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

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

***Abstract*—The applied research in remote sensing images has been pushed by convolutional neural network (CNN). Because of the fixed size of the perceptual field, CNN is unable to model global semantic relevance. Modeling global semantic information is possible with the self-attentive Transformer-based model. However, the method of patch computation used by Transformer for self-attentive computation ignores the spatial information inside each patch. To address these issues, we offer the STransFuse model as a new semantic segmentation method for remote sensing images. It is a model that combines the benefits of Transformer with CNN to improve the segmentation quality of various remote sensing images. We employ a staged model to extract coarse-grained and fine-grained feature representations at various semantic scales, unlike earlier techniques based on Transformer model fusion. In order to take full advantage of the features acquired at different stages, we designed an adaptive fusion module. This module adaptively fuses the semantic information between features at different scales employing a self-attentive mechanism. The overall accuracy (OA) of our proposed model on the Vaihingen dataset is 1.36% higher than the baseline, and 1.27% improvement in OA over baseline on the Potsdam dataset. When compared to other advanced models, the STransFuse model performs admirably.**

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

Manuscript received July 26, 2021; revised September 14, 2021; accepted October 10, 2021. Date of publication October 14, 2021; date of current version November 10, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61966035, in part by the National Science Foundation of China under Grant U1803261, in part by the Xinjiang Uygur Autonomous Region Innovation Team under Grant XJEDU2017T002, in part by the Autonomous Region Graduate Innovation Project under Grant XJ2021G062, and in part by the Autonomous Region Graduate Innovation Project under Grant XJ2020G074. *(Corresponding author: Yurong Qian.)*

Liang Gao, Minhang Yang, Long Chen, Yaling Wan, and Yurong Qian are with the College of Software, Xinjiang University, Urumqi 830008, China, and with the 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: gaoliang@stu.xju.edu.cn; yangminhang@stu.xju.edu.cn; ry19chenlong@stu.xju.edu.cn; wyl@stu.xju.edu.cn; qyr@xju.edu.cn).

Hui Liu is with the College of Information Science and Engineering, Xinjiang University, Urumqi 830014, China, and with the 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).

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

Digital Object Identifier 10.1109/JSTARS.2021.3119654

I. Introduction

A PIXEL-LEVEL classification challenge, semantic segmentation of remote sensing images, is an essential problem for remote sensing research. High-quality and high-resolution remote sensing images are now readily available because of advancements in remote sensing and sensor technology. However, remotely sensed images contain complex ground information with interclass and intraclass variation, which makes the research of remotely sensed images challenging.

In recent years, with the development of deep learning, remote sensing image research has also made great progress. In the field of computer vision, CNN-based enhanced models [1], [2] have demonstrated exceptional performance. Fully convolutional networks (FCNs) [3], developed on the basis of CNN, achieve pixel-level classification and further promote the application of deep learning models in the field of image segmentation. However, the scales of features contained in remotely sensed images vary greatly, and ground objects of the same class may have different shapes and sizes, and ground objects of different classes may have similar characteristics. Models using only spectral information are not sufficient to effectively distinguish ground objects [4], and multiscale contextual information is needed to assist in identification [5]. Therefore, how to effectively obtain the contextual information of images is a problem worth researching.

Due to the fixed receptive fields of convolutional kernels, FCNs are unable to collect visual contextual information well. To solve the problem of the lack of sensory fields, researchers use pooling methods. Through a deeper network, the model collects high-level feature maps rich in semantic information and decreases the feature maps' resolution to obtain a global representation of the feature information. However, in the process of continuous subsampling, the model loses some information. Several researchers [6]–[8] have tried to solve the above problem using the fusion of multiscale contextual information. Chen *et al.* [7] improved the atrous spatial pyramid pooling (ASPP) module to capture multiscale contextual information by combining atrous convolution. Unet [9] with an encoder–decoder structure acquires feature map information at different levels by skip-linking in order to enhance the representation of feature map information. Pyramid scene parsing network (PSPNet) [8] aggregates the ability to extract global contextual information based on different regions through a pyramid pooling module.

Although the integration of multiscale contextual information aids in the collection of ground objects at many scales, it ignores ground object relationships [10]. Each ground object has different relationships with other ground objects, and these relationships can be used for better contextual modeling.

Besides, some researchers [11], [12] tried to use self-attention [13] to solve the problem of the lack of model-receptive fields. Fu *et al.* [11] designed compact position attention module and compact channel attention module based on the self-attentive mechanism to model semantic relevance from spatial and channel dimensions, respectively. The Transformer method based on sequence-to-sequence prediction [14]–[18] has shown outstanding performance in the field of computer vision. Transformer discards convolutional operations in its structure and adopts a structure of pure attention mechanism. Transformer, unlike CNN, is capable of acquiring global contextual knowledge through self-attention. There are experiments [14] that demonstrate that in the case of large-scale pretraining experiments, Transformer is able to reach state-of-the-art in image tasks such as image classification, image recognition, and semantic segmentation.

In this article, we explore the application potential of Transformer for semantic segmentation in the context of remote sensing. Interestingly, when we transfer some Transformer models that achieve outstanding results on public datasets to remote sensing image datasets, we find that these network models do not produce satisfactory results. This is because during the input of patches to the Transformer network, patches are compressed into a sequence of 1-D and the structural information in patches is lost. In the encoder stage of the Transformer network, the detailed information cannot be recovered effectively by upsampling, which eventually leads to poor segmentation of the model. Inspired by the Unet network [9], we fused the feature maps of different stages, which were used to obtain semantic contextual information and spatial contextual information of the images.

To this end, we propose a model for semantic segmentation of remote sensing images, STransFuse. STransFuse uses the architecture of Swin Transformer [19] combined with CNN. The Swin Transformer branch acquires features in the form of shift windows to build self-attentiveness. The CNN branch acquires spatial contextual information. The success achieved by Transformer relies on the training of a large amount of data. However, the limited image datasets acquired in the field of remote sensing greatly limit the application of Transformer in the field of remote sensing. Inspired by the paper [14], we used Resnet34 with pretrained weights as the network backbone of the CNN branch, combined with Swin Transformer, to obtain the rich feature information of remote sensing images. At the same time, the designed structure of the staged fusion model allows the model to acquire rich features, and the perceptual field of the model increases as the layers of the STransFuse model are deepened and used to acquire contextual information, and our main contributions are as follows.

1) A framework for Swin Transformer and CNN parallelism was designed with STransFuse model. First, STransFuse brings the global semantic information of feature maps into the model through Transformer network and uses CNN to extract the spatial contextual information of low-level feature maps. Second, STransFuse avoids the difficulties of gradient disappearance and feature map information loss by not requiring the creation of very deep networks. Then, it is experimentally demonstrated that the adopted two-branch structure makes the model efficient in terms of performance and computational speed. Finally, the designed phased fusion structure allows the high-level feature map to contain richer feature information.
2) A feature map fusion module is designed to improve the model feature representation by adaptively fusing feature maps with self-attentive structures.
3) The proposed STransFuse model achieves a comparatively good result on the Vaihingen and Potsdam datasets.

The rest of this article is organized as follows. 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 adaptive fusion module (AFM) design are explored. The datasets used in the studies, as well as the experimental parameters, are described in Section IV. Section V presents complete ablation studies and experimental comparison between the STransFuse model and some state-of-the-art models to validate the proposed module. Finally, Section VI concludes this article.

II. Related Works

A. Semantic Segmentation of Remote Sensing Images

Remote sensing images are widely used in many application fields, including crop yield estimation [20], military reconnaissance, and natural disaster monitoring [21]. 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 require 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. For remote sensing images, deep learning-based semantic segmentation algorithms [22]–[25] have sprouted up. AFNet [26] employed the scale-feature attention module and scale-layer attention module to better tackle the difference between intraclass and interclass in remote sensing images. Pan *et al.* [27] introduced a conditional generative adversarial network that actively generates new sample images while extracting advanced spatial information from previous training images. To overcome the problem of cloud segmentation in remote sensing images, Yao *et al.* [28] presented a multiscale feature extraction and content-aware recombination network. Mou *et al.* [29] designed the spatial relation module and the channel relation module, through which the relationship between any two spatial locations or feature maps is learned and inferred. The contextual information collected by these models, however, is insufficient since the spatial dependence of ground objects in high-resolution remotely sensed images also plays a significant role. As a result,

the segmentation of certain finely structured objects (e.g., car) remains poor.

B. Contextual Information

To increase the accuracy of image semantic segmentation, it is critical to understand how to properly extract the image's contextual information. FCNs [3] 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. Some researchers have attempted to alleviate the problem of a lack of sensory field by fusing multiscale contextual information. Unet [9] 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 research [30] offered an axial-attention model to enlarge the perceptual area and alleviate the problem of remote contextual information being lost due to convolution. Although the multiscale context fusion approach aids in the acquisition of contextual information, it ignores the pixels' relationship. The self-noticing-based Transformer is able to model global semantic relevance, and some researchers have tried to use Transformer's approach to obtain contextual information of images. Feature Pyramid Transformer [16] contains a fully active feature interaction across both space and scales, by designing a pyramid-like structure to expand the perceptual field.

The contextual information in remote sensing images describes the relationship between objects. Contextual information of remote sensing images is difficult to obtain because of the high resolution of remote sensing images and the unbalanced proportion of the ground object contained in the images. Remote sensing images are generally difficult to process directly, and often require preprocessing (cropping, normalization) of the images. 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 class of a ground object. When the common model executes the convolution operation, the proportion of large-scale ground objects in the patch can be substantially higher than that of small-scale ground objects, causing small-scale ground objects to be heavily influenced by large-scale ground objects. According to the characteristics that remote sensing images have intraclass differences and interclass differences, balancing the accuracy and efficiency of remote sensing image processing has become a hot spot in the current research of remote sensing images. Our research focuses on increasing the model's feature extraction capabilities and the integrated use of feature maps from multiple phases of the model due to the complexity of remote sensing images. The STransFuse model was created based on these findings.

C. Transformer

The Transformer was originally used in the realm of natural language processing (NLP) [13]. 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) [14], 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 [31]–[36] tried many ways to make the Transformer more successful in the field of computer vision, inspired by the construction of the visual Transformer model. Semantic SEgmentation Transformer (SETR) [37] 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) [15] is a Transformer that was developed by Facebook AI researchers and applied to a visual model. It is 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. However, these visual transformers usually treat the image as a series of patches, ignoring the intrinsic structural information within each patch. The Transformer-in-Transformer (TNT) model [38] 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 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 ground objects of remote sensing images are quite different from those of ordinary images. As inspired by the ViT model [14], we combined the pretrained 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

We present the STransFuse model as a new semantic segmentation method for effectively obtaining global semantic context information and spatial context information in remote sensing images. We use Swin Transformer and CNN to handle the images, fuse the feature maps at different stages, and finally restore the feature maps to their original size. In Section III-B, we will introduce the overall structure of STransFuse. Then, the details of Swin Transformer are given in Section III-C. Finally, the AFM is described in Section III-D.

B. STransFuse Overall Architecture

In order to improve the accuracy of model segmentation, semantic and spatial contextual information in images is essential. CNN is limited by the fixed size of convolutional kernels and cannot model global semantic information. Transformer can obtain global semantic information by self-attentive computation, but self-attentive computation needs to stretch the patches into 1-D tokens, and the spatial information inside the patches

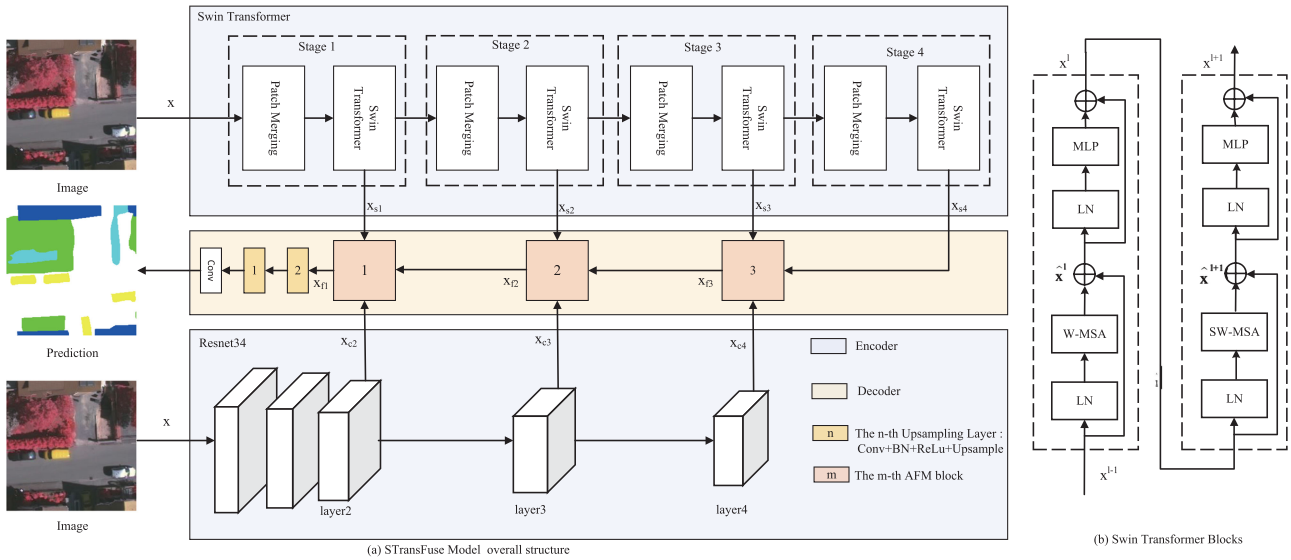


Fig. 1. (a) The overall structure of STransfuse model. (b) The detail of Swin Transformer blocks.

is lost. In order to solve these problems, we designed models with two encoders to extract features. We used Resnet34 with training weights as the encoder for the CNN branch. The feature representation capability of the model is improved by fusing the features extracted from the Transformer branch and the Resnet34 branch.

As shown in Fig. 1(a), the image $x \in R^{H\times W\times C}$ is input into the Swin Transformer network and the Resnet34 network, respectively, where H represents the height of the image, W represents the width of the image, and C represents the number of channels in the image.

The feature maps extracted by the model at different stages are of different sizes, corresponding to different semantic scales of feature granularity. In order to enable the feature maps to contain rich semantic information as well as feature detail information, we use a staged fusion strategy.

There are four 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. Patch merging splits the image into nonoverlapping patches by sliding the window on the input image. 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. The resolution of the output feature map is $\frac{H}{4}\times\frac{W}{4}$. The above steps 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, as illustrated in Fig. 2. Patch merging applies linear embedding layer to change the output dimension to 2C, and applies Swin Transformer for feature transformation. In "Stage2," the resolution of the output feature map is maintained at $\frac{H}{8}\times\frac{W}{8}$. "Stage 3" and "Stage 4" are similar to "Stage 2," and the output feature map resolutions are $\frac{H}{16}\times\frac{W}{16}$ and $\frac{H}{32}\times\frac{W}{32}$, respectively.

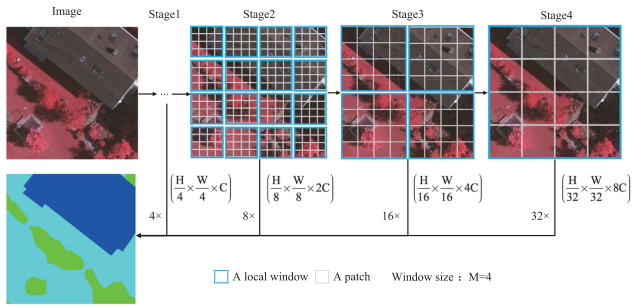


Fig. 2. Swin Transformer builds hierarchical feature maps by merging image patches in deeper layers.

The images are input to the Resnet34 network to get the feature maps, which are output by layer2 to layer4 as feature maps x_{c2}, x_{c3}, and x_{c4} respectively, and the sizes of these feature maps are $\frac{H}{8}\times\frac{W}{8}$, $\frac{H}{16}\times\frac{W}{16}$, 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 by a factor of two, and the feature map is restored to the size of the input image.

C. Swin Transformer Block

The image's contextual information is critical for improving semantic segmentation accuracy, and long-range semantic information can be employed as a discrimination aid, allowing the model to rely on more than just the image's spectral information. Therefore, we introduced the Transformer network as a subnetwork in the model for feature extraction. There have been many studies demonstrating that self-attention in Transformer can model global semantic information [11], [39], [40].

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. The images are segmented by nonoverlapping windows. Each window contains M × M patches. It is worth noting that to ensure that the images (h × w) are divisible by the window (M × M), we use a padding method and mask the padding values when computing attention. In this case, the computational complexity of multihead self attention (MSA) is shown in (1), and the computational complexity of window MSA (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 C denotes the dimension, and *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*.

Swin Transformer is the replacement of MSA in the Transformer module with W-MSA. 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.

In summary, the process of calculating the feature map in the W-MSA module is shown below:

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

The feature map then passes through a linear batch layer and a fully connected layer with the following equations:

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

where $\hat{x}^{l}$ denotes the output characteristics of the W-MSA module of l block, and x^{l} represents the output characteristics of the MLP module after l block. LN denotes layer normalization and MLP denotes multilayer perceptron.

Because of the sliding-window segmentation operation performed by W-MSA, the cropped patches do not overlap and there is a lack of effective information interaction between the windows. A shifted window MSA (SW-MSA) network exists to further increase the model's performance. SW-MSA performs window shifting compared to W-MSA. The idea of SW-MSA is to move the image cyclically up and cyclically left by half the window size. The area on the image beyond the window will be moved to the lower and right side of the window, respectively. Then, by slicing the window according to W-MSA on top of the shift, we will get a different window-slicing method than W-MSA. The formula for SW-MSA is shown below:

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

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

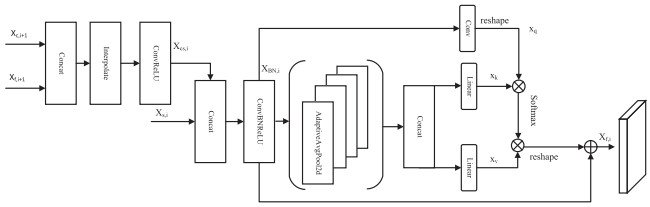


Fig. 3. Detail display of AFM.

where $\hat{x}^{l+1}$ denote the output characteristics of the SW-MSA module of l+1 block, and x^{l+1} denote the MLP module of l+1 block.

D. AFM Block

To efficiently fuse the encoded features from CNN and Swin Transformer, we designed an AFM based on the self-attentive mechanism. The feature weight matrix is obtained by self-attentive calculation to selectively enhance spatial details or suppress other regions, thus enhancing the differentiation ability of dense prediction. The structure of AFM is shown in Fig. 3. We will perform the fusion of features with the following equation:

$$x_{cs,i} = \text{Re}\,LU(Conv(Interpolate(concat(x_{f,i+1}, x_{c,i+1})))) \tag{7}$$

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

$$x_q = Conv(x_{BN,i}) \tag{9}$$

where $x_{f,i}$ represents the feature matrix of the output of the ith stage of the AFM, $x_{c,i}$ represents the feature matrix of the output of the ith layer of the CNN, $x_{s,i}$ represents the feature matrix of the output of stage i of Swin Transformer, $x_{BN,i}$ represents the ith AFM block fusion feature map, and x_q is the query in the self-attentive calculation.

The feature fusion is computationally intensive due to the use of feature maps from three branches. To alleviate this problem, we add the AdaptiveAvgPool2d method to the AFM, enabling the AFM to construct a relationship between each pixel and some convergence centers. By this collection of feature vectors from a subset of pixels in the input tensor, the AFM is made computationally acceptable. The formula for calculating semantic relevance in AFM is shown below:

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

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

$$x_{f,i} = (x_q \otimes x_k) \otimes x_v \oplus x_{BN,i} \tag{12}$$

where x_k represents the key in self-attention calculation, x_v represents the value in self-attention calculation, $x_q \bigotimes x_k$ gets the self-attention weight matrix, $(x_q \bigotimes x_k) \bigotimes x_v$ obtains the weighted feature matrix, and add the weighted feature matrix and the fusion feature matrix to obtain $x_{f,i}$. The feature value of each position in $x_{f,i}$ is the weighted sum of the features of all positions and the original features. Thus, AFM is able to selectively aggregate contextual information based on the attentional feature map in the global view, improving the model's

ability to discriminate between dense pixels. The information of each pixel will be passed to the pixel associated with its semantics, which improves the semantic consistency.

IV. Dataset Description and Design of Experiments

The ISPRS Vaihingen and Potsdam datasets are utilized for testing to validate the effectiveness of the suggested method. We began with a brief description of the datasets and an introduction of the experiment's specifics in this section.

A. Dataset

1) Vaihingen: There are 33 patches in the Vaihingen dataset [41]. 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, 16 patches were used as the training set (image id: 1, 3, 5, 7, 11, 13, 15, 17, 21, 23, 26, 28, 30, 32, 34, 37), and the other 17 as the test 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 [42] semantic annotation collection consists of 38 patches, each with a GSD of 5 cm and a resolution of 6000×6000 pixels. We used RGB images as the dataset instead of nDSM or DSM data. We followed the official division principle and used 13 (because the provided label dataset is missing 03_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, which was also used in the papers [25], [27], [43], [44]. We used intersection over union (IoU) for each class, F1-score for each class, mean intersection over union (mIoU), mean F1-score, and 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 (15)

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

The F1-score formula for each class is defined as shown in (16)

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

The mean F1-score is obtained by averaging the F1-score of each class. 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 $\mathbb{N}_{\text{p}}$ represents the prediction set, and $\mathbb{N}_{\text{gt}}$ 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)

$$W_{\text{class}} = \frac{1}{\log(P_{\text{class}} + c)}, c = 1.12. \tag{18}$$

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 11 G.

V. Experimental

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

A. Ablation Studies

In this section, we discuss the experimental performance of the single-branch model and the two-branch model. Also, we discuss the experimental results of putting Resnet34 in serial and parallel with Transformer. In addition to this, we will discuss the effectiveness of the designed phased integration strategy. Finally, we also discuss the effectiveness of the proposed AFM through experimental results. We perform experimental comparisons in the Vaihingen and Potsdam datasets.

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, WHERE THE METRICS OA, MF1, AND IOU ARE IN PERCENTAGES; T REFERS TO THE TIME OF MODEL TRAINING IN MINUTES (MIN). PARA REPRESENTS THE NUMBER OF PARAMETERS OF THE MODEL IN M

Method	$x_{s,1}$	$x_{s,2}$	$x_{s,3}$	$x_{s,4}$	concat	AFM	Para (M)	Vaihingen mF1	Vaihingen mIoU	Vaihingen OA	Vaihingen T(min)	Potsdam mF1	Potsdam mIoU	Potsdam OA	Potsdam T(min)
FCNs							22.06	76.57	63.85	84.71	**292**	80.36	69.19	85.44	**1566**
Swin_xs4				√			36.29	71.77	58.87	83.4	626	76.26	63.93	82.46	2374
Swin	√	√	√	√	√		33.38	73.04	60.18	83.79	653	76.82	64.82	82.95	2811
Res34+Swin	√	√	√	√	√		59.42	77.17	64.76	84.91	1176	80.16	68.89	85.28	1989
Swin+Res_xs4				√	√		59.3	76.93	64.61	85.27	8078	81.24	70.27	86.15	2529
Swin+Res34	√	√	√	√	√		80.78	77.95	66	85.94	1555	81.61	70.92	86.47	2747
STransFuse	√	√	√	√		√	86.99	**78.67**	**66.66**	**86.07**	909	**82.08**	**71.46**	**86.71**	2399

1) Quantitative Analysis in Ablation Studies: The experimental results of the ablation experiments are shown in Table I, and FCN [3] (Resnet34) was chosen as the baseline model for comparison.

Table I summarizes 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 stage 4 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 stages of Swin Transformer network and Resnet34. Swin+Res_xs4 represents the fusion using only the feature maps $x_{s,4}$ and $x_{c,4}$. 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 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 (FCNs, Swin_xs4, Swin, Res34+Swin) perform worse for semantic segmentation of images than two-branch network models (Swin+Res34, Swin+Res_xs4, STransFuse).

At the same time, we compared the experimental results of using only the features ($x_{s,4}$) output from the final stage of Swin Transformer and fusing features of different stages ($x_{s,1}$, $x_{s,2}$, $x_{s,3}$, $x_{s,4}$). The testing results demonstrate that the Swin model, which combines the feature maps of several stages of the Swin Transformer, can enhance metric OA by 0.49% in the Potsdam dataset.

Then, we compared the experimental effects of connecting the Resnet34 network with pretrained weights to the Transformer model in series (Res34+Swin) and in parallel (Swin+Res34, Swin+Res_xs4, STransFuse). The parallel network model performs better in the experiments, as seen in Table I. In the Potsdam dataset, Swin+Res34 is 1.19% OA higher than Res34+Swin. The effectiveness of the staged fusion strategy can be found by comparing the experimental results of Swin+Res_xs4 and Swin+Res34.

Finally, we compared the AFM and concat modules' performance. As shown in Table I, the model (STransFuse) employing our proposed AFM for feature map fusion performs 0.24% OA better in the Potsdam dataset than the model (Swin+Res34) 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 images, it stretches the patch into a 1-D 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.

Besides, Table I shows that our design STransFuse model has a shorter training time than other parallel models (Swin+Res34, Swin+Res_xs4).

2) Qualitative Analysis in Ablation Studies: It can be seen clearly from Fig. 4 that the STransFuse model segmented better than the baseline network FCNs, and that the STransFuse model did not misclassify the ground objects with shading effects in a 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 a row (b).

B. Visualization Analysis

To further illustrate the ability of our STransFuse model to effectively acquire feature information in remotely sensed images, we compared the recognition capabilities of the benchmark model FCNs and STransFuse models for different classes of the ground objects. We visualized the last convolutional layer in the FCNs model and the STransFuse model using the class activation mapping (CAM) approach, respectively. CAM, originally proposed in the paper [45], is a weighted linear sum of these visual patterns that are at different spatial locations. It is the multiplication of the weights corresponding to a certain class by the layer corresponding to the feature map, normalized by the heat map, as shown in the second row of the heat map in Fig. 5.

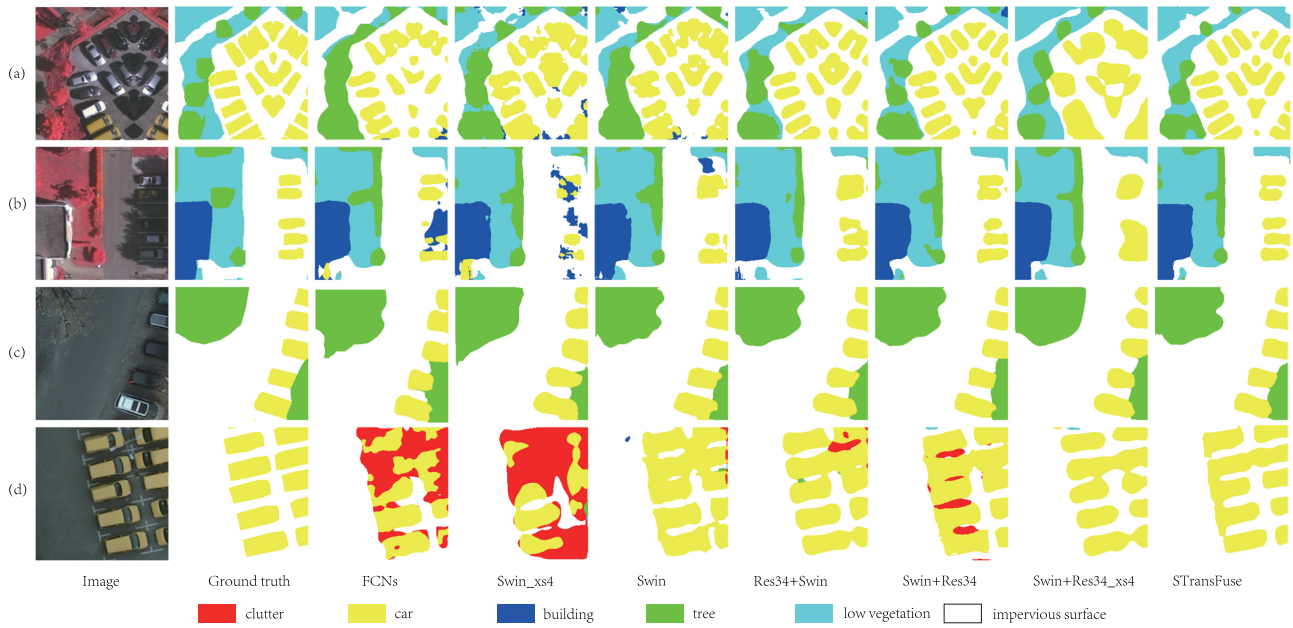


Fig. 4. The result figures of ablation studies visualized in different datasets. (a) and (b) Results in the Vaihingen dataset. (c) and (d) Results in the Potsdam dataset.

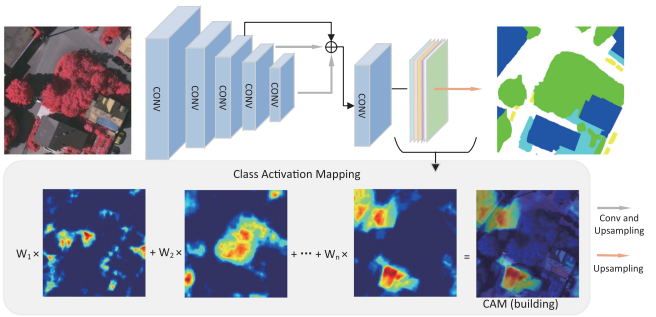


Fig. 5. Class activation mapping: The predicted class scores are mapped back to the previous convolutional layers to generate CAMs. The highlighted areas in the image (red) represent the areas that the model focuses on for specific classes, while the areas in the image that are dark (dark blue) represent the areas that the model does not focus on.

By simply upsampling the CAM to the size of the input image, we can identify those regions that the model focuses on. As shown in the figure, we show the CAM implementation process for the building class, using the FCNs model as an example.

As shown in Fig. 6, 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 the building. Because the ground objects are captured using an overhead view, the lack of ground object height information in the image causes the tops of buildings and impervious surfaces to have a similar texture representation. Therefore, the FCNs model appeared the phenomenon of "car flying on the roof of the building" in the recognition image. However, due to the use of self-attention, the STransFuse model modeled the long-range semantic correlation and determined the class information of similar semantics. Therefore, the STransFuse model can recognize semantic information better. In the column where the class car is located, the FCNs did not identify all cars and were not accurate enough in the already identified car boundary information, compared to the STransFuse model which is also good at identifying ground objects with small scale like the car. In the column where the impervious surface is located, it is shown that FCNs recognized some car's semantic information as impervious surface. This is because the car occupies a smaller proportion of the image compared to the impervious surface, and impervious surfaces enclose the car. There is no correlation between car and car. This is a common interclass imbalance in remote sensing images, which occurs because remote sensing images often span a wide range of locations, and larger objects can fill a larger proportion of the image, whereas smaller scale ground objects can 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 classes [22]. The Transformer branch we use can effectively solve this type of problem. The Transformer can obtain a weight matrix by self-attentive computation, which adaptively enhances or attenuates the feature values, making the class represented by the pixel values more accurate. The small interclass differences between tree and low vegetation can be seen through Fig. 6. 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 ground objects.

C. *Window Size Impact Analysis*

In this section, we discuss the effect of the size of the local window on the experimental results. We first compare the effect of local windows on the model performance when the size of M is 4, 8, 7, and 10. Then, we compare the effect of local windows

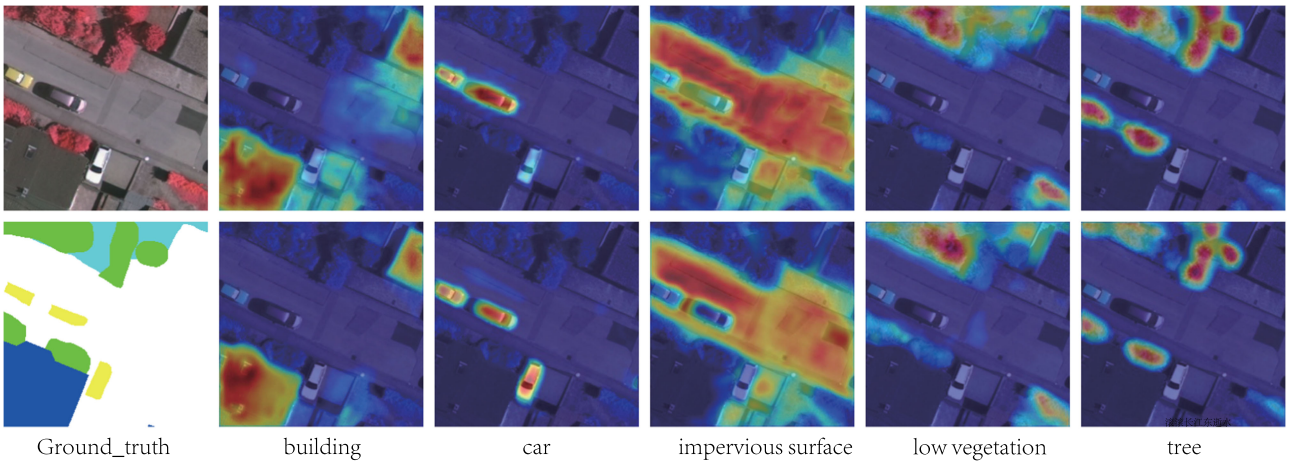


Fig. 6. The class activation mapping of different classes of features. The image of the first row is generated by FCNs and the second row is for STransFuse.

TABLE II
DISCUSSION ON THE EFFECT OF WINDOW SIZE ON EXPERIMENTAL RESULTS. THE BOLD VALUES ARE THE BEST, AND THE UNDERLINED VALUES ARE THE SECOND BEST, PARA (M), T (MIN)

Method	M	Patch size	Para	Vaihingen mF1	Vaihingen mIoU	Vaihingen OA	Vaihingen T	Potsdam mF1	Potsdam mIoU	Potsdam OA	Potsdam T
STransFuse	4	4×4	86.97	78.59	66.67	**86.15**	862	81.71	71.02	86.38	2897
	8	4×4	86.99	78.76	66.74	85.95	1006	81.69	71.04	86.49	2952
	10	4×4	87.01	**79.24**	**67.22**	85.91	1336	81.8	71.2	86.52	1950
	7	4×4	86.99	78.67	66.66	86.07	909	**82.08**	**71.46**	**86.71**	2399
	7	8×8	86.99	78.92	66.92	85.96	1377	81.91	71.35	86.63	1392

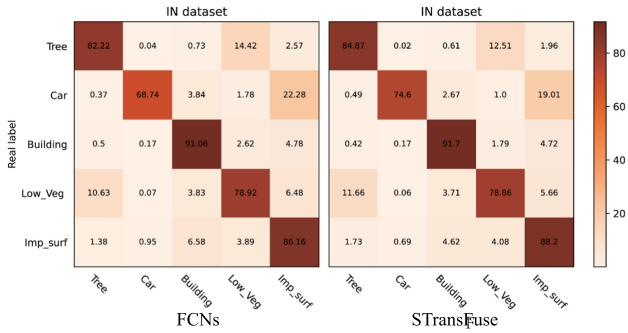


Fig. 7. Confusion matrixes of a sample of Vaihingen dataset with FCNs and STransFuse.

on model performance when the patch size is 4 × 4 and 8 × 8. The experimental results of the model on the Vaihingen and Potsdam datasets are shown in Table II.

As we can see in Table II, the difference in *M* size has an impact on the parameters of the model, the segmentation performance, and the training time. A larger value of *M* represents a larger window for local computation and therefore a larger number of parameters for the model. In the Vaihingen dataset, the model with a window size of 10 × 10 achieves the best performance on the metrics mF1 and mIoU with 79.24% and 67.22%, respectively. In addition to this, the size of the patches also had an impact on the experimental results. As can be seen in Table II, the model with a patch size of 8 × 8 was able to achieve 78.92% mF1 and 66.92% mIoU in Vaihingen and a suboptimal result in the Potsdam dataset.

D. Confusion Matrix

Fig. 7 shows the confusion matrix generated after the completion of the test on the Vaihingen dataset. The proportion of accurately predicted classes of the images to total predicted classes 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. 7, and small-scale car are easily labeled as large-scale impervious surface. To some extent, the STransFuse model solves this problem.

E. Evaluation and Comparisons on the Vaihingen Dataset

We compared the performance of our STransFuse model with other state-of-the-art models (Deeplabv3+, Unet, PSPNet) based on pretrained Resnet34 on the Vaihingen Dataset. Then, there are models based on Transformer improvements (BoTNet, SETR_PUP, TNT, Transunet), the design details of which are all mentioned in Sections I and II. Furthermore, we have undertaken experimental comparisons with models that have used the same dataset as ours in recent years, and the design details of these models are presented below.

1) Scale-aware network (SAN): Lin *et al.* [22] presented SAN in 2019. SAN uses a resampling approach with the aim of enabling pixels to adjust their position to different scales of ground objects and to implicitly introduce spatial attention by using resampled maps as weighted maps.
2) Dilated convolutions' merging network (DDCM-Net): Liu *et al.* [46] proposed DDCM-Net in 2020. The proposed DDCM-Net consists of a dense convolution of atrous images and different atrous rates, which effectively extends the perceptual field of the model.

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

Method	Tree		Car		Building		Low Veg		Imp surf		mF1	mIoU	OA
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU			
Deeplabv3+ [47]	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 [9]	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
PSPNet [8]	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 [48]	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 [37]	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 [38]	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 [49]	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
SAN [22]	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
DDCM-Net [46]	85.49	74.66	77.89	63.78	91.61	84.51	78.25	64.28	88.21	78.9	**79.23**	**67.18**	85.99
HCANet [5]	85.28	74.33	**78.18**	**64.17**	**91.79**	**84.82**	78.16	64.15	88.21	78.9	78.17	66.24	85.94
Capsules–Unet [4]	70.9	54.92	2.56	1.3	61.32	44.22	32.54	19.43	69.18	52.88	39.44	52.88	61.47
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**

3) Context aggregation network (HCANet): Bai *et al.* [5] presented HCANet in 2021. HCANet has an encoder–decoder structure similar to that of UNet. The researcher designed the compact atrous spatial pyramid module to extract contextual information for multiple semantic features and the compact atrous spatial pyramid+ module to aggregate contextual information.
4) Capsules-Unet: Guo *et al.* [4] presented this model in 2020. Capsules are incorporated into the U-net architecture for remote sensing image classification, and capsules are used to encapsulate the multidimensional properties of objects in order to train better models.

To be fair, we put these models in the same experimental setting and did not use other kinds of fancy tuning methods.

1) Quantitative Comparison: Table III shows the results of the comparative experiments. It can be seen from Table III that the STransFuse model can achieve the best results. Although Deeplabv3+ [47] 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 Fig. 10. The Deeplabv3+ model's overall efficiency is low. BoTNet [48] replaced the last three bottleneck blocks in Resnet with a global attention module, and was implicitly regarded as multihead 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 [38] has poor experimental results. The Transformer was used as an encoder in the Transunet model [49] 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 (FCNs, Deeplabv3+, Unet, SAN, PSPNet) and Transformer-based improved models (BoTNet, SETR_PUP, TNT, Transunet), the STransFuse model achieved better performance.

Comparing the models using the same dataset as ours, the model we designed improves 1.28% over SAN in OA, 0.08% over DDCM-Net in OA, 0.13% over HCANet in OA, and 24.6% over Capsules-Unet in OA. Overall, our STransFuse model, which gets the first best score on OA, achieves the second best score on mF1 and mIoU.

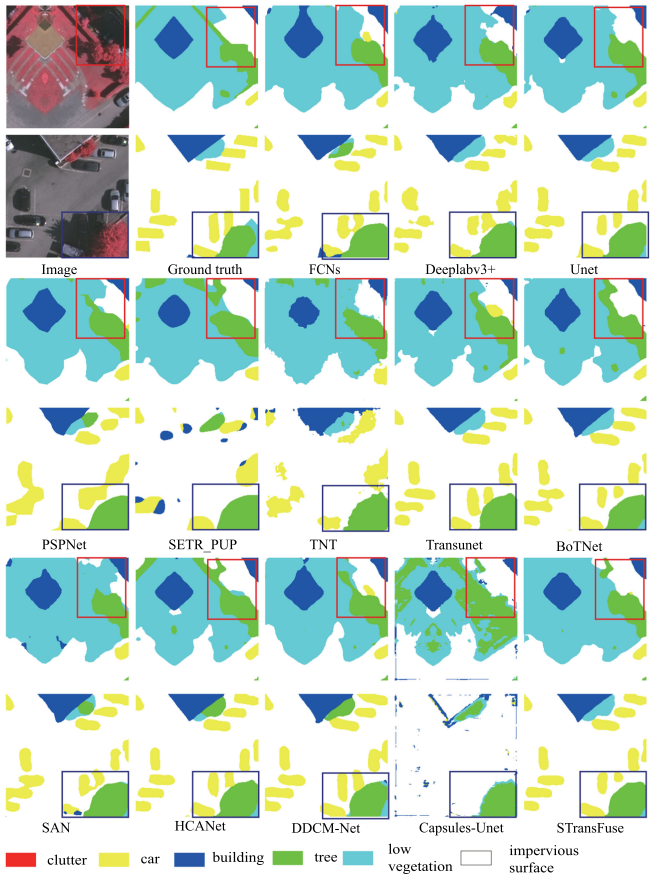


Fig. 8. Visualization results on Vaihingen dataset.

2) Qualitative Comparison: On the Vaihingen dataset, the qualitative comparison results are displayed in Fig. 8. As shown in Fig. 8, the STransFuse is capable of recognizing a variety of target classes. It benefits from the Transformer network's capabilities, such as improved global context modeling

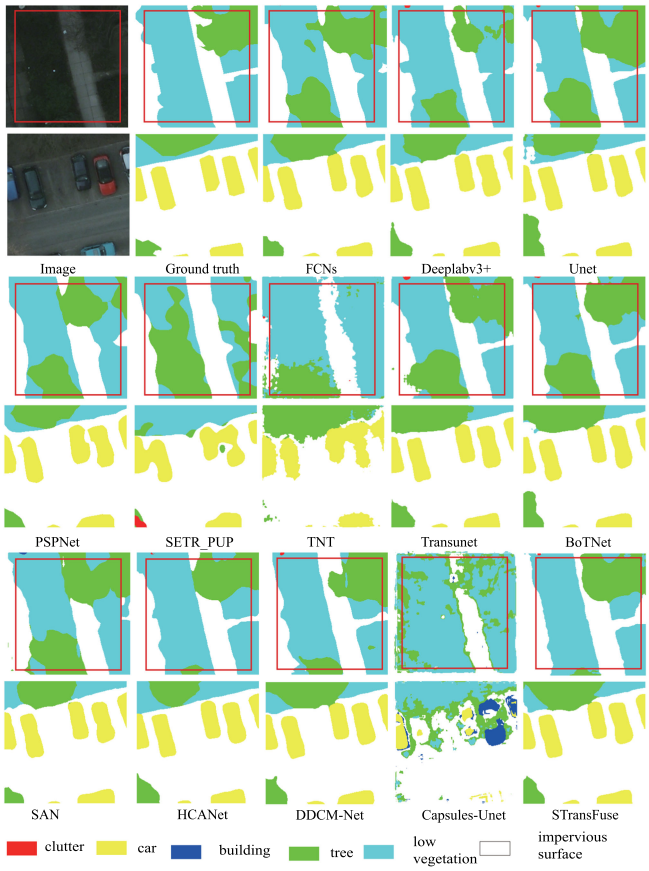


Fig. 9. Visualization results on Potsdam dataset.

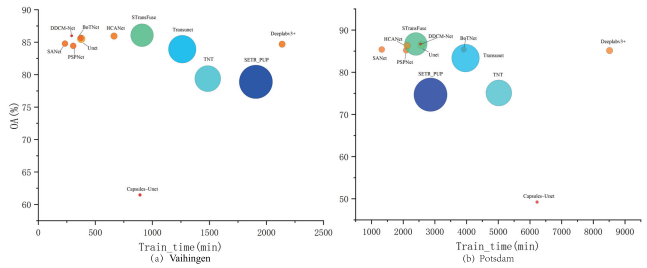


Fig. 10. Efficiency comparison of different models on Vaihingen and Potsdam datasets. The vertical axis represents the overall accuracy. The horizontal axis indicates the training time of the model. The size of the circle indicates the number of model parameters (note that except for TNT model and Transunet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

efficiency without sacrificing low-level detailed localization capability. Furthermore, we discovered that the model with a pure Transformer encoder (SETR_PUP, TNT) incorrectly recognizes a building as an impervious surface. The reason for this phenomenon may be that the two classes of ground objects, building and impermeable surface, have similar characteristics. When Transformer stretches the patch into a 1-D token, the difference between the building and impermeable surface feature values in the token is not significant. When calculating the similarity, the self-attention judges the two types of ground objects as the same type. When the characteristics are very different, the Transformer can distinguish them.

Besides, we can see that the shadows in the image have a large impact on the model recognition performance. For example, in the red box area in Fig. 8, the shadow of the building makes it difficult for the model to extract the features of the road. As can be seen in Fig. 8, shading has a significant impact on all models. The shadow region of an image can be thought of as a low-illumination image [50], with buried ground object information and blurring edges as issues. Then, in the blue box of the image, we can see that there is a tree and low vegetation tightly surrounded by each other, and the human eye's recognition ability has made it difficult to discern the border between the tree and the low vegetation. Without the inclusion of new information, the model has a harder time distinguishing between two types of ground object boundaries. The two issues mentioned above are equally tough to solve in the field of remote sensing image applications research.

F. Evaluation and Comparisons on the Potsdam Dataset

1) Quantitative Comparison: Table IV shows that the STransFuse model is able to get the highest OA score on the Potsdam Dataset. Comparing the models using the same dataset as ours, the model we designed improves 1.32% over SANet in OA, 0.02% over DDCM-Net in OA, 0.38% over HCANet in OA, and 37.48% over Capsules-Unet in OA. Overall, our STransFuse model achieves the first best score on OA and the second best score on mF1 and mIoU.

2) Qualitative Comparison: The results of the qualitative comparison of the models are displayed in Fig. 9, where it can be shown that the STransFuse model performed well for various sizes of ground objects. The STransFuse model determines more precisely the borders of ground objects 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 multiscale remote sensing images with high accuracy.

The effect of low-illumination images on model performance is more pronounced in the Potsdam dataset, as shown in the red boxed area in Fig. 9. The model is unable to extract features adequately due to the image's overall low brightness. At the same time, the image features tree and low vegetation with identical spectral information, making model recognition much more challenging. We can see that all of the models have low ground object recognition accuracy in this image, and the misclassification problem is more pronounced. Although our STransFuse model achieves more accurate segmentation, there are still a small number of misclassifications. Therefore, our study still needs further research and exploration for remote sensing images with low illumination and low interclass variation.

G. 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 the STransFuse model (green circles) improved OA with a small

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

Method	Tree		Car		Building		Low Veg		Imp surf		mF1	mIoU	OA
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU			
Deeplabv3+	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	**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
PSPNet	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	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	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	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	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
SAN	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
DDCM-Net	83.83	72.16	89.03	80.23	93.89	88.48	82.87	70.75	89.71	81.35	81.97	71.42	86.69
HCANet	83.82	72.14	89.24	80.56	93.75	88.24	82.87	70.76	89.22	80.54	81.58	71.02	86.33
Capsules–Unet	24.36	13.87	24.84	14.18	48.92	32.38	57.99	40.83	57.4	40.25	35.7	23.65	49.23
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**

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. Because Transformer is based on self-attention for semantic computation, the number of model parameters improved based on Transformer is large, and our designed STransFuse model can balance the number of model parameters and experimental performance. The experimental efficiency of different models on Potsdam is shown in Fig. 10(b), and it can be observed from Fig. 10(b) that the STransFuse model reached a better OA in a shorter period of time.

VI. CONCLUSION

In this article, we propose a model of fusing Swin Transformer and CNN, STransFuse. This two-branch model can combine the advantages of Transformer network and CNN. Transformer is able to model the global semantic relevance of the input image. CNN with pretrained weights are capable of acquiring spatial contextual information of the images. And the designed model structure of phased fusion can make full use of coarse-grained and fine-grained feature information at different semantic scales, enabling the model to have excellent feature representation. In addition, we provide an attention fusion module that can adaptively fuse the output features from the transformer and CNN, resulting in a feature map input to the model's decoder that incorporates rich semantic and spatial contextual information. The prediction result of our STransFuse model in Vaihingen and Potsdam datasets gives a competitive result compared to other advanced models. Transformer still has great potential for application in computer vision, and we will continue to study the application of Transformer in remote sensing field in future.

ACKNOWLEDGMENT

The authors would like to thank the International Society for Photogrammetry and Remote Sensing for the dataset provided, and the anonymous reviewers for their voluntary and constructive comments that helped to improve this article.

REFERENCES

[1] X. Sun, P. Wang, C. Wang, Y. Liu, and K. Fu, "PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 173, pp. 50–65, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271620303555

[2] Q. He, X. Sun, Z. Yan, and K. Fu, "DABNet: Deformable contextual and boundary-weighted network for cloud detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, 2021, doi: 10.1109/TGRS.2020.3045474.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[4] Y. Guo, J. Liao, and G. Shen, "A deep learning model with capsules embedded for high-resolution image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 214–223, 2021.

[5] H. Bai, J. Cheng, X. Huang, S. Liu, and C. Deng, "HCANet: A hierarchical context aggregation network for semantic segmentation of high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, 2021, doi: 10.1109/LGRS.2021.3063799.

[6] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2015, pp. 234–241.

[10] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia, "Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 96–107, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271618300261

[11] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2547–2560, Jun. 2021.

[12] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 593–602.

[13] A. Vaswani *et al.*, "Attention is all you need," 2017, *arXiv:1706.03762*.

[14] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[16] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 323–339.

[17] J. He, L. Zhao, H. Yang, M. Zhang, and W. Li, "HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 165–178, Jan. 2020.

[18] H. Lin *et al.*, "Cat: Cross attention in vision transformer," 2021, *arXiv:2106.05786*.

[19] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

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

[21] W. Shi, M. Zhang, H. Ke, X. Fang, Z. Zhan, and S. Chen, "Landslide recognition by deep convolutional neural network and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 4654–4672, Jun. 2021.

[22] J. Lin, W. Jing, and H. Song, "SAN: Scale-aware network for semantic segmentation of high-resolution aerial images," 2019, *arXiv:1907.03089*.

[23] 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 Geosci. Remote Sens. Lett.*, to be published, doi: 10.1109/LGRS.2020.3021210.

[24] S. Xiang, Q. Xie, and M. Wang, "Semantic segmentation for remote sensing images based on adaptive feature selection network," *IEEE Geosci. Remote Sens. Lett.*, to be published, doi: 10.1109/LGRS.2021.3049125.

[25] A. Li, L. Jiao, H. Zhu, L. Li, and F. Liu, "Multitask semantic boundary awareness network for remote sensing image segmentation," *IEEE Trans. Geosci. Remote Sens.*, 2021, doi: 10.1109/TGRS.2021.3050885.

[26] R. Liu, L. Mi, and Z. Chen, "AFNet: Adaptive fusion network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7871–7886, Sep. 2021.

[27] 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 Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7854–7870, Sep. 2021.

[28] 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, doi: 10.3390/sym13010028.

[29] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12416–12425. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Mou_A_Relation-Augmented_Fully_Convolutional_Network_for_Semantic_Segmentation_in_Aerial_CVPR_2019_paper.html

[30] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 108–126.

[31] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," 2021, *arXiv:2102.10662*.

[32] H. Chen *et al.*, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12299–12310. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Pre-Trained_Image_Processing_Transformer_CVPR_2021_paper.html

[33] X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sens.*, vol. 13, no. 3, p. 498, 2021, doi: 10.3390/rs13030498.

[34] L. Yuan *et al.*, "Tokens-to-token VIT: Training vision transformers from scratch on imagenet," 2021, *arXiv:2101.11986*.

[35] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two pure transformers can make one strong GAN, and that can scale up," 2021, *arXiv:2102.07074*.

[36] W. Wang, C. Chen, M. Ding, J. Li, H. Yu, and S. Zha, "TransBTS: Multimodal brain tumor segmentation using transformer," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2021*, M. de Bruijne et al., Eds. Cham, Switzerland: Springer, 2021, pp. 109–119.

[37] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Zheng_Rethinking_Semantic_Segmentation_From_a_Sequence-to-Sequence_Perspective_With_Transformers_CVPR_2021_paper.html

[38] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," 2021, *arXiv:2103.00112*.

[39] 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 Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7557–7569, Nov. 2020.

[40] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," 2021, *arXiv:2105.15203*.

[41] ISPRS, Semantic Labeling Contest-Vaihingen (2018). Accessed: Sep. 4, 2021. [Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/benchmark/2d-sem-label-vaihingen/

[42] ISPRS, "Semantic Labeling Contest-Potsdam (2018). Accessed: Sep. 4, 2021. [Online]. Available: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[43] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021.

[44] 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 Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2021.3055950.

[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Zhou_Learning_Deep_Features_CVPR_2016_paper.html

[46] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated convolutions' merging network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6309–6320, Sep. 2020.

[47] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[48] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16519–16529. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Srinivas_Bottleneck_Transformers_for_Visual_Recognition_CVPR_2021_paper.html

[49] J. Chen *et al.*, "TransUnet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.

[50] S. Qian, Y. Shi, H. Wu, J. Liu, and W. Zhang, "An adaptive enhancement algorithm based on visual saliency for low illumination images," *Appl. Intell.*, May 2021, doi: 10.1007/s10489-021-02466-4.

Liang Gao received the bachelor's degree in computer science and technology from Zaozhuang University, Zaozhuang, China, in 2019. He is currently working toward the master's degree in software engineering with Xinjiang University, Urumqi, China.

His research interests include deep learning and remote sensing image semantic segmentation.

Hui Liu received the B.S. degree in software engineering from Xinjiang University, Urumqi, China, in 2014, and the master's degree in software engineering in 2017 from the College of Software, Xinjiang University, where she is currently working toward the Ph.D. degree in computer science and technology.

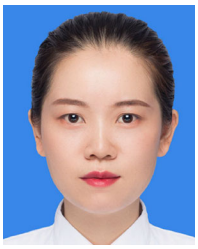
Her research interests include deep learning and opportunistic networks and the processing of remote sensing image data.

Minhang Yang received the bachelor's degree in software engineering from the Xi'an University of Technology, Xi'an, China, in 2019. She is currently working toward the master's degree in software engineering with Xinjiang University, Urumqi, China.

Her research interests include deep learning and multilabel image classification.

Long Chen received the bachelor's degree in geographic information science from the Shandong University of Science and Technology, Qingdao, China, in 2018. He is currently working toward the master's degree in software engineering with Xinjiang University, Urumqi, China.

His research interests include deep learning and single-image super-resolution.



Yaling Wan received the bachelor's degree in communication engineering, in 2014, from Xinjiang University, Urumqi, China, where she is currently working toward the master's degree in software engineering.

Her research interests include deep learning and hyperspectral image classification.

Zhengqing Xiao received the Ph.D degree in geographic information system from Beijing Normal University, Beijing, China, in 2011.

He is currently with the College of Mathematics and System Sciences, Xinjiang University, Urumqi, China. His research interests include Big Data analysis, image processing, and complex system modeling.

Yurong Qian received the bachelor's and master's degrees in computer science and technology from Xinjiang University, Urumqi, China, in 2000, and the doctorate degree in biology from Nanjing University, Nanjing, China, in 2010.

From 2012 to 2013, she was a Postdoctoral Fellow with the Department of Electronics and Computer Engineering, Hanyang University, South Korea, and is currently a Professor with the School of Software, Xinjiang University. Her research interests include computational intelligence such as Big Data processing, image processing, and artificial neural networks.

Dr. Qian 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, China.