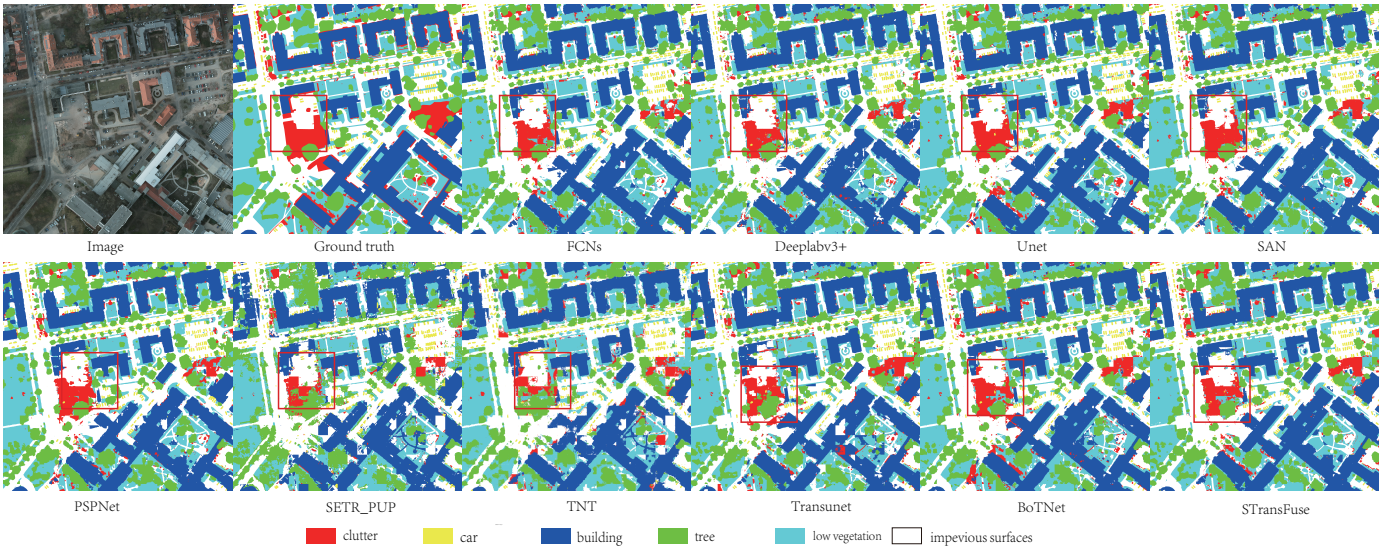

Fig. 9. Visualization results on Potsdam Dataset original image

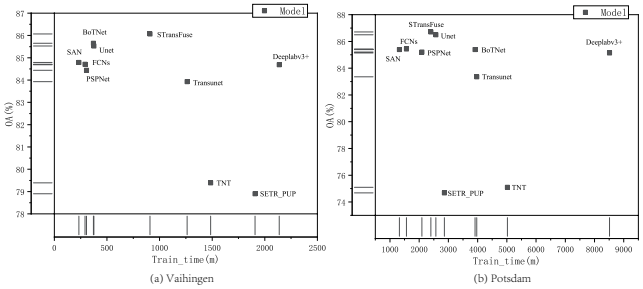

Fig. 10. Efficiency comparison of different models on Vaihingen and Potsdam datasets. The vertical axis represents the overall accuracy. The horizontal axis indicates the training time of the model (please note that except for TNT model and Transunet model, the batch size is 12, the default value of other models is 16)

the STransFuse model improved OA with a small increase in training time. Because of the ASPP module, the Deeplabv3+ model takes longer time in training and is thus less efficient. It is also seen from Fig. 10 that when applied directly to the semantic segmentation of remote sensing images, the performance of the model based on improved Transformer model is low. The experimental efficiency of different models on Potsdam is shown in Fig. 10 (b), and it can be observed from (b) that the STransFuse model reached a better OA in a shorter period of time.

VI. Conclusion

In this article, the STransFuse model, a combination of a Fusing Swin Transformer and CNN network, is proposed. This two-branch model takes the advantages of both Transformer network and CNN. Transformer can model the global correlation of input patches. CNN with pre-training weights can make up for the shortcomings of fewer remote sensing image training sets. The proposed model structure can make full use of feature information and detail information in all stages of feature maps to generate outstanding feature representation in the model. In addition, we created an AFM that can adaptively fuse features from the Transformer and CNN networks, resulting in a feature map input to the decoder that incorporates rich semantic and spatial context information. In comparison with some state-of-the-art models, the result from STransFuse model using the Vaihingen and Potsdam datasets is competitive. In the future, we will continue to research Transformer's applicability in the realm of remote sensing image processing to explore its potential.

Acknowledgment

The author thanks the International Society for Photogrammetry and Remote Sensing for the dataset provided.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017. [Online]. Available: https://arxiv.org/abs/1706.05587

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, Conference Proceedings, pp. 234–241.

[6] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, pp. 2547–2560, 2021.

[7] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 593–602.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, Conference Proceedings, pp. 213–229.

[11] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *European Conference on Computer Vision*. Springer, 2020, Conference Proceedings, pp. 323–339.

[12] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "Hsi-bert: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 165–178, 2019.

[13] H. Lin, X. Cheng, X. Wu, F. Yang, D. Shen, Z. Wang, Q. Song, and W. Yuan, "Cat: Cross attention in vision transformer," *arXiv preprint arXiv:2106.05786*, 2021. [Online]. Available: https://arxiv.org/abs/2106.05786

[14] Y. Fan, Y. Qian, L. Yang, and Z. HUANG, "Cotton recognition method for remote sensing image based on bp neural network," *Comput. Eng. Design*, vol. 5, no. 16, pp. 1356–1360, 2017. [Online]. Available: https://en.cnki.com.cn/Article_en/CJFDTotal-SJSJ201705044.htm

[15] W. Shi, M. Zhang, H. Ke, X. Fang, Z. Zhan, and S. Chen, "Landslide recognition by deep convolutional neural network and change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, pp. 4654 – 4672, 2020.

[16] J. Lin, W. Jing, and H. Song, "San: Scale-aware network for semantic segmentation of high-resolution aerial images," *arXiv preprint arXiv:1907.03089*, 2019. [Online]. Available: https://arxiv.org/abs/1907.03089

[17] Y. Chong, X. Chen, and S. Pan, "Context union edge network for semantic segmentation of small-scale objects in very high resolution remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.

[18] S. Xiang, Q. Xie, and M. Wang, "Semantic segmentation for remote sensing images based on adaptive feature selection network," *IEEE Geoscience and Remote Sensing Letters*, 2021.

[19] A. Li, L. Jiao, H. Zhu, L. Li, and F. Liu, "Multitask semantic boundary awareness network for remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1 – 14, 2021.

[20] R. Liu, L. Mi, and Z. Chen, "Afnet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1 – 16, 2020.

[21] X. Pan, J. Zhao, and J. Xu, "Conditional generative adversarial network-based training sample set improvement model for the semantic segmentation of high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1 – 17, 2020.

[22] Z. Yao, J. Jia, and Y. Qian, "Mcnet: Multi-scale feature extraction and content-aware reassembly cloud detection model for remote sensing images," *Symmetry*, vol. 13, no. 1, p. 28, 2021.

[23] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 416–12 425.

[24] W. Chen, X. Zhu, R. Sun, J. He, R. Li, X. Shen, and B. Yu, "Tensor low-rank reconstruction for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, Conference Proceedings, pp. 52–69.

[25] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7557–7569, 2020.

[26] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *European Conference on Computer Vision*. Springer, 2020, Conference Proceedings, pp. 108–126.

[27] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," *arXiv preprint arXiv:2102.10662*, 2021. [Online]. Available: https://arxiv.org/abs/2102.10662

[28] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 299–12 310.

[29] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sensing*, vol. 13, no. 3, p. 498, 2021.

[30] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," *arXiv preprint arXiv:2101.11986*, 2021. [Online]. Available: https://arxiv.org/abs/2101.11986

[31] Y. Jiang, S. Chang, and Z. Wang, "Transgan: Two pure transformers can make one strong gan, and that can scale up," *arXiv preprint arXiv:2102.07074*, 2021. [Online]. Available: https://arxiv.org/abs/2102.07074

[32] W. Wang, C. Chen, M. Ding, J. Li, H. Yu, and S. Zha, "Transbts: Multimodal brain tumor segmentation using transformer," *arXiv preprint arXiv:2103.04430*, 2021. [Online]. Available: https://arxiv.org/abs/2103.04430

[33] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, and P. H. Torr, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[34] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *arXiv preprint arXiv:2103.00112*, 2021. [Online]. Available: https://arxiv.org/abs/2103.00112

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[36] ISPRS, "Semantic labeling contest—potsdam," 2018. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[37] G. Deng, Z. Wu, C. Wang, M. Xu, and Y. Zhong, "Ccanet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1 – 20, 2021.

[38] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[39] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 519–16 529.

[40] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021. [Online]. Available: https://arxiv.org/abs/2102.04306

Liang Gao received a bachelor's degree in computer science and technology from Zaozhuang University, Shandong Province in 2019. At present, he is studying for a master's degree in software engineering at Xinjiang University. His research fields include deep learning and remote sensing image semantic segmentation.

Hui liu received his B.S. degree in Software Engineering from Xinjiang University, China in 2014. She received the Master of Engineering from the college of software, Xinjiang University, Urumqi, China, in 2017. She is currently pursuing a PH.D. degree in computer science and technology, Xinjiang University, Urumqi, China. Her research interests include deep learning and Opportunistic networks and the processing of remote sensing image data.

Minhang Yang graduated from Xi 'an University of Technology with a bachelor's degree in Software Engineering in 2019. She is currently pursuing a master's degree in software engineering at Xinjiang University. Her research interests include deep learning and multi-label image classification.

Long Chen received his bachelor's degree in geographic information science from Shandong University Of Science And Technology in 2018. At present, he is studying for a master's degree in software engineering at Xinjiang University. His research interests include deep learning and single image super resolution.

Yaling Wan received her bachelor's degree in communication engineering from Xinjiang University in 2014. At present, she is studying for a master's degree in software engineering at Xinjiang University. Her research interests include deep learning and hyperspectral image classification.

Zhengqing Xiao received his Ph.D from Beijing Normal University in 2011. He's currently working at the College of Mathematics and System Sciences of Xinjiang University. His research interests include big data analysis, image processing and complex system modelling.

Yurong Qian received bachelor's and master's degrees in computer science and technology from Xinjiang University in 2000 and a doctorate in biology from Nanjing University in 2010. From 2012 to 2013, she worked as a postdoctoral fellow in the Department of Electronics and Computer Engineering, Hanyang University, South Korea, and is currently a professor at the School of Software, Xinjiang University, China. She is a senior member of the Chinese Computer Federation. In 2015, she was trained as a young scientific and technological innovation talent by the Science and Technology Department of Xinjiang Province in China; her research interests include computational intelligence such as big data processing, image processing, and artificial neural networks.