测绘出版社 SURVEYING AND MAPPING PRESS

# CFM-UNet: A Joint CNN and Transformer Network via Cross Feature Modulation for Remote Sensing Images Segmentation

Min WANG[1], Peidong WANG[2]

1. *School of Urban Construction Engineering, Guangzhou City Polytechnic, Guangzhou* 510800, *China*; 2. *Guangdong Mechanical & Electrical Polytechnic, Guangzhou* 510515, *China*

**Abstract**: The semantic segmentation methods based on CNN have made great progress, but there are still some shortcomings in the application of remote sensing images segmentation, such as the small receptive field can not effectively capture global context. In order to solve this problem, this paper proposes a hybrid model based on ResNet50 and swin transformer to directly capture long-range dependence, which fuses features through Cross Feature Modulation Module (CFMM). Experimental results on two publicly available datasets, Vaihingen and Potsdam, are mIoU of 70.27% and 76.63%, respectively. Thus, CFM-UNet can maintain a high segmentation performance compared with other competitive networks.

**Key words**: remote sensing images; semantic segmentation; swin transformer; feature modulation module

Citation: Min WANG, Peidong WANG. CFM-UNet: A Joint CNN and Transformer Network via Cross Feature Modulation for Remote Sensing Images Segmentation[J]. Journal of Geodesy and Geoinformation Science, 2023, 6(4): 40-47. DOI:10.11947/j.JGGS.2023.0404.

## 1 Introduction

With the rapid development of satellite and Remote Sensing(RS) technology, the ability of high-resolution RS image acquisition has been greatly improved. And it becomes extremely important to extract the information of interest from the image accurately. Semantic segmentation has become an important method for RS images analysis to obtain application-worthy information, which can provide data support for precision agriculture, desertification detection, traffic supervision, urban planning and land resource management, etc[1].

In recent years, there have been breakthroughs in deep learning. Convolutional Neural Networks (CNNs), which effectively extract high-level abstract features through nonlinear structures, have been widely used in the field of image analysis and have made a great impact. The Fully Convolutional Networks (FCNs)' s proposal[2] has led to further breakthroughs in semantic segmentation of RS images. After FCNs, a large number of semantic segmentation networks with excellent performance have been proposed one after another, including U-Net[3], Deeplab V3+[4], DANet[5], UperNet[6]. U-Net fuses the high and low level semantic information to improve the classification effect of the semantic details of object boundaries and improves the segmentation performance of the network. In order to integrate spatial features, Deeplab V3 + made use of a decoder based on Deeplab V3, significantly raising network performance. DANet improves segmentation performance through both parallel channel attention and spatial attention. In order to obtain context information, PSPNet[7] and its upgraded version of UperNet adopt pyramid pooling module. However, these methods further capture global context information from local features obtained by CNNs, rather than directly capture global information. Thus, in RS images with complex backgrounds, it is not easy to ef-

Received date: 2023-09-18; accepted date: 2023-11-23
Foundation support: Young Innovative Talents Project of Guangdong Ordinary Universities (No. 2022KQNCX225); School-level Teaching and Research Project of Guangzhou City Polytechnic (No. 2022xky046)
First author: Min WANG
E-mail: wangmin@gcp.edu.cn
Corresponding author: Peidong WANG
E-mail: wpd_jxnc@163.com

fectively capture global scene information[8-9].

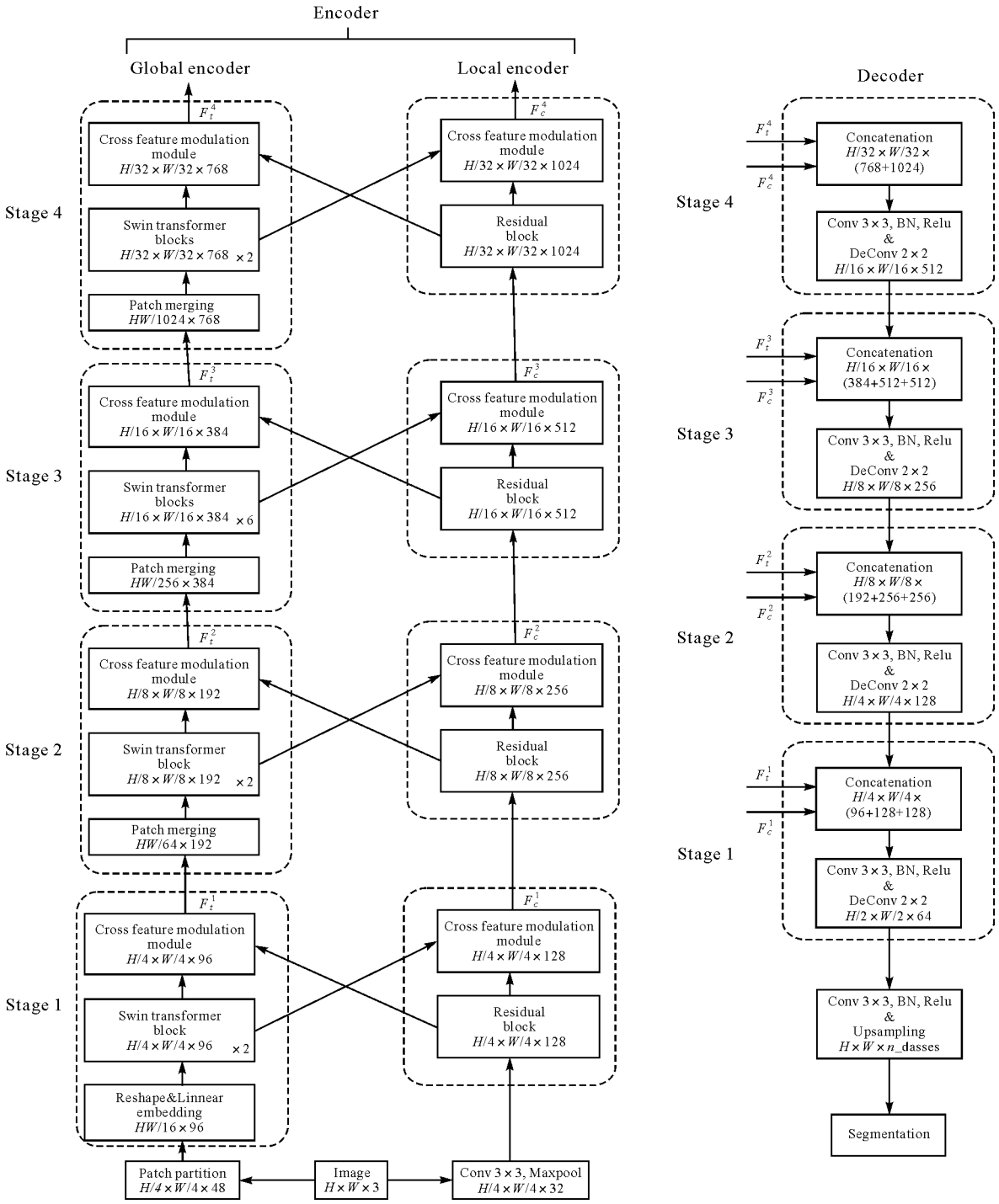



Fig.1 The whole architecture of CFM-UNet

Recently, with the development of deep learning, transformer has gained wide attention in the field of Computer Vision (CV) and opened a new era in the field of vision[10]. Benefit by the excellent global modeling ability of transformer, swin transformer[11] has been proposed, showing great potential in some semantic segmentation tasks. Up to now, the swin transformer-based methods have achieved great success in segmentation of medical images[12-13]. In RS images segmentation, ST-UNet[14] is proposed to improve the RS images segmentation performance by aggregating CNN features and transformer features. Although the design of CNNs and transformer combined structures has made significant progress, it is

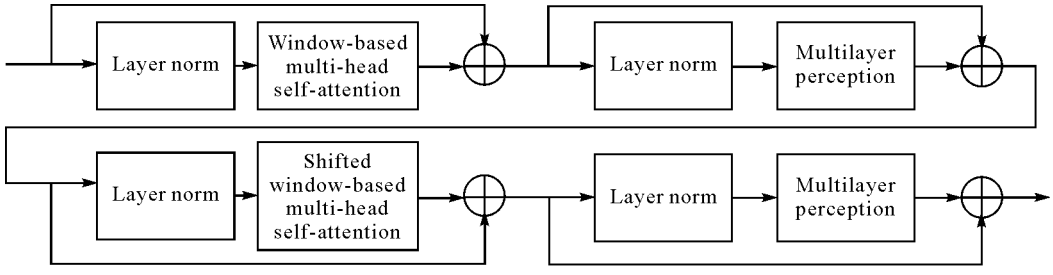
still an open issue. In this paper, we introduce cross feature modulation into the field of semantic segmentation of RS images for the first time, and propose a novel network structure (CFM-UNet) for RS images segmentation by combining the advantages of global receptive field of Swin Transformer and high-precision local features of CNNs, which improves the performance of RS images segmentation.

## 2 Methodology

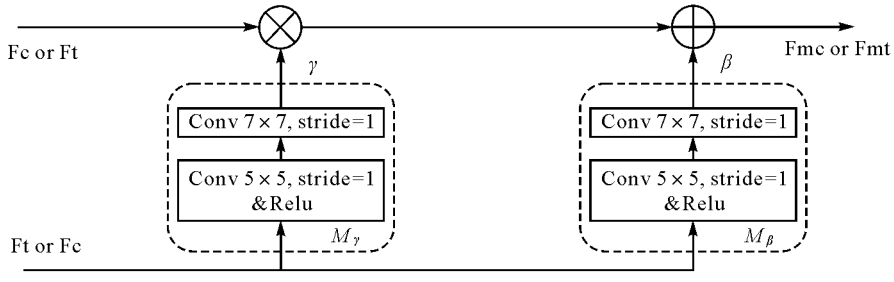
As shown in Fig. 1, the whole architecture of our CFM-UNet is constructed based on the structure of encoder and decoder. Especially, CFM-UNet creates a parallel encoder structure composed of a residual network based on CNN (called local encoder) and a Swin Transformer (called global encoder), which transmits information through the Cross Feature Modulation Module to fully capture discriminant features of RS images.

### 2.1 Network structure

The structure of the encoder can be divided into a global encoder and a local encoder. The following describes the global encoder, namely the Swin Transformer encoder. For an input RS image $X \in \mathbb{R}^{H\times W\times 3}$, where H denotes the length of the image, $W$ denotes the width of the image, and 3 denotes the image dimension. When the RS image passes through the patch partition layer, it is divided into $4\times 4$ non-overlapping patches. The size of each patch tensor is $H/4 \times W/4 \times 48$, and then the patch tensor is projected onto C dimensions by the linear embedding layer. Next, the Swin Transformer block searches from the local detail information of the image to the global contextual information. Finally, the one-stage feature map tensor is output. In the next stage, the patch merging layer merges each set of $2\times 2$ neighboring patches in the feature map, and the output dimension is set to twice the output dimension of the previous stage by reducing the dimensionality of the space in exchange for more channels. As in the first stage, the network then passes through the Swin Transformer block. Similarly, the remaining two stages of the global encoder are constructed. As shown in Fig.2(a), in order to enhance the across-windows information connection, W-MSA and SW-MSA are alternately implemented in successive Swin Transformer blocks. In the local encoder, namely the CNN-based residual network, the input RS image $X$ is first fed to ResNet50 so that the network can obtain the CNN features of four coding stages, and in this paper ResNet50 is compressed by half on the channel. Due to the constraints of the experimental conditions, we apply ResNet50 as the backbone of the CNN branch.




(a) The architecture of swin transformer blocks




(b) The architecture of CFMM

Fig.2 Architecture of some modules

The original Swin Transformer and residual network are described in the previous section, and in this paper each stage of the encoder is followed by a cross feature modulation module, which will be described below.

After the above four stages in two parallel encoders, we get feature $F_t^4 \in \mathbb{R}^{(H/32)\times(W/32)\times 768}$, $F_c^4 \in \mathbb{R}^{(H/32)\times(W/32)\times 1024}$, which is sent to the decoder after a concatenational layer and a convolutional layer. In each decoder stage, like UNet, firstly CFM-UNet concatenates the global encoder features ($F_t^1$, $F_t^2$, $F_t^3$) and the local encoder features ($F_c^1$, $F_c^2$, $F_c^3$) one by one for each stage. Then, CFM-UNet concatenates the encoder features and decoder features by skip connection. At last, through a 3 × 3 convolutional layer the number of channels of the feature is reduced.

## 2.2 Cross Feature Modulation Module(CFMM)

We found that the transformer's features have special signatures such as global attention but coarse textures differing from the CNN's features that have local attention but clear details. Inspired by the research direction of style transfer and conditional image enhancement[15-17], a cross feature modulation is attached after each encoder stage. Concretely, we use transformer features as conditional information to predict modulation parameters, which then modulates CNN features. On the contrary, we also use CNN features as conditional information to predict modulation parameters, which then modulates transformer features. The obtained feature maps are then sent to the next stage respectively. With the above cross modulation method, the CFMM exchanges the prior information extracted from the transformer branch and the CNN branch with each other, and forms a variety of series and parallel structures of transformer and CNN blocks at different depths, which enhances the information flow and feature selectivity in the network differently from some existing methods. In this way, we expect to transfer transformer's global attention to CNN features without destroying the details of CNN features and to obtain the inductive bias of CNNs to accelerate transformer convergence, which can be expressed as

$$F_{mt,mc}^s = M_{\gamma_{c,t}}^s(F_{c,t}^s) \otimes F_{t,c}^s \oplus M_{\beta_{c,t}}^s(F_{c,t}^s) \quad (1)$$

where $F_{mt}$, $F_{mc}$, respectively, represent the features of the global encoder and local encoder after modulation; $S$ denotes the stage level, which can take the values from 1 to 4. $\gamma$ and $\beta$ indicate the scaling and shifting matrices of affine transformation, which both have the same size with the corresponding dimension of CNN features $F_c$ or transformer features $F_t$; and $M_\gamma(\cdot)$ and $M_\beta(\cdot)$ are the modulation parameters generation blocks that contain two convolutional layers controlled on CNN features $F_c$ or transformer features $F_t$. Here, we use one 5×5 convolutional layer and one 7×7 convolutional layer instead of the original algorithm's two 3×3 convolutional layers. $\otimes$ and $\oplus$ denote the element-wise multiplication and element-wise addition, respectively. This cross feature modulation module is shown in Fig.2(b).

**Tab.1 Segmentation results of some methods on the Vaihingen dataset** (%)

| Methods | IoU | | | | | Evaluation indicator |
|---|---|---|---|---|---|---|
| | Low vegetation | Tree | Car | Impervious surface | Building | mIoU |
| FCN[2] | 54.80 | 70.38 | 39.92 | 73.22 | 78.97 | 63.46 |
| UNet[3] | 57.23 | 71.63 | 48.29 | 72.91 | 81.68 | 66.35 |
| Deeplab V3+[4] | 56.09 | 71.54 | 50.30 | 74.85 | 83.01 | 67.16 |
| UperNet[5] | 55.65 | 71.31 | 47.26 | 73.45 | 81.50 | 65.84 |
| DANet[6] | 56.88 | 71.21 | 42.68 | 73.54 | 81.40 | 65.14 |
| TransUNet[15] | 55.07 | 71.08 | 55.13 | 73.27 | 81.01 | 67.11 |
| Swin-UNet[12] | 49.48 | 67.12 | 30.78 | 69.31 | 73.37 | 58.01 |
| ST-UNet[14] | 57.79 | 72.53 | **61.48** | 76.36 | 82.98 | 70.23 |
| CFM-UNet(ours) | **58.08** | **72.80** | 60.71 | **76.46** | **83.33** | **70.27** |

**Tab.2 Segmentation results of some methods on the Potsdam dataset** (%)

| Methods | IoU | | | | | Evaluation indicator |
|---|---|---|---|---|---|---|
| | Low vegetation | Tree | Car | Impervious surface | Building | mIoU |
| FCN[2] | 66.10 | 63.19 | 74.34 | 77.41 | 83.52 | 72.91 |
| UNet[3] | 64.59 | 65.44 | 76.16 | 77.10 | 82.83 | 73.22 |
| Deeplab V3+[4] | 67.53 | 63.05 | 78.05 | 79.01 | 84.76 | 74.48 |
| UperNet[5] | 65.65 | 60.40 | 76.57 | 76.95 | 83.93 | 72.70 |
| DANet[6] | 66.46 | 63.47 | 75.28 | 77.35 | 83.45 | 73.20 |
| TransUNet[15] | 67.16 | 64.10 | 79.33 | 78.61 | 85.60 | 74.96 |
| Swin-UNet[12] | 59.03 | 50.96 | 71.15 | 71.45 | 75.02 | 65.52 |
| ST-UNet[14] | 67.89 | 66.37 | **79.77** | 79.19 | 86.63 | 75.97 |
| CFM-UNet(ours) | **69.49** | **68.32** | 78.89 | **79.58** | **86.86** | **76.63** |

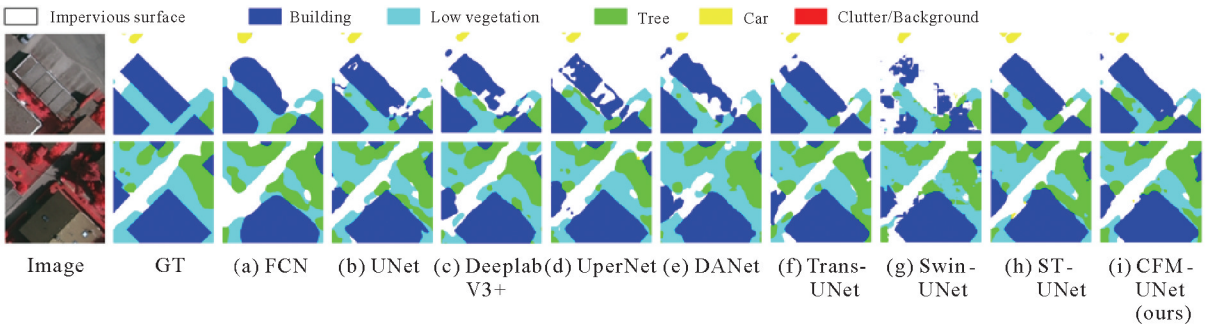



Fig.3 Examples of semantic segmentation results on the Vaihingen dataset

## 3 Experimental Results

### 3.1 Datasets

In this paper, we use two state-of-the-art airborne image datasets from the city classification and 3D building reconstruction test programs provided by ISPRS[18]. The datasets employ Digital Surface Models (DSM) generated from high-resolution orthorectified photographs and corresponding dense image matching techniques. Both dataset areas cover urban scenes: Vaihingen is a relatively small village with many individual buildings and small multi-story buildings. Following Literatures [19], [14], [20], [21] and [22], we choose 11 images numbered 1, 3, 5, 7, 13, 17, 21, 23, 26, 32, and 37 as the training set, 5 images numbered 11, 15, 28, 30, and 34 as the testing set, and crop them to 256 × 256, respectively. Potsdam is a typical historical city with large building blocks, narrow streets and dense settlement structures. With reference to the previous Literatures [8], [14] and [19], we utilize 14 color rgb images numbered 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, and 7_13 as the testing set, and the remaining 24 color rgb images as the training set. Similarly, we divide these images into 256 × 256. Each dataset has been manually classified into the six most common land cover categories. Following Literatures [14] and [23], we neglect the category of "Clutter/Background" when calculating evaluation indicator on the above two datasets.

### 3.2 Implementation details

(1) Training settings

Based on the Pytorch framework, our network is built. All experiments are implemented on a single GPU "NVIDIA Geforce RTX 3090 24-GB GPU". The batch size and the maximum epoch is set to 8 and 100, respectively. Following Literature [14], we use Stochastic Gradient Descent (SGD) optimizer with weight decay of 1e-4 and momentum term of 0.9 to train the network. Besides, we apply 0.01 as the initial learning rate and "Poly" as the decay strategy.

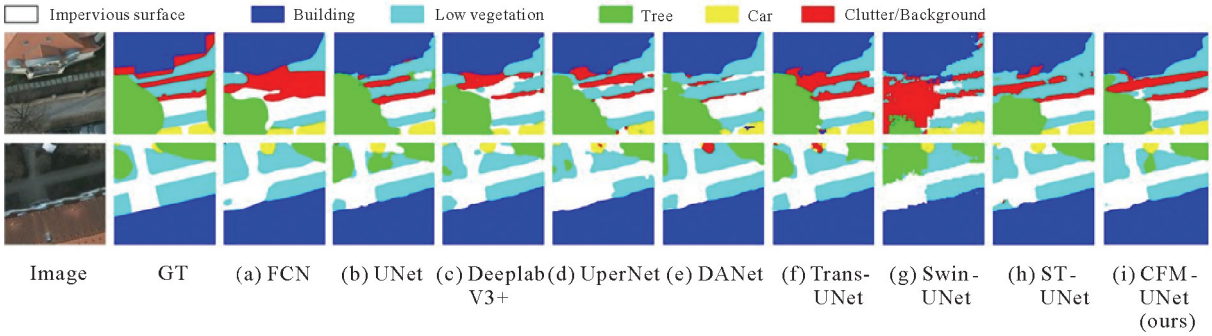



Fig.4 Examples of semantic segmentation results on the Potsdam dataset

(2) Loss function

Following Literatures [14], [24] and [25], in order to eliminate the impact of category imbalance, we use the joint loss function consisting of dice loss $L_{\text{Dice}}$ and the cross-entropy loss $L_{\text{CE}}$ to supervise the model. The joint loss function L is expressed as below

$$L = L_{\text{CE}} + L_{\text{Dice}} \tag{2}$$

(3) Evaluation criteria

Following the conventional semantic segmentation evaluation method, the experimental results are analyzed in this paper using the mean Intersection over Union (mIoU). mIoU is implemented as shown in equation

$$\text{mIoU} = \frac{1}{n+1}\sum_{i=0}^{n}\frac{\text{TP}}{\text{FN}+\text{FP}+\text{TP}}$$

where TP denotes the number of positive categories with correct prediction results; FP denotes the number of positive categories with incorrect prediction results, and FN denotes the number of negative categories with incorrect prediction results. The larger the mIoU value, the better the model segmentation performance.

### 3.3 Semantic segmentation results and analysis

(1) Performance comparison

To be fair, the CNN backbone networks of all models involved in the comparison were Resnet50 without pre-training. The training settings are the same for all models. In our proposed method, the dimensions $C$ and the number $N$ of Swin Transformer blocks in each stage can be obtained by following the standard Swin Transformer blocks: $C = \{96, 192, 384, 768\}$, $N = \{2, 2, 6, 2\}$. Our model outperforms the second best model by 0.04% on Vaihingen dataset and 0.66% on Potsdam dataset in mIoU, outperforming the majority of ResNet-based models. The detailed experimental results are presented in Tab.1 and Tab.2. And Fig.3 and Fig.4 shows the visualized prediction results of several semantic segmentation methods involved in Tab.1 and Tab.2. It can be observed that Swin-UNet lacks spatial location information, resulting in many semantic fragments in its segmentation results. Compared with other models, CFM-UNet reduces segmentation errors, especially for ground objects with high similarity.

(2) Efficiency analysis

For the comprehensive comparisons, Tab.3 lists the computational complexity, model parameters, speed and accuracy of all models in the same operating environment. Due to the parallel structure of CNN and transformer, our method has a larger number of parameters but higher accuracy.

(3) Ablation study

We performed ablation experiments by comparing the performance of removing the introduced cross feature modulation and the present method on Vaihingen dataset. Compared to the model without the cross feature modulation, the proposed model has a greater improvement on both datasets in mIoU. The validity of the cross feature modulation module is experimentally verified. The specific experimental results are shown in Tab.4.

In addition, in order to explore the effects of different components of the proposed method, we compare cross feature modulation with single-direction feature modulation and the feature fusion strategy in Literature [26] on Vaihingen dataset.

Defining the single-direction feature modulation of the CNN to the transformer as C>>T and the single-direction feature modulation of the transformer to the CNN as T>>C, the results are shown in Tab.4.

**Tab.3 Comparison of computational complexity, model parameters, speed and accuracy on Vaihingen dataset**

| Methods | FLOPs(G) | Parameters(MB) | Speed(FPS) | mIoU/(%) |
|---|---|---|---|---|
| FCN[2] | 6.2 | 22.70 | 370 | 63.46 |
| UNet[3] | 7.1 | 25.13 | 210 | 66.35 |
| Deeplab V3+[4] | 14.8 | 38.48 | 69 | 67.16 |
| UperNet[5] | 37.1 | 102.13 | 58 | 65.84 |
| DANet[6] | 13.1 | 45.36 | 107 | 65.14 |
| TransUNet[15] | 36.2 | 100.44 | 33 | 67.11 |
| Swin-UNet[12] | 6.5 | 25.89 | 52 | 58.01 |
| ST-UNet[14] | 52.3 | 160.97 | 6 | 70.23 |
| CFM-UNet(ours) | 66.1 | 209.71 | 5 | 70.27 |

**Tab.4 Ablation experiment of the proposed modules on the Vaihingen dataset** (%)

| Methods | Evaluation indicator |
|---|---|
| | mIoU |
| CFM-UNet without CFMM | 67.32 |
| CFM-UNet without C>>T | 67.43 |
| CFM-UNet without T>>C | 68.56 |
| CFM-UNet replaces CFMM with fusion strategy in Literature [26] | 69.14 |
| CFM-UNet | 70.27 |

From the results, the performance of CFM-UNet without T>>C is better than that of CFM-UNet without C>>T, so this is mainly because the transformer branch is working. Compared to CFM-UNet, the single-direction feature modulation network uses fewer cross connections and reduces the information flow within the network, resulting in lower accuracy. However, CFMM performs better than the fusion strategy with the similar cross structure in Literature [26], demonstrating its effectiveness.

## 4 Conclusion

In this paper, we propose a semantic segmentation model for RS images based on swin transformer, which applies the cross feature modulation module to combine the respective advantages of CNNs and transformer to improve the segmentation performance. Through experiments, the mIoU of CFM-UNet on two publicly available datasets, Vaihingen and Potsdam, are 70.27% and 76.63%, respectively, which can maintain a high segmentation performance compared with other competitive networks. Although the proposed network has achieved some progress in performance, it still cannot meet the real-time segmentation requirement, which will be the focus of the next stage of research.

## References

[1] ZUO Zongcheng, ZHANG Wen, ZHANG Dongying. A remote sensing image semantic segmentation method by combining deformable convolution with conditional random fields [J]. Journal of Geodesy and Geoinformation Science, 2020, 3(3): 39-49.

[2] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA: IEEE, 2015: 3431-3440.

[3] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]// Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich: Springer, 2015: 234-241.

[4] CHENL C, ZHU Yukun, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 833-851.

[5] FU Jun, LIU Jing, TIAN Haijie, et al. Dual attention network for scene segmentation[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA: IEEE, 2019: 3146-3154.

[6] XIAO Tete, LIU Yingcheng, ZHOU Bolei, et al. Unified perceptual parsing for scene understanding[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 432-448.

[7] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan, et al. Pyramid scene parsing network[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI: IEEE, 2017: 6230-6239.

[8] MOU Lichao, HUA Yuansheng, ZHU Xiaoxiang. Relation matters: relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(11): 7557-7569.

[9] L Wenjie, LI Yu, Z Quanhua. High-resolution remote sensing image segmentation using minimum spanning tree tessellation and RHMRF-FCM algorithm[J]. Journal of Geodesy and Geoinformation Science, 2020, 3(1): 52-63.

[10] DOSOVITSKIYA, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[C]//Proceedings of the 9th International Conference on Learning Representations. [S. l.]: OpenReview. net, 2021: 1-5.

[11] LIU Z, et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986.

[12] CAO Hu, WANG Yueyue, CHEN J, et al. Swin-Unet: unet-like pure transformer for medical image segmentation[C]// Proceedings of the European Conference on Computer Vision. Tel Aviv: Springer, 2023: 205-218.

[13] LIN Ailiang, CHEN Bingzhi, XU Jiayu, et al. DS-TransUNet: dual swin transformer U-Net for medical image segmentation [J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 4005615.

[14] HE Xin, ZHOU Yong, ZHAO Jiaqi, et al. Swin transformer embedding UNet for remote sensing image semantic segmentation[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 4408715.

[15] CHEN J, LU Y, YU Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[EB/OL]. [2023-09-01]. https://www.cs.jhu.edu/~alanlab/Pubs21/chen2021transunet.pdf.

[16] JIANG Liming, ZHANG Changxu, HUANG Mingyang, et al. TSIT: a simple and versatile framework for image-to-image translation [C] // Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 206-222.

[17] WANG Xintao, YU Ke, DONG Chao, et al. Recovering realistic texture in image super-resolution by deep spatial feature transform[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 606-615.

[18] ISPRS 2D semantic labeling dataset[EB/OL]. [2021-06-10]. https: // www2. isprs. org/commissions/comm2/wg4/benchmark/semantic-labeling/.

[19] MAGGIORI E, TARABALKA Y, CHARPIAT G,et al. High-resolution aerial image labeling with convolutional neural networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(12): 7092-7103.

[20] LIU Yu, NGUYEN D M, DELIGIANNIS N, et al. Hourglass-shape network based semantic segmentation for high resolution aerial imagery[J]. Remote Sensing, 2017, 9(6): 522.

[21] VOLPI M, TUIA D. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(2): 881-893.

[22] MARCOS D, VOLPI M, KELLENBERGER B,et al. Land cover mapping at very high resolution with rotation equivariant CNNs: towards small yet accurate models[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 145: 96-107.

[23] LI Xiangtai, HE Hao, LI Xia, et al. PointFlow: flowing semantics through points for aerial image segmentation[C]// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN: IEEE, 2021: 4217-4226.

[24] FIDONL, LI Wenqi, GARCIA-PERAZA-HERRERA L C, et al. Generalised Wasserstein dice score for imbalanced multi-class segmentation using holistic convolutional networks[C]//Proceedings of the 3rd International MICCAI Brainlesion Workshop. Quebec City: Springer, 2017: 64-76.

[25] ZHU Qingtian, ZHENG Yumin, JIANG Yulai, et al. Efficient multi-class semantic segmentation of high resolution aerial imagery with dilated LinkNet[C]//Proceedings of 2019 IEEE International Geoscience and Remote Sensing Symposium. Yokohama: IEEE, 2019: 1065-1068.

[26] PENG Zhiliang, HUANG Wei, GU Shanzhi, et al. Conformer: local features coupling global representations for visual recognition[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 357-366. DOI: 10.1109/ICCV48922.2021.00042.