A Novel Transformer based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images

Journal:	*Geoscience and Remote Sensing Letters*
Manuscript ID	GRSL-00776-2021
Manuscript Type:	Letters
Sub-topic:	Image Processing, Analysis and Classification
Date Submitted by the Author:	12-May-2021
Complete List of Authors:	Wang, Libo; Wuhan University, School of Remote Sensing and Information Engineering Li, Rui; Wuhan University, School of Remote Sensing and Information Engineering Duan, Chenxi; Wuhan University, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing Zhang, Ce; Lancaster University, Lancaster Environment Centre; Centre for Ecology and Hydrology, Centre of Excellence in Environmental Data Science Meng, Xiaoliang; Wuhan University, School of Remote Sensing and Information Engineering Fang, Sheng; Wuhan University, School of Remote Sensing and Information Engineering
Key Words:	Vegetation and Land Surface < Methodologies and Applications to

A Novel Transformer based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images

Libo Wang, Rui Li, Chenxi Duan, Ce Zhang, Xiaoliang Meng and Shenghui Fang

***Abstract*—The Fully Convolutional Network (FCN) with an encoder-decoder architecture has been the standard paradigm for semantic segmentation. The encoder-decoder architecture utilizes an encoder to capture multi-level feature maps, which are incorporated into the final prediction by a decoder. As the context is crucial for precise segmentation, tremendous effort has been made to extract such information in an intelligent fashion, including employing dilated/atrous convolutions or inserting attention modules. However, these endeavours are all based on the FCN architecture with ResNet or other backbones, which cannot fully exploit the context from the theoretical concept. By contrast, we introduce the Swin Transformer as the backbone to extract the context information and design a novel decoder of densely connected feature aggregation module (DCFAM) to restore the resolution and produce the segmentation map. The experimental results on two remotely sensed semantic segmentation datasets demonstrate the effectiveness of the proposed scheme.**

***Index Terms*—semantic segmentation, fine-resolution remote sensing images, transformer**

I. Introduction

As an effective method to extract features automatically and hierarchically from images, the convolutional neural network (CNN) has become the common framework for computer vision (CV) related tasks. For semantic segmentation, the Fully Convolutional Network (FCN) [3] is the first proven and effective end-to-end CNN structure. Specifically, there are two symmetric paths in the FCN and its variants: a contracting path, i.e., the encoder, for extracting features, and an expanding path, i.e., the decoder, for exacting positions [6]. The contracting path, by definition, gradually downsamples the resolution of feature maps to reduce the computational consumption, while the expanding path can learn more semantic meaning via a progressively increasing receptive field. Benefit from its translation equivariance and locality, the FCN enhances the segmentation performance significantly and influences the entire field. Specifically, the translation equivariance underpins the generalization capability of the model to unseen data, while the locality reduces the complexity of the model by sharing parameters.

The outcome of FCN, although encouraging, appears to be coarse due to the over-simplified design of the decoder. Subsequently, more elaborate encoder-decoder structures were proposed [9-11], thus increasing the accuracy further. However, the long-range dependency is limited by the locality property of FCN-based methods, which is critical for segmentation in unconstrained scene images. There are two types of methods to address the issue, either modifying the convolution operation or utilizing the attention mechanism. The former aiming to enlarge the receptive fields using large kernel sizes [13], dilated convolutions [16], or feature pyramids [2], whereas the latter [4, 17, 18] focuses on integrating attention mechanisms [20] with the FCN architecture to capture long-range dependencies of the feature maps. Nevertheless, both methods fail to liberate the network from the dependence of the FCN structure. More recently, several inspiring advances [21-23] attempt to avoid convolution operations completely by employing attention-alone models, thereby achieving feature maps with long-range dependencies effectively.

For natural language processing (NLP), the dominant architecture is the Transformer [20], which adopts the multi-head attention to model long-range dependencies for sequence modeling and transduction tasks. The tremendous breakthrough in the natural language domain inspires researchers to explore the potential and feasibility of Transformer in the computer vision field. Obviously, the successful application of Transformer will become the first and foremost step to integrate computer vision and NLP, thereby providing a universal and uniform artificial intelligence (AI) scheme.

The pioneering work of Swin Transformer [23] presents a hierarchical feature representation scheme that demonstrates impressive performances with linear computational complexity. In this Letter, we *first* introduce the Swin Transformer for semantic segmentation of fine-resolution remote sensing images. Most importantly, we propose a densely connected

This work was funded by National Natural Science Foundation of China (NSFC) under grant number 41971352. *(Libo Wang and Rui Li contributed equally to this work). (Corresponding author: Shenghui Fang.)*

L. Wang, R. Li, X. Meng and S. Fang are with School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: wanglibo@whu.edu.cn; lironui@whu.edu.cn; xmeng@whu.edu.cn; shfang@whu.edu.cn).

C. Duan is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China; chenxiduan@whu.edu.cn (e-mail: chenxiduan@whu.edu.cn).

C. Zhang is with Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, United Kingdom; UK Centre for Ecology & Hydrology, Library Avenue, Lancaster, LA1 4AP, United Kingdom (e-mail: c.zhang9@lancaster.ac.uk).

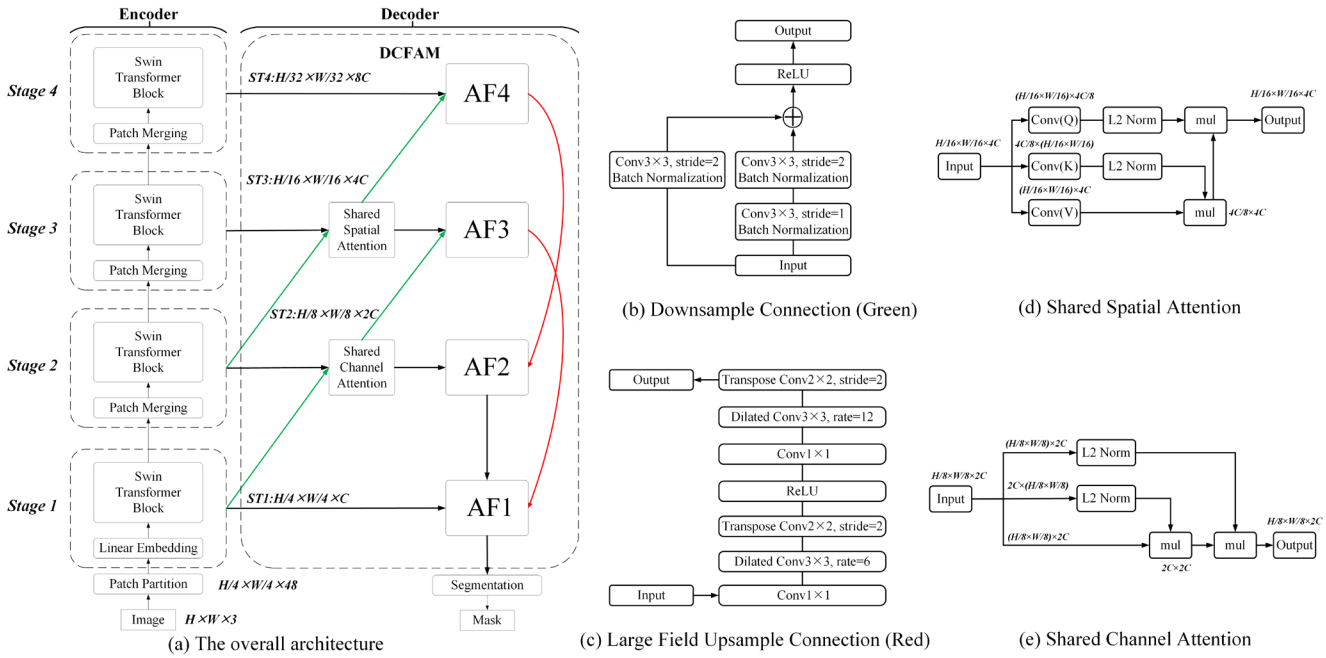

Fig. 1 (a) The overall architecture of DCST, (b) Downsample Connection, (c) Large Field Upsample Connection, (d) Shared Spatial Attention, and (e) Shared Channel Attention.

feature aggregation module (DCFAM) to extract multi-scale relation-enhanced semantic features for precise segmentation. Combining Swin Transformer and DCFAM, a novel semantic segmentation scheme of Densely Connected Swin Transformer (DCST) is established.

II. Methodology

The overall architecture of our DCST is constructed based on the encoder-decoder structure, where the Swin Transformer is introduced as the encoder while the proposed DCFAM is selected as the decoder.

A. Swin Transformer

As shown in Fig.1 (a), the Swin Transformer backbone [23] first utilizes a patch partition module to split the input RGB image into non-overlapping patches as "tokens". The feature of each patch is set as a concatenation of the raw pixel RGB values. Subsequently, this raw-valued feature is fed into the multistage feature transformation. In stage 1, a linear embedding layer is deployed to project features to an arbitrary dimension C. Thereafter, the specially designed Swin Transformer blocks, which can maintain the number of tokens (i.e., $H/4 \times W/4$), are adopted to extract semantic features. In the remaining stages, the number of tokens is gradually reduced by patch merging layers along with the increasing depth of the network to produce a hierarchical representation. Proceed by the four stages, four hierarchical Swin Transformer features ($\boldsymbol{ST_1, ST_2, ST_3, and\ ST_4}$) with different sizes are created.

By choosing diverse hyper-parameters, i.e., the dimensions C and the number of Swin Transformer blocks in each stage, four Swin Transformer backbones with different complexities can be obtained:

- Swin-T: $C = 96$, block numbers = {2, 2, 6, 2}
- Swin-S: $C = 96$, block numbers = {2, 2, 18, 2}
- Swin-B: $C = 128$, block numbers = {2, 2, 18, 2}
- Swin-L: $C = 192$, block numbers = {2, 2, 18, 2}

In this letter, we choose Swin-S pre-trained on the ImageNet as the backbone of the encoder, with the number of parameters (50M) comparable to ResNet-101 (45M).

B. Densely Connected Feature Aggregation Module

Multi-scale and confusing geospatial objects appear frequently in fine-resolution remote sensing images, which seriously affects the quality of segmentation. To handle this issue, we propose a novel DCFAM method for feature representation. To be specific, we design a Shared Spatial Attention (SSA) and a Shared Channel Attention (SCA) to enhance the spatial-wise and channel-wise relationship of the semantic features based on our previous work of linear attention mechanism [17]. Besides, multi-level features are further integrated using the Downsample Connection and the Large-field Upsample Connection for improving multi-scale representation. As shown in Fig.1, the DCFAM connects the four hierarchical transformer features with cross-scale connections (i.e., Downsample Connection and Large Field Upsample Connection) and attention blocks (i.e., Shared Spatial Attention and Shared Channel Attention), generating four aggregation features (i.e., $\boldsymbol{AF_1, AF_2, AF_3, and\ AF_4}$). Considering the efficiency of the model, we only leverage $\boldsymbol{AF_1}$ for final segmentation.

Downsample Connection: The Downsample connection aims to connect the low-level and high-level transformer features for fusion, which can be defined as follow:

$$D_i^j(\boldsymbol{X}) = f_\sigma(f_\delta(\boldsymbol{X}) + f_\mu(f_\theta(\boldsymbol{X}))) \quad (1)$$

where $\boldsymbol{X}$ is the input vector, f_σ is a ReLU activation function, f_δ and f_μ are a 3×3 convolution layer with a stride of 2, and f_θ is a standard 1×1 convolution layer. i and j denote the number of the input channel and output channel, respectively.

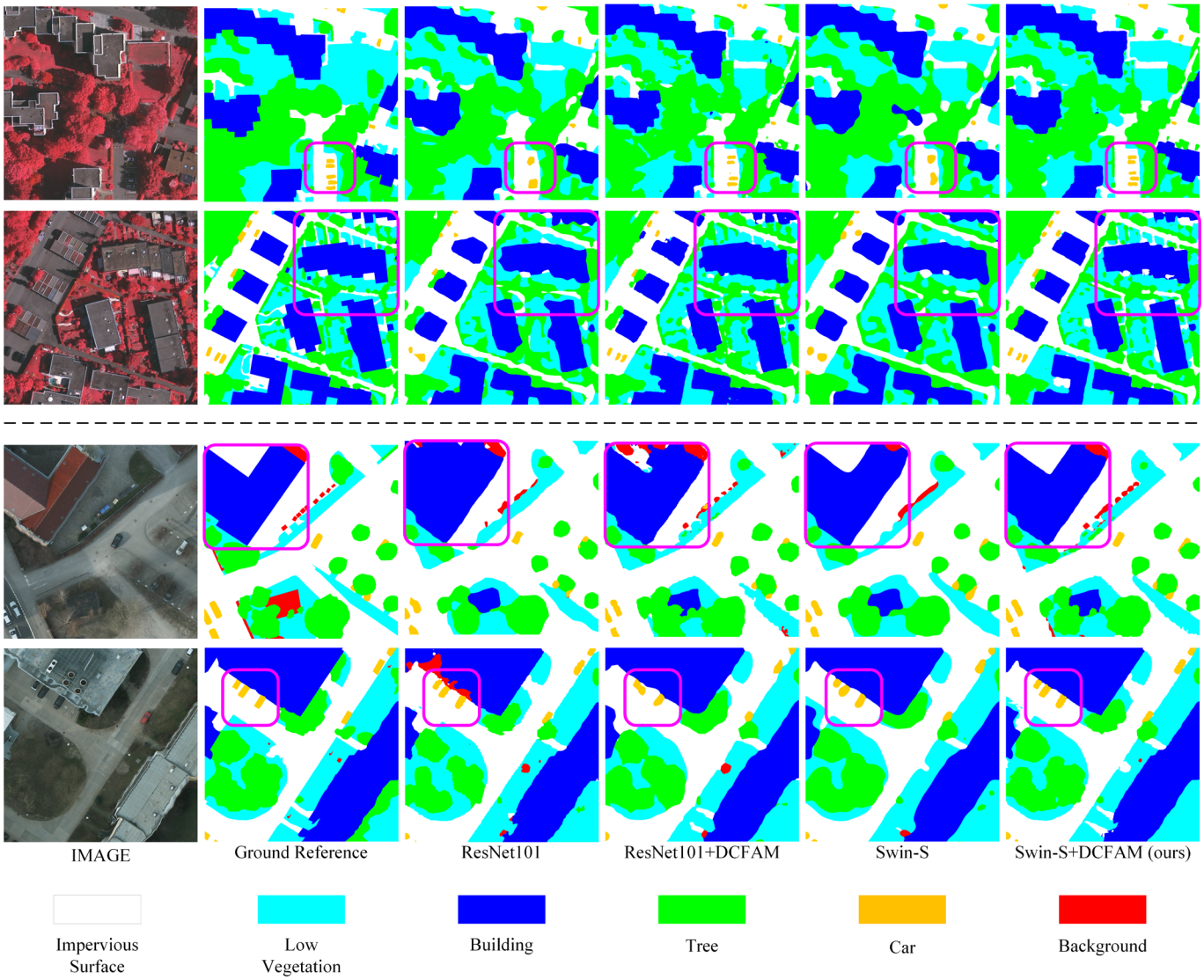

Fig. 2 Enlarged visualization of results on the Vaihingen dataset (Top) and Potsdam dataset (Bottom).

Large field Upsample Connection: To capture multi-scale context effectively, we embedded the dilated convolution into the Large filed Upsample Connection formulated as:

$$LU_m^n(\boldsymbol{X}) = f_\varphi^{12}(f_\sigma(f_\varphi^6(\boldsymbol{X}))) \tag{2}$$

where f_φ^{12} is a composite function that contains a standard 1×1 convolution, a dilated convolution with a dilated rate of 12, and a standard transpose convolution. Similarly, f_φ^6 has a dilated rate of 6. m and n represent the number of the input channel and output channel, respectively.

Shared Spatial Attention: Based on the linear attention mechanism [17], we utilize the Shared Spatial Attention to model the long-range dependencies in the spatial dimension defined as:

$$SSA(\boldsymbol{X}) = \frac{\sum_n V(\boldsymbol{X})_{c,n} + \left(\frac{Q(\boldsymbol{X})}{\|Q(\boldsymbol{X})\|_2}\right)\left(\left(\frac{K(\boldsymbol{X})}{\|K(\boldsymbol{X})\|_2}\right)^T V(\boldsymbol{X})\right)}{N + \left(\frac{Q(\boldsymbol{X})}{\|Q(\boldsymbol{X})\|_2}\right)\sum_n\left(\frac{K(\boldsymbol{X})}{\|K(\boldsymbol{X})\|_2}\right)^T_{c,n}} \tag{3}$$

where $\mathrm{Q}(\boldsymbol{X})$, $\mathrm{K}(\boldsymbol{X})$, and $\mathrm{V}(\boldsymbol{X})$ represent the convolutional operation to generate the *query* matrix $\boldsymbol{Q} \in \mathbb{R}^{N\times D_k}$, *key* matrix $\boldsymbol{K} \in \mathbb{R}^{N\times D_k}$, and *value* matrix $\boldsymbol{V} \in \mathbb{R}^{N\times D_v}$. N is the number of pixels in the input feature maps. c and n indicate the channel dimension and the flattened spatial dimension.

Shared Channel Attention: Similarly, the Shared Channel Attention is designed to extract the long-range dependencies among the channel dimension:

$$SCA(\boldsymbol{X}) = \frac{\sum_c R(\boldsymbol{X})_{c,n} + \left(R(\boldsymbol{X})_{c,n}\left(\frac{R(\boldsymbol{X})}{\|R(\boldsymbol{X})\|_2}\right)^T\right)\frac{R(\boldsymbol{X})}{\|R(\boldsymbol{X})\|_2}}{N + \left(\frac{R(\boldsymbol{X})}{\|R(\boldsymbol{X})\|_2}\right)^T\sum_c\left(\frac{R(\boldsymbol{X})}{\|R(\boldsymbol{X})\|_2}\right)^T_{c,n}} \tag{4}$$

where $R(\boldsymbol{X})$ indicate the reshape operation to flatten the spatial dimension. The detailed information about our previous work on the linear attention mechanism can be referred to [17].

Feature aggregation: The four aggregation features (***AF₁, AF₂, AF₃, and AF₄***) can eventually be computed by the following equations:

TABLE I
THE EXPERIMENTAL RESULTS ON THE VAIHINGEN DATASET.

Method	Backbone	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)
DeepLabV3+ [1]	ResNet101	92.38	95.17	84.29	89.52	86.47	89.57	90.56	81.47
PSPNet [2]	ResNet101	92.79	95.46	84.51	89.94	**88.61**	90.26	90.85	82.58
DANet [4]	ResNet101	91.63	95.02	83.25	88.87	87.16	89.19	90.44	81.32
EaNet [7]	ResNet101	93.40	**96.20**	*85.60*	**90.50**	*88.30*	**90.80**	*91.20*	-
DDCM-Net [5]	ResNet50	92.70	95.30	83.30	89.40	*88.30*	89.80	90.40	-
CASIA2 [14]	ResNet101	93.20	96.00	84.70	89.90	86.70	90.10	91.10	-
V-FuseNet [12]	FuseNet	91.00	94.40	84.50	89.90	86.30	89.20	90.00	-
DLR_9 [19]	-	92.40	95.20	83.90	89.90	81.20	88.50	90.30	-
Ours	Swin-S	**93.60**	*96.18*	**85.75**	*90.36*	87.64	*90.71*	**91.63**	**83.22**

TABLE II
THE EXPERIMENTAL RESULTS ON THE POTSDAM DATASET.

Method	Backbone	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA (%)	mIoU (%)
DeepLabV3+ [1]	ResNet101	92.95	95.88	87.62	88.15	96.02	92.12	90.88	84.32
PSPNet [2]	ResNet101	93.36	96.97	87.75	88.50	95.42	92.40	91.08	84.88
DDCM-Net [5]	ResNet50	92.90	96.90	87.70	*89.40*	94.90	92.30	90.80	-
CCNet [8]	ResNet101	93.58	96.77	86.87	88.59	*96.24*	92.41	91.47	85.65
AMA_1	-	93.40	96.80	87.70	88.80	96.00	*92.54*	91.20	-
SWJ_2	ResNet101	**94.40**	*97.40*	*87.80*	87.60	94.70	92.38	*91.70*	-
V-FuseNet [12]	FuseNet	92.70	96.30	87.30	88.50	95.40	92.04	90.60	-
DST_5 [15]	FCN	92.50	96.40	86.70	88.00	94.70	91.66	90.30	-
Ours	Swin-S	*94.19*	**97.57**	**88.57**	**89.62**	**96.31**	**93.25**	**92.00**	**87.56**

$$\boldsymbol{AF_4} = \boldsymbol{ST_4} + D_{384}^{768}(SSA(D_{192}^{384}(\boldsymbol{ST_2})) \quad (5)$$

$$\boldsymbol{AF_3} = SSA(\boldsymbol{ST_3}) + D_{192}^{384}(SCA(D_{96}^{192}(\boldsymbol{ST_1})) \quad (6)$$

$$\boldsymbol{AF_2} = SCA(\boldsymbol{ST_2}) + LU_{768}^{192}(\boldsymbol{AF_4}) \quad (7)$$

$$\boldsymbol{AF_1} = \boldsymbol{ST_1} + U(\boldsymbol{AF_2}) + LU_{384}^{96}(\boldsymbol{AF_3}) \quad (8)$$

Here, U is a bilinear interpolation upsample operation with a scale factor of 2. Capitalising on the benefits provided by cross-scale connections and attention blocks, the final segmentation feature $\boldsymbol{AF_1}$ is abundant in multi-scale and contextual information.

III. EXPERIMENTAL RESULTS

A. Dataset

We test the effectiveness of the proposed scheme on the well-known ISPRS Vaihingen and Potsdam semantic labelling datasets. There are 33 tiles extracted from true orthophoto and the co-registered normalized DSMs in the Vaihingen dataset with an average size of 2494 × 2064 pixels. The Potsdam dataset contains 38 tiles and the size of each tile is 6000 × 6000. Following previous pieces of literature [5, 12, 14], in the Vaihingen dataset, we use the benchmark organizer defined 16 images for training and 17 for testing, while the setting in the Potsdam dataset is 24 tiles for training and 14 tiles for testing. We do not employ DSMs in our experiments to reduce computation.

B. Experimental Setting

All of the experiments are implemented with PyTorch on a single RTX 3090, and the optimizer is set as AdamW with a 0.0003 learning rate. The soft cross-entropy is used as the loss function. For each method, the overall accuracy (OA), mean Intersection over Union (mIoU), and F1-score (F1) are chosen as evaluation indices:

$$OA = \frac{\sum_{k=1}^{N} TP_k}{\sum_{k=1}^{N} TP_k + FP_k + TN_k + FN_k}, \quad (9)$$

$$mIoU = \frac{1}{N}\sum_{k=1}^{N} \frac{TP_k}{TP_k + FP_k + FN_k}, \quad (10)$$

$$precision = \frac{1}{N}\sum_{k=1}^{N} \frac{TP_k}{TP_k + FP_k}, \quad (11)$$

$$recall = \frac{1}{N}\sum_{k=1}^{N} \frac{TP_k}{TP_k + FN_k}, \quad (12)$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}, \quad (13)$$

where TP_k, FP_k, TN_k, and FN_k indicate the true positive, false positive, true negative, and false negatives, respectively, for the specific object indexed as class k. OA is computed for all categories including the background.

C. Semantic Segmentation Results and Analysis

1) Performance Comparison: The experimental results on the Vaihingen and Potsdam datasets among state-of-the-art methods are listed in Table I and Table II. The quantitive indices demonstrate the effectiveness of the proposed segmentation scheme (DCST) constructed by the Swin-S and DCFAM. To be specific, our proposed DCST method achieves 90.71% in mean F1-score, 91.63% in OA, and 83.22% in mIoU for the Vaihingen dataset, with 93.25%, 92.00%, and 87.56% for the Potsdam dataset, outperforming the majority of ResNet-based methods with highly competitive accuracy. Benefit from the global context information modeled by the Swin-Transformer and the DCFAM, the performance of our scheme not only superiors to recent contextual information aggregation methods designed initially for natural images such as DeepLabV3+ and PSPNet, but also prevail over the lastest multi-scale feature aggregation models proposed for remote sensing images, such as EaNet and DDCM-Net.

2) Ablation Study: As we not only propose a novel feature aggregation model but also introduce a brand-new backbone for segmentation, it is valuable to conduct the ablation study and

investigate the contribution of each part upon accuracy. For the ablation study, we select the ResNet-101 and the Swin-S with the direct upsample operation as the baseline. As shown in Table III, the substitution of the backbone from the ResNet-101 to the Swin-S yields a 3% increase in the Vaihingen dataset and a 4.05% increase in the Potsdam dataset for the mIoU index, showing the superiority of the Swin-S. Meanwhile, as the DCFAM can refine the feature maps effectively by capturing the long-range dependencies and multi-scale information, the utilization of DCFAM enhances the performance dramatically compared with baseline methods. For example, the increase in mIoU for the ResNet baseline in the two datasets is 6.95% and 5.90%, respectively. The integration of Swin-S and DCFAM achieves the highest accuracy (Table III), whose performance can also be observed in Fig. 2.

TABLE III
ABLATION STUDY ON THE VAIHINGEN AND POTSDAM DATASETS.

Dataset	Method	Mean F1	OA (%)	mIoU (%)
Vaihingen	ResNet101	85.31	89.59	75.48
	ResNet101+DCFAM	90.22	91.04	82.43
	Swin-S	87.54	90.50	78.48
	Swin-S+DCFAM	**90.71**	**91.63**	**83.22**
Potsdam	ResNet101	88.66	89.24	79.97
	ResNet101+DCFAM	92.28	90.81	85.87
	Swin-S	91.20	90.54	84.02
	Swin-S+DCFAM	**93.25**	**92.00**	**87.56**

IV. CONCLUSION

In this Letter, for the first time, we introduce Transfomer into semantic segmentation of fine-resolution remote sensing images and we develop a densely connected feature aggregation module to capture multi-scale relation-enhanced semantic features, thereby increasing the segmentation accuracy. Numerical experiments conducted on the ISPRS Vaihingen and Potsdam datasets demonstrate the effectiveness of our scheme in segmentation accuracy. We envisage this pioneering Letter could inspire researchers and practitioners in this field to explore the potential and feasibility of the Transformer more widely in the remote sensing and Earth observation domain.

REFERENCE

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

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

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

[4] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146-3154.

[5] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Dense dilated convolutions' merging network for land cover classification," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 58, no. 9, pp. 6309-6320, 2020.

[6] L. Wang *et al.*, "SaNet: Scale-aware Neural Network for Semantic Labelling of Multiple Spatial Resolution Aerial Images," 2021.

[7] X. Zheng, L. Huan, G.-S. Xia, and J. Gong, "Parsing very high resolution urban scene images by learning deep ConvNets with edge-aware loss," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 170, pp. 15-28, 2020.

[8] Z. Huang *et al.*, "CCNet: Criss-Cross Attention for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2020.

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

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence,* vol. 39, no. 12, pp. 2481-2495, 2017.

[11] R. Li, C. Duan, S. Zheng, C. Zhang, and P. M. Atkinson, "MACU-Net for Semantic Segmentation of Fine-Resolution Remotely Sensed Images," *IEEE Geoscience and Remote Sensing Letters,* pp. 1-5, 2021, doi: 10.1109/lgrs.2021.3052886.

[12] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 140, pp. 20-32, 2018.

[13] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters--improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4353-4361.

[14] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS journal of photogrammetry and remote sensing,* vol. 145, pp. 78-95, 2018.

[15] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," *arXiv preprint arXiv:1606.02585,* 2016.

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

[17] R. Li, S. Zheng, C. Duan, J. Su, and Z. Ce, "Multistage Attention ResU-Net for Semantic Segmentation of Fine-Resolution Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters,* vol. 10.1109/LGRS.2021.3063381, 2021.

[18] R. Li, S. Zheng, C. Duan, and J. Su, "Multi-Attention-Network for Semantic Segmentation of High-Resolution Remote Sensing Images," *arXiv preprint arXiv:2009.02130,* 2020.

[19] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 135, pp. 158-172, 2018.

[20] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998-6008.

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

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

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