Class-guided Swin Transformer for Semantic Segmentation of Remote Sensing Imagery

Journal:	*Geoscience and Remote Sensing Letters*
Manuscript ID	GRSL-01132-2022
Manuscript Type:	Letters
Sub-topic:	Image Processing, Analysis and Classification
Date Submitted by the Author:	24-Jun-2022
Complete List of Authors:	Meng, Xiaoliang; Wuhan University, School of Remote Sensing and Information Engenering Yang, Yuechi; Wuhan University, School of Remote Sensing and Information Engineering Wang, Libo; Wuhan University, School of Remote Sensing and Information Engineering Wang, Teng; Ministry of Natural Resources South China Sea Bureau, Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical Area of South China Li, Rui; University of Warwick, School of Engineering Zhang, Ce; Lancaster University, Lancaster Environment Centre; Centre for Ecology and Hydrology, Centre of Excellence in Environmental Data Science
Key Words:	Optical Data < Processing, Sensors and Systems for

Class-guided Swin Transformer for Semantic Segmentation of Remote Sensing Imagery

Xiaoliang Meng, Yuechi Yang, Libo Wang*, Teng Wang, Rui Li and Ce Zhang

***Abstract*—Semantic segmentation of remote sensing images plays a crucial role in a wide variety of practical applications, including land cover mapping, environmental protection, and economic assessment. In the last decade, convolutional neural network (CNN) is the mainstream deep learning based method of semantic segmentation. Compared with conventional methods, CNN-based methods learn semantic features automatically, thereby achieving strong representation capability. However, the local receptive field of the convolution operation limits CNN-based methods from capturing global information. In contrast, Vision Transformer demonstrates its great potential in global information modelling and obtains superior results in semantic segmentation. Inspired by this, in this Letter, we propose a class-guided Swin Transformer (CG-Swin) for semantic segmentation of remote sensing images. Specifically, we adopt a Transformer-based encoder-decoder structure, which introduces the Swin Transformer backbone as the encoder and designs a class-guided Transformer block to construct the decoder. The experimental results on ISPRS Vaihingen and Potsdam datasets demonstrate the significant breakthrough of the proposed method over ten benchmarks, outperform both advanced CNN-based and recent Vision Transformers based approaches.**

***Index Terms*—Fully Transformer network, class-guided mechanism, semantic segmentation, remote sensing.**

I. Introduction

Semantic segmentation is one of the fundamental tasks in remote sensing imagery interpretation, which plays a vital role in widespread applications [4], such as land cover classification [7], urban planning [9], and ecological assessment [12]. Over the last decade, the convolutional neural network (CNN) has become the most popular framework for semantic segmentation due to its automatic and strong feature learning. The Fully Convolutional Network (FCN) [1] first proposes an end-to-end CNN structure for semantic segmentation. The outcome of FCN, although encouraging, appears to be coarse due to the over-simplified decoding process. To address such issue, the U-Net [16] presents an encoder-decoder framework with two symmetric paths, i.e. a contracting path for encoding semantic features and an expanding path for decoding and restoring the resolution. Benefiting from this encoder-decoder design, U-Net and its variants [18] obtain impressive segmentation performance. Subsequently, many CNN-based segmentation methods adopt the encoder-decoder structure and make great progress in the remote sensing community.

However, the convolutional kernel with a fixed size leads to a local receptive field, which makes CNNs incapable to model the long-range dependency or global context inherently [19]. As for semantic segmentation, the local pixel classification can be more accurate if with the support of global contextual information [20]. To circumvent this issue, some studies modified the convolutional operation [2, 3] or applied attention mechanisms [21]. The former aims to enlarge the receptive fields by using large kernel sizes, dilated convolutions, or feature pyramids, whereas the latter introduces spatial or point-wise attention modules into CNNs to better capture contextual information. Nevertheless, these methods fail to liberate the network from the dependence of a convolutional encoder, thus, biased toward local interactions.

More recently, several inspiring advances attempt to avoid convolution operations completely by a novel architecture, i.e. Vision Transformer (ViT) [22]. ViTs formulate the semantic segmentation task as a sequence-to-sequence problem and extract semantic features with long-range dependencies effectively, which obtain impressive performances on remotely sensed image classification [23-29]. Especially, the Swin Transformer [30] presents a hierarchical structure, which further enhances its potential for semantic segmentation and achieves state-of-the-art results.

Driven by this, in this Letter, we propose a class-guided Swin Transformer with a Transformer-based encoder-decoder architecture for remote sensing image segmentation. We introduce the Swin Transformer backbone as the encoder. Since category-based information is a vital factor for precise semantic

This Research was supported by the Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical Area of South China, Ministry of Natural Resources (No. 2022NRMZ02) (Xiaoliang Meng and Yuechi Yang contributed equally to this work; Corresponding author: *Libo Wang*).

X. Meng, Y. Yang, and L. Wang are with School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: xmeng@whu.edu.cn; yangyuechi@whu.edu.cn; wanglibo@whu.edu.cn).

T. Wang is with Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical Area of South China, Ministry of natural resources, Surveying and Mapping Institute Lands and Resource Department of Guangdong Province, and Guangdong Science and Technology Collaborative Innovation Center for Natural Resources, Guangzhou 510663, China (e-mail: wangteng43@hotmail.com).

R. Li is with the Intelligent Control & Smart Energy (ICSE) Research Group, School of Engineering, University of Warwick, Coventry CV4 7AL, UK (e-mail: rui.li.4@warwick.ac.uk).

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).

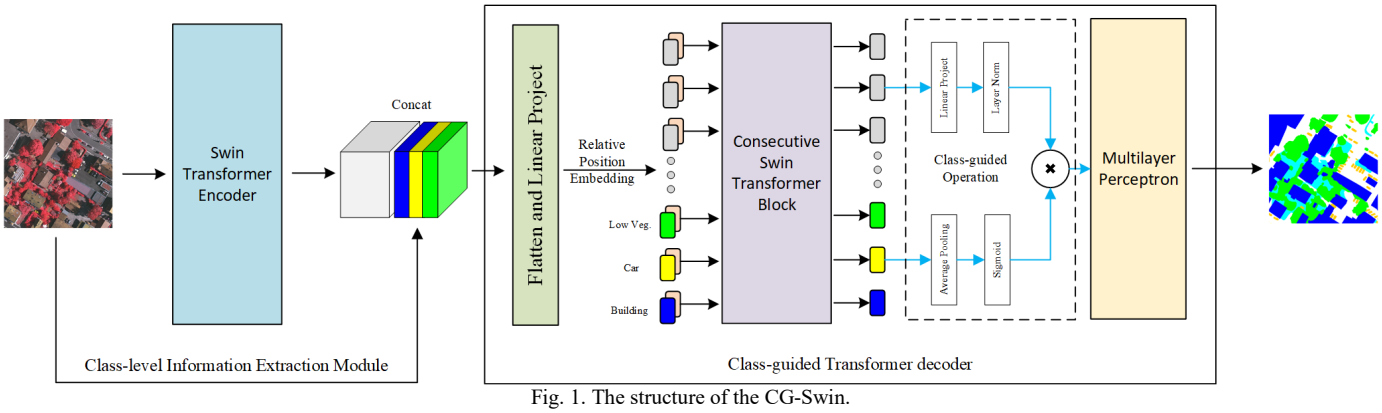

Fig. 1. The structure of the CG-Swin.

segmentation [31], we develop a class-guided Transformer decoder. Specifically, we first apply stacked convolution layers to extract the class-level semantic feature automatically from the input remote sensing image, then process it jointly with the extracted feature from the Swin Transformer encoder. The performance of the proposed method is tested on two popular remote sensing image segmentation datasets, i.e. ISPRS Vaihingen and Potsdam datasets, and compares with ten benchmark approaches comprehensively, including both advanced CNN-based and recent Vision Transformers based approaches.

II. Methodology

The proposed CG-Swin follows the fully Transformer-based encoder-decoder architecture, which obtains pixel-level class annotations from non-overlapping image patches. The overview of the CG-Swin is shown in Fig. 1.

A. Encoder

The small version of the Swin Transformer (Swin-S) [30] is selected as the encoder due to its excellent trade-off between efficiency and accuracy. As shown in Fig. 1, the encoder takes an image $X \in \mathbb{R}^{3\times H\times W}$ as the input and generates four semantic features at different channel dimensions and resolutions. Thus, we utilize standard 3×3 convolution layers and up-sampled operations to unify the shape of the four feature maps. Finally, we merge the four feature maps and generate an encoding feature $X_e \in \mathbb{R}^{d\times\frac{H}{4}\times\frac{W}{4}}$, where d is the channel dimension and set as 96. This encoding feature is fed into the decoder for further processing.

B. Decoder

The class-guided Transformer decoder is composed of three stages. In the first stage, it applies the class-level information extraction module to capture the category-based semantic feature. The class-level information extraction module employs three 3×3 convolution layers to capture the class-level context automatically. Each convolution layer is equipped with a Batch Normalization operation and a ReLU activation function. The extracted class-level feature can be denoted by $X_c \in \mathbb{R}^{k\times\frac{H}{4}\times\frac{W}{4}}$, where k is the number of categories.

In the second stage, we first flatten the concatenated feature to a 1D vector $X_{Cat} \in \mathbb{R}^{\frac{HW}{16}\times(d+k)}$ and linearly project it to a sequence of tokens. Then, the learnable relative position embedding $X_{Pos} \in \mathbb{R}^{\frac{HW}{16}\times(d+k)}$ is embedded with the X_{Cat} for improving positional information. A consecutive Swin Transformer block with two layers is selected to optimize the X_{Cat}. The consecutive Swin Transformer block can be defined as:

$$\overline{X}^1 = WMSA(LN(X^0)) + X^0 \quad (1)$$

$$X^1 = MLP(LN(\overline{X}^1)) + \overline{X}^1 \quad (2)$$

$$\overline{X}^2 = SWMSA(LN(X^1)) + X^1 \quad (3)$$

$$X^2 = MLP(LN(\overline{X}^2)) + \overline{X}^2 \quad (4)$$

where X^0 denotes the input vector X_{Cat}, $\overline{X}^1$ and $\overline{X}^2$ represents the output vector of the window-based multi-head self-attention module (WMSA) and the shifted window-based multi-head self-attention(SWMHSA) [30]. X^1 and X^2 denotes the middle vector and the output vector, respectively. *LN* is the Layer Norm operation, and *MLP* denotes the multilayer perceptron.

Finally, we separate the optimized X_{Cat} into the X_e and X_c and apply the class-guided operation to enhance the context interaction between them. The class-guided operation consists of two parallel paths, i.e. a compression path and an attention path. The compression path employs the Linear Projection and Layer Norm operation to reduce the channel dimension of the X_e to the number of categories. The attention path applies an average pooling layer and a Sigmoid activation function to generate a class-guided attention matrix. The total function of the class-guided operation can be defined as:

$$S(X_e, X_c) = LN(LP_k^d(X_e)) \odot Sigmoid(AvgPool(X_c)) \quad (5)$$

Where *LN* and *LP* denote the Layer Norm and Linear Projection, respectively. ⊙ represents the dot product operation. Processed by the class-guided operation, the compressed semantic feature X_e gains rich class-level information, which means that each channel dimension of X_e accurately characterized a category of ground objects. For precise per-pixel classification, it is crucial to obtain such a category-based semantic feature in the decoding period.

In the third stage, we utilize a multilayer perceptron to refine the category-based semantic feature and restore its resolution to the original, thereby producing the segmentation mask.

TABLE I
THE EXPERIMENTAL RESULTS ON THE VAIHINGEN DATASET.

Method	Backbone	Imp. surf.	Building	Low veg.	Tree	Car	Mean F1	OA	mIoU
FCN [1]	ResNet50	89.73	93.17	80.57	88.89	71.55	84.78	87.99	73.45
PSPNet [2]	ResNet101	92.79	95.46	84.51	89.94	88.61	90.26	90.85	82.58
DeepLabV3+ [3]	ResNet101	92.38	95.17	84.29	89.52	86.47	89.57	90.56	81.47
DANet [5]	ResNet101	91.63	95.02	83.25	88.87	87.16	89.19	90.44	81.32
DDCM-Net [6]	ResNet50	92.70	95.30	83.30	89.40	88.30	89.80	90.40	-
CASIA2 [11]	ResNet101	93.20	96.00	84.70	89.90	86.70	90.10	91.10	-
V-FuseNet [8]	FuseNet	91.00	94.40	84.50	89.90	86.30	89.20	90.00	-
UFMG_4 [15]	-	91.10	94.50	82.90	88.00	81.30	87.70	89.40	-
MANet [13]	ResNet50	93.02	95.47	84.64	89.98	**88.95**	90.41	90.96	82.71
BANet [14]	ResT-Lite	92.23	95.23	83.75	89.92	86.76	89.58	90.48	81.35
CMFNet [17]	-	92.36	**97.17**	80.37	90.32	85.47	89.24	90.43	80.82
CG-Swin(ours)	Swin-S	**93.55**	96.24	**85.70**	**90.59**	87.98	**90.81**	**91.68**	**83.39**

TABLE II
THE EXPERIMENTAL RESULTS ON THE POTSDAM DATASET.

Method	Backbone	Imp.surf.	Building	Low veg.	Tree	Car	MeanF1	OA	mIoU
FCN [1]	ResNet50	90.84	95.59	84.10	84.75	84.95	88.05	88.02	79.53
DeepLabV3+ [3]	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 [6]	ResNet50	92.90	96.90	87.70	89.40	94.90	92.30	90.80	-
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 [8]	FuseNet	92.70	96.30	87.30	88.50	95.40	92.04	90.60	-
DST_5 [10]	FCN	92.50	96.40	86.70	88.00	94.70	91.66	90.30	-
MANet [13]	ResNet50	93.40	96.96	88.32	89.36	96.48	92.90	91.32	86.95
BANet [14]	ResT-Lite	93.13	96.37	87.31	88.01	95.79	92.12	90.76	85.62
CMFNet [17]	-	92.84	**97.63**	88.00	87.40	95.68	92.10	91.16	85.63
CG-Swin(ours)	Swin-S	94.07	97.42	**88.53**	**89.74**	**96.61**	**93.28**	**91.93**	**87.61**

III. EXPERIMENTAL RESULTS

A. Dataset

We select the ISPRS Vaihingen and Potsdam semantic labelling datasets to test the effectiveness of the proposed method. There are 33 tiles with an average size of 2494×2064 pixels in the Vaihingen dataset and 38 tiles with a size of 6000×6000 pixels in the Potsdam dataset. Following previous pieces of literature [6], in the Vaihingen dataset, we use the officially provided 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. The image tiles are cropped into 1024×1024 px patches as the input. As for data augmentation, we use the random scale with a range from 0.5 to 1.5, random flip and random crop. Besides, the multi-scale strategy is applied in the testing period.

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 learning rate of 6e-4. For each method, the overall accuracy (OA), mean Intersection over Union (mIoU), and F1-score (F1) are selected as evaluation metrics:

$$OA = \frac{\sum_{k=1}^{N} TP_k}{\sum_{k=1}^{N} TP_k+FP_k+TN_k+FN_k}, \quad (6)$$

$$mIoU = \frac{1}{N}\sum_{k=1}^{N} \frac{TP_k}{TP_k+FP_k+FN_k}, \quad (7)$$

$$precision = \frac{1}{N}\sum_{k=1}^{N} \frac{TP_k}{TP_k+FP_k}, \quad (8)$$

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

$$F1 = 2 \times \frac{precision \times recall}{precision+recall}, \quad (10)$$

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) Quantitative Comparison: The experimental results on the Vaihingen and Potsdam datasets are listed in Table I and Table II. The quantitive metrics illustrate the effectiveness of the proposed method. Specifically, the CG-ViT achieves 90.81% in mean F1 score, 91.68% in OA, and 83.39% in mIoU for the Vaihingen dataset, with 93.28%, 91.93%, and 87.61% for the Potsdam dataset, outperforming the advanced CNN-based methods like DDCM-Net [6] and UFMG_4 [15]. Benefitting from the category-level semantic information modelled by the class-guided Transformer block, the performance of our method also outperforms recent Vision Transformers designed initially for remote sensing images, such as BANet [14] and CMFNet [17]. The visual comparisons with the FCN further illustrate the advantages of the proposed method (Fig. 2).

2) Ablation Study: To investigate the contribution of the class-guided mechanism to accuracy, we conduct the ablation study on the Vaihingen and Potsdam datasets. As shown in Table III, the utilization of the class guidance provides significant improvements, i.e. 2.21% and 1.65% mIoU for the Vaihingen and Potsdam test sets, respectively.

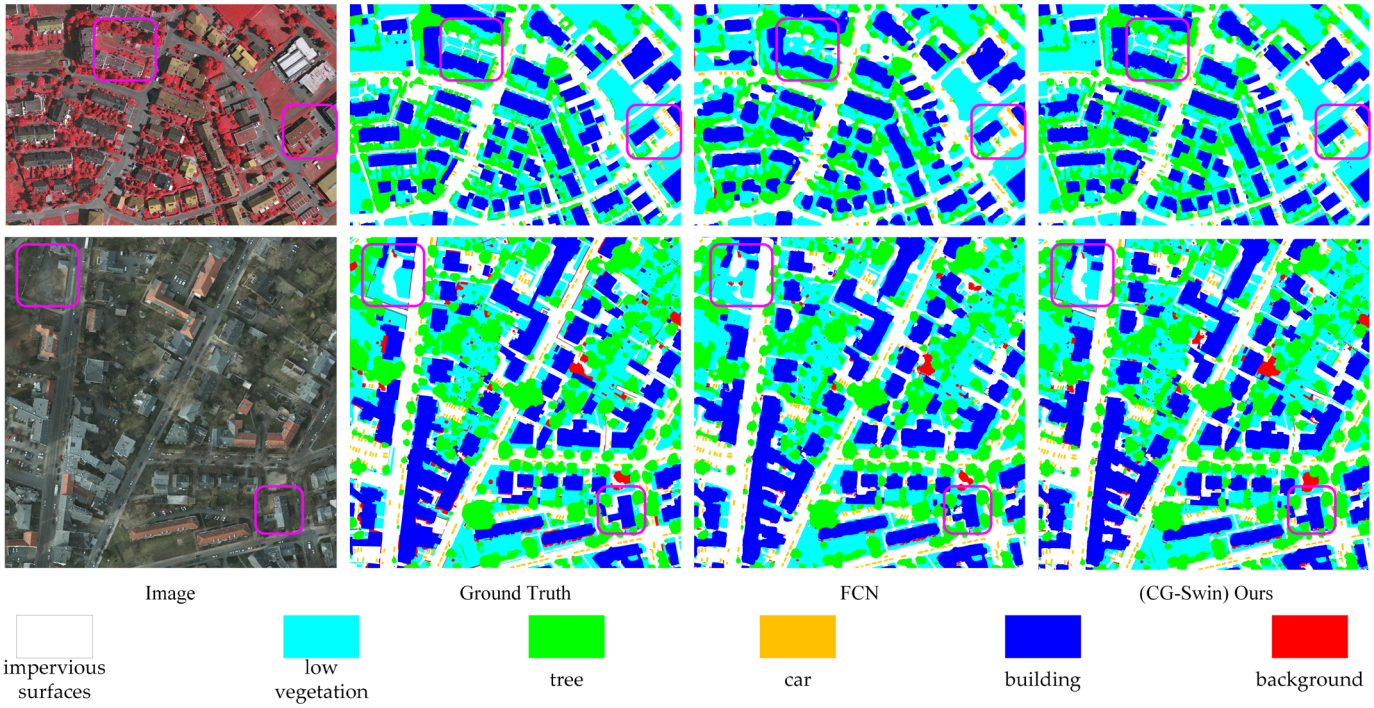

Fig. 2. Predicted results of the CG-ViT and FCN on the Vaihingen dataset (Top) and the Potsdam dataset (Bottom).

TABLE III
ABLATION STUDY ON THE VAIHINGEN AND POTSDAM DATASETS.

Dataset	Method	mIoU
Vaihingen	CG-Swin without class guidance	81.18
	CG-Swin with class guidance	83.39
Potsdam	CG-Swin without class guidance	85.96
	CG-Swin with class guidance	87.61

IV. CONCLUSION

In this Letter, we construct a fully Transformer network, namely the class-guided Swin Transform er (CG-Swin), for semantic segmentation of remote sensing images. We design a class-guided mechanism and combined it with the Swin Transformer block to construct a class-guided Transformer decoder. Numerical experiments conducted on the ISPRS Vaihingen and Potsdam datasets with benchmark comparison demonstrate the benefits of the class-guided mechanism and the effectiveness of the proposed method in segmentation accuracy.

REFERENCE

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

[2] 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] 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.

[4] X. X. Zhu *et al.*, "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources," *IEEE Geoscience and Remote Sensing Magazine,* vol. 5, no. 4, pp. 8-36, 2017, doi: 10.1109/MGRS.2017.2762307.

[5] 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.

[6] 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.

[7] 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 Journal of Photogrammetry and Remote Sensing,* vol. 145, pp. 96-107, 2018/11/01/ 2018, doi: https://doi.org/10.1016/j.isprsjprs.2018.01.021.

[8] 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.

[9] M. Y. Yang, S. Kumaar, Y. Lyu, and F. Nex, "Real-time Semantic Segmentation with Context Aggregation Network," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 178, pp. 124-134, 2021.

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

[11] 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.

[12] X.-Y. Tong *et al.*, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sensing of Environment,* vol. 237, p. 111322, 2020.

[13] R. Li *et al.*, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing,* 2021.

[14] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images," *Remote Sensing,* vol. 13, no. 16, p. 3065, 2021.

[15] K. Nogueira, M. Dalla Mura, J. Chanussot, W. R. Schwartz, and J. A. Dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 57, no. 10, pp. 7503-7520, 2019.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Cham, 2015: Springer International Publishing, in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234-241.

[17] X. Ma, X. Zhang, and M.-O. Pun, "A Crossmodal Multiscale Fusion Network for Semantic Segmentation of Remote Sensing Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 15, pp. 3463-3474, 2022.

[18] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 162, pp. 94-114, 2020.

[19] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7262-7272.

[20] R. Li, S. Zheng, C. Zhang, C. Duan, L. Wang, and P. M. Atkinson, "ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 181, pp. 84-98, 2021/11/01/ 2021, doi: https://doi.org/10.1016/j.isprsjprs.2021.09.005.

[21] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794-7803.

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

[23] L. Wang, R. Li, C. Duan, C. Zhang, X. Meng, and S. Fang, "A Novel Transformer Based Semantic Segmentation Scheme for Fine-Resolution Remote Sensing Images," *IEEE Geoscience and Remote Sensing Letters,* vol. 19, pp. 1-5, 2022, doi: 10.1109/LGRS.2022.3143368.

[24] D. Hong *et al.*, "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Transactions on Geoscience and Remote Sensing,* 2021.

[25] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An Empirical Study of Remote Sensing Pretraining," *IEEE Transactions on Geoscience and Remote Sensing,* 2022.

[26] L. Ding *et al.*, "Looking Outside the Window: Wider-Context Transformer for the Semantic Segmentation of High-Resolution Remote Sensing Images," *arXiv e-prints,* p. arXiv: 2106.15754, 2021.

[27] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote-Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 60, pp. 1-19, 2022.

[28] P. Deng, K. Xu, and H. Huang, "When CNNs meet vision transformer: A joint framework for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters,* vol. 19, pp. 1-5, 2021.

[29] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Transactions on Geoscience and Remote Sensing,* 2021.

[30] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012-10022.

[31] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 60, pp. 1-18, 2021.