Integrating Spatial Details With Long-Range Contexts for Semantic Segmentation of Very High-Resolution Remote-Sensing Images

Jiang Long, Mengmeng Li, *Member, IEEE*, and Xiaoqin Wang

***Abstract*— This letter presents a cross-learning network (i.e., CLCFormer) integrating fine-grained spatial details within long-range global contexts based upon convolutional neural networks (CNNs) and transformer, for semantic segmentation of very high-resolution (VHR) remote-sensing images. More specifically, CLCFormer comprises two parallel encoders, derived from the CNN and transformer, and a CNN decoder. The encoders are backboned on SwinV2 and EfficientNet-B3, from which the extracted semantic features are aggregated at multiple levels using a bilateral feature fusion module (BiFFM). First, we used attention gate (ATG) modules to enhance feature representation, improving segmentation results for objects with various shapes and sizes. Second, we used an attention residual (ATR) module to refine spatial features's learning, alleviating boundary blurring of occluded objects. Finally, we developed a new strategy, called auxiliary supervise strategy (ASS), for model optimization to further improve segmentation performance. Our method was tested on the WHU, Inria, and Potsdam datasets, and compared with CNN-based and transformer-based methods. Results showed that our method achieved state-of-the-art performance on the WHU building dataset (92.31% IoU), Inria building dataset (83.71% IoU), and Potsdam dataset (80.27% MIoU). We concluded that CLCFormer is a flexible, robust, and effective method for the semantic segmentation of VHR images. The codes of the proposed model are available at https://github.com/long123524/CLCFormer.**

***Index Terms*— Auxiliary supervise, CLCFormer, convolutional neural networks (CNNs), semantic segmentation, transformer, very high-resolution (VHR) images.**

I. Introduction

SEMANTIC segmentation has been widely used in the remote-sensing community for thematic information extraction, such as land cover and land use mapping and change detection [1], [2], [3], [4]. Among many, a particular interest has been drawn to extracting fine-grained land cover types from very high-resolution (VHR) remote-sensing images [5], [6], [7], [8].

Manuscript received 28 November 2022; revised 9 February 2023 and 8 March 2023; accepted 24 March 2023. Date of publication 28 March 2023; date of current version 6 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42001283 and in part by the Science and Technology Program of Fujian Province of China under Grant 2022N0019 and Grant 2022C0024. *(Corresponding author: Mengmeng Li.)*

The authors are with the Key Laboratory of Spatial Data Mining and Information Sharing of the Ministry of Education, Academy of Digital China (Fujian), Fuzhou University, Fuzhou 350002, China (e-mail: 205527028@fzu.edu.cn; mli@fzu.edu.cn; wangxq@fzu.edu.cn).

Digital Object Identifier 10.1109/LGRS.2023.3262586

Conventional methods for VHR semantic segmentation extract handcrafted features of texture, spectral, and geometry and then apply machine-learning methods, for example, random forests and support vector machines, to identify the interest of objects [9]. These methods, however, fail to fully use high-level semantic features at the fine granularity. Recently, convolutional neural networks (CNNs) have been successfully used for the semantic segmentation of VHR images, which automatically extract high-level semantic features with fine-grained spatial details [10], [11], [12]. Popular networks are mainly based upon encoder–decoder architectures, in which the encoder extracts multiscale semantic features, and the decoder is used to refine extracted features [11]. Generally, CNN-based models are insufficient to capture long-range global context information, leading to a low segmentation accuracy for objects with small sizes, shadow occlusion, and high interclass similarity. To deal with this issue, many studies have used attention mechanisms to strengthen semantic feature learning [13], [14], [15], [16]. These CNN methods, however, are still difficult to extract spatial and spectral information in a long-range global context.

In recent years, increasing attention has been drawn to using vision transformer (ViT), which uses a self-attention to model global context information [17], [18]. Usually, ViT-based methods require huge training costs, challenging model training. An interesting work is by [19], who proposed a Swin transformer using a shifted window to model global contexts between image patches, lowering computational complexity. So far, the Swin transformer has also been used for semantic segmentation of VHR images [20], [21], [22], due to its robust ability in global context modeling. Although great success has been made using existing transformer methods, we believed that deep integration of fine-grained spatial details at the local scale can further boost the performance of semantic segmentation.

In this study, we proposed a cross-learning network integrating fine-grained spatial details within long-range global contexts based upon the transformer and CNN, called CLCFormer, for semantic segmentation of VHR images. Differing from [22], [23], and [24], we fused semantic features derived from the CNN and transformer at multiscales rather than only using the global-scale features from the transformer. Moreover, we designed more effective attention modules (i.e., bilateral feature fusion module (BiFFM), attention gate (ATG)

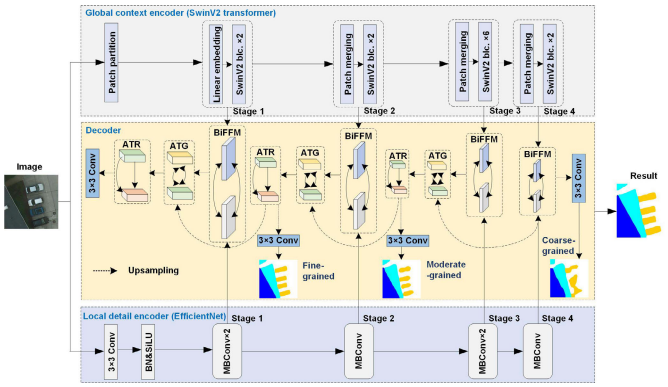

Fig. 1. Overall architecture of CLCFormer for semantic segmentation of VHR images. It consists of SwinV2 transformer blocks (SwinV2 block) [25], MBConv blocks [26], ATR modules, ATG modules, and BiFFM.

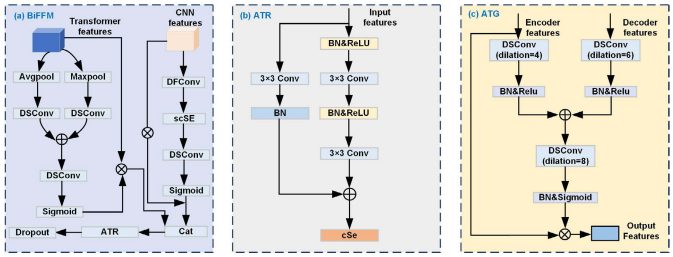

Fig. 2. Details of the proposed BiFFM, ATR module, and ATG module.

block, and attention residual (ATR) block) based on expert knowledge to strengthen semantic learning and representation. The contributions of our study lie in the following.

1) It presents an effective network for semantic segmentation of VHR images. The BiFFM is developed to aggregate discriminative features from the CNN and transformer to distinguish objects with high interclass similarity.
2) It designs a ATG module to enhance semantic feature learning at different scales to characterize objects with various shapes and sizes, and the ATR to enhance meaningful semantic extraction, alleviating boundary blurring caused by shadow occlusion.
3) It develops an auxiliary supervising strategy (ASS) for model optimization to accelerate the model's convergence.

II. Methods

The overall architecture of the proposed CLCFormer is illustrated in Fig. 1. It is characterized by dual encoders and one decoder, and the ability to capture long-range global contexts while maintaining fine-grained spatial details.

A. CLCFormer Architecture

CLCFormer has dual encoders based upon the transformer and CNN (Fig. 1). The use of these two encoders is to strengthen the ability of CNNs to capture long-range global contexts and to remedy the insufficient extraction of fine-grained details by the transformer in turn. We created a BiFMM [Fig. 2(a)] to aggregate extracted semantic features at different levels. Moreover, we developed an ATG module [Fig. 2(c)] using atrous convolutions to capture contextual semantics within a large neighborhood. Next, an ATR module [Fig. 2(b)] with channel attention is used to improve feature representation. Last, we designed an auxiliary supervise strategy (ASS) to construct multiple auxiliary outputs at different levels to optimize model training.

Specifically, the transformer encoder uses a popular SwinV2 backbone pretrained on the ImageNet [25]. It first splits an image into nonoverlapping patches (i.e., tokens) and then uses a linear embedding layer to project the features of each patch into dimension C. Next, SwinV2 blocks based upon windows multihead self-attention (W-MSA) and shift W-MSA (SW-MSA) are used to extract global contextual features. The resolution of a patch is gradually reduced by patch merging layers when increasing the depth of the network. For an input image $I \in R^{H\times W\times 3}$, we obtained feature maps $F_i \in R^{(H/2^i)\times(W/2^i)\times C_i}$ at four levels (Fig. 1), where $C_i \in \{96, 192, 384, 768\}$ and $i \in \{2, 3, 4, 5\}$.

The CNN encoder uses a robust and lightweight EfficientNet-B3 backbone pretrained on the ImageNet [26]. More specifically, it first uses a 3×3 convolution to extract shallow features, followed by a batch normalization (BN) and Sigmoid weighted linear unit (SiLU) function. Subsequently, a series of mobile inverted bottleneck convolution (MBConv) blocks are used to capture multilevel spatial details. By doing so, the CNN-based encoder reduces huge parameters while preserving fine-grained spatial details. Similarly, four feature maps $F_i \in R^{(H/2^i)\times(W/2^i)\times C_i}$ at different levels are obtained (Fig. 1), where $C_i \in \{48, 96, 136, 232\}$ and $i \in \{2, 3, 4, 5\}$.

Moreover, we constructed an enhanced version of CLCFormer by replacing the SwinV2 backbone with SwinV2-S [25] for the transformer encoder, denoted as CLCFormer-S.

B. Bilateral Feature Fusion Module

It is known that CNN-based encoders are difficult to model global context information due to the locality of convolution operations. By contrast, transformer-based encoders are robust in extracting context information with a long-range, however, they are insufficient in capturing fine-grained spatial features. Here, we used a BiFFM [Fig. 2(a)] to fuse these discriminative semantic cues at different levels to improve the identification of target objects with high interclass similarity.

Let T_n and C_n be the features derived from the transformer and CNN encoders at the nth level. For CNN-based features, we first used a 3×3 deformable convolution (DFConv) to capture the shape variety of target objects in VHR images. A concurrent spatial and channel squeeze and channel excitation (scSE) attention module is used to enhance related features with the target objects [27]. Next, we used depthwise separable convolutions (DSConv) to further refine obtained features without increasing computation complexity. Last, we obtained refined features P_C by element-wise multiplication

$$P_C = \delta(\mathrm{DSC}(\mathrm{scSE}(\mathrm{DFC}(C_n)))) \times C_n. \quad (1)$$

For transformer-based features, we first used two pooling operations to capture comprehensive channel cues.

TABLE I
COMPARISON OF EXTRACTION RESULTS BETWEEN THE PROPOSED METHOD AND EXISTING METHODS ON THE WHU DATASET (%)

Methods	P	R	F1	IoU
SiU-Net [31]	93.80	93.90	93.85	88.40
B-FGC-Net [10]	95.03	94.49	94.76	90.04
BOMSC-Net [5]	95.14	94.50	94.80	90.15
STT [18]	-	-	94.97	90.48
EU-Net [11]	94.98	95.10	95.04	90.56
Swin-U-S [20]	95.04	95.18	95.11	90.68
MAP-Net [13]	95.62	94.81	95.21	90.86
LFENET [14]	-	-	95.30	91.01
CBRNet [8]	95.31	95.70	95.51	91.40
BuildFormer [21]	**95.65**	95.40	95.53	91.44
CLCFormer	95.42	96.03	95.72	92.14
CLCFormer-S	95.45	**96.09**	**95.77**	**92.31**

TABLE II
COMPARISON OF EXTRACTION RESULTS BETWEEN THE PROPOSED METHOD AND EXISTING METHODS ON THE INRIA DATASET (%)

Methods	P	R	F1	IoU
SiU-Net [31]	84.60	82.10	83.33	71.40
BOMSC-Net [5]	87.93	87.58	87.75	78.18
LFENET [14]	-	-	88.37	79.16
B-FGC-Net [10]	87.82	89.12	88.46	79.31
STT [18]	-	-	87.99	79.42
ASF-Net [15]	-	-	-	80.20
DS-Net [32]	-	-	-	80.73
CBRNet [8]	89.93	89.20	89.56	81.10
BuildFormer [21]	**90.75**	88.81	89.77	81.44
CLCFormer	90.04	89.83	89.93	83.24
CLCFormer-S	90.66	**89.98**	**90.32**	**83.71**

More specifically, we applied max- and average-pool layers to obtain statistical features at channel levels and fed them to the DSConv layer, called S_{AM}. This process can be described by the following equation:

$$S_{\mathrm{AM}} = \mathrm{DSC}(\mathrm{AvgPool}(T_n)) + \mathrm{DSC}(\mathrm{MaxPool}(T_n)). \quad (2)$$

We then further refined the features of each channel by multiplying S_{AM} with T_n

$$S_F = \delta(\mathrm{DSC}(\mathrm{ReLU}(S_{\mathrm{AM}}))) \times T_n. \quad (3)$$

Last, the refined features S_F and P_C obtained from the transformer and CNN were concatenated in CLCFormer. Moreover, we used an ATR module to enhance feature representation, and a dropout layer to avoid model overfitting.

C. ATG Module and ATR Module

Conventional encoder–decoder networks like ResUNet [28] use skip connections to aggregate low- and high-level semantics. Such practice, however, fails to effectively utilize rich semantic cues from different scales, decreasing the segmentation accuracy for objects with various shapes and sizes. To deal with this issue, inspired by [29], we constructed an ATG based upon attention mechanisms to enhance the spatial relations between encoder features E_F and decoder features D_F [Fig. 2(c)], and to suppress irrelevant cues. ATG uses DSConv with different dilated rates to capture more contextual information at different scales while maintaining a low computation complexity. Furthermore, we used an element-wise multiplication operation to refine the obtained encoder features F_{ED}:

$$F_{\mathrm{ED}} = \delta(\mathrm{BN}(\mathrm{DSC}(\mathrm{BR}((\mathrm{DSC}(E_F) + \mathrm{DSC}(D_F)))))) \times E_F \quad (4)$$

where δ indicates a Sigmoid function and BR represents the BN layer and rectified linear unit (ReLU).

Moreover, we used ATR to enhance the extraction of spatial features I_F with fine-grained details to alleviate boundary blurring caused by shadow occlusion. The ATR module is detailed in Fig. 2(b). Here, the ATR is formulated as follows:

$$O_F = \mathrm{cSe}(\sigma(\mathrm{BR}(\sigma(\mathrm{BR}(I_F))))) + \mathrm{BN}(\sigma(I_F))) \quad (5)$$

where σ is a 3×3 convolution, and *cSe* is the spatial squeeze and channel excitation module used for enhancing salient features at the channel level [27].

III. EXPERIMENTAL RESULTS

A. Datasets

We tested the proposed method on three public datasets: WHU, Inria, and Potsdam datasets. The WHU dataset consists of 8189 image tiles, where 7436 tiles were used for model training, 1036 tiles for validation, and 2416 tiles for testing. Each image tile has a size of 512 × 512 pixels with a spatial resolution of 0.3 m. The Inria dataset contains 360 image tiles (0.3 m) collected from five cities (i.e., Kitsap, Tyrol, Austin, Chicago, and Vienna). Each image tile has a size of 5000 × 5000 pixels. For the Inria building dataset, we chose the first five image titles of each city for testing and the rest for training, as suggested by the official dataset instruction. We then cropped the image tiles of Inria dataset to 512 × 512 pixels for experimental convenience. The Potsdam dataset consists of 38 image tiles, each with a size of 6000 × 6000 pixels and a spatial resolution of 0.05 m. Following [22], we used 24 tiles for training and 14 tiles for testing, where each tile was also cropped to 256 × 256 pixels.

B. Implementation Details

To enlarge the training set, we conducted a data augmentation for all datasets, including vertical and horizontal flips. An AdamW optimizer with an initial learning rate of 10^{-4} is used to train the CLCFormer from scratch. Besides, we used a weighted cross-entropy loss and intersection over union loss to train the model for WHU and Inria datasets, while a soft cross-entropy loss and Dice loss for CLCFormer on the Potsdam dataset, according to [22] and [23]. High-level features usually include more robust semantic information compared to low- and middle-level features [23], [30]. Here, we gradually increased the weights of different levels and set them to 0.15:0.15:0.3:0.4. Last, we applied a test time augmentation strategy in the testing processing, to further improve segmentation results [21]. All experiments are conducted on an NVIDIA GeForce RTX 3090 24 GB GPU, using a batch size of 8 and an epoch of 25.

TABLE III

QUANTITATIVE COMPARISON OF SEGMENTATION RESULTS BY USING THE PROPOSED METHOD WITH THE SOTA METHODS ON THE POTSDAM DATASET

Methods	FLOPs	Parameters	IoU(%)					MIoU(%)	MF1(%)
			Impervious surface	Building	Low vegetation	Tree	Car		
SegViT [33]	24.16G	108.04M	70.67	77.02	61.23	52.96	57.65	63.91	77.63
ST-UNet [22]	19.69G	160.97M	72.21	78.42	66.42	64.12	67.37	69.71	82.05
CATS [24]	29.84G	91.22M	72.81	80.23	65.39	66.83	67.99	70.65	82.69
ResUNet [28]	94.56G	62.74M	75.59	80.65	67.66	69.00	79.59	74.50	85.28
TransFuse [23]	9.54G	**26.30M**	80.21	86.81	70.35	73.13	79.42	77.98	87.51
SSformer [34]	25.93G	87.50M	80.50	89.18	71.82	73.93	**80.46**	79.18	88.25
CLCFormer	**7.79G**	38.07M	80.85	88.57	72.78	74.28	77.84	78.86	88.07
CLCFormer-S	11.89G	54.11M	**82.47**	**90.04**	**73.02**	**75.64**	80.19	**80.27**	**88.94**

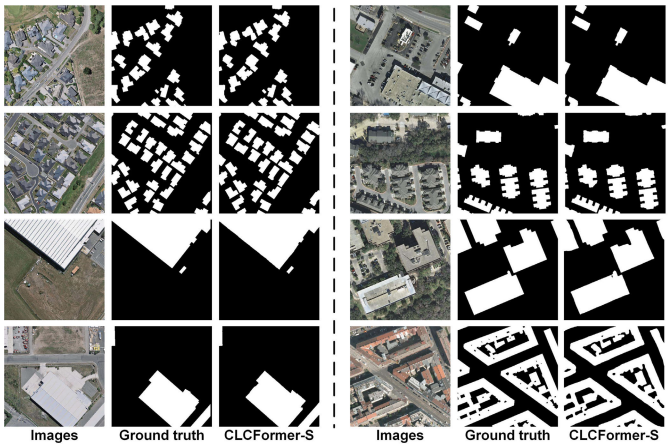

Fig. 3. Extracted buildings by our method on the WHU (left) and Inria (right) datasets.

C. Accuracy Assessment and Performance Evaluation

We applied the commonly used precision (P), recall (R), F1-score, mean F1-score (MF1), and mean intersection over union (MIoU) measures to evaluate the performance of the proposed method [22].

D. Result Analysis

1) Performance evaluation: To verify the robustness of our method, we compared it with state-of-the-art (SOTA) methods on the WHU, Inria, and Potsdam datasets. The building extraction results were evaluated in Tables I and II. These tables show that CLCFormer obtained a higher accuracy on the WHU and Inria building datasets, compared with the other existing methods. We observed that CLCFormer outperformed transformer-based methods, that is, STT, Swin-U-S, and BuildFormer, on the WHU and Inria datasets for all measures, and performed better than CNN-based methods as well. It implies that our method captured more robust building features, reducing false positives and negatives. Moreover, the enhanced CLCFormer variant (i.e., CLCFormer-S) yielded the highest accuracy on the WHU dataset (Table I). We also see that CLCFormer-S obtained the highest IoU of 83.71% and F1 of 90.32% on the Inria dataset (Table II), outperformed the recent CBRNet by 2.61% and 0.76%, and the advanced BuildFormer by 2.27% and 0.55%, respectively. Fig. 3 shows the building extraction results on the two datasets. Clearly, the extracted buildings by CLCFormer-S were close to the ground truth on the two datasets.

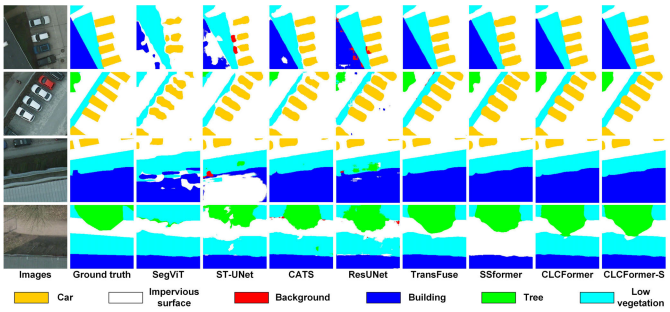

Fig. 4. Examples of semantic segmentation using different methods on the Potsdam dataset.

TABLE IV

ABLATION STUDY OF THE PROPOSED MODULES AND STRATEGY ON THE WHU DATASET (%)

Methods	P	R	F1	IoU
Baseline	94.65	95.09	94.87	90.65
Baseline+ASS	94.22	95.90	95.05	90.91
Baseline+ASS+BiFFM	95.02	95.28	95.15	91.28
Baseline+ASS+BiFFM+ATR	94.63	96.12	95.37	91.51
Baseline+ASS+BiFFM+ATR+ATG	95.42	96.03	95.72	92.14

For the Potsdam dataset, the proposed CLCFormer produced satisfying results with an MIoU of 78.86% and an MF1 of 88.07%, exceeding the recent ST-UNet by 9.15% and 6.02%, respectively (Table III). CLCFormer-S further improved the extraction for all categories. This implies that our method is flexible and can be replaced by many advanced backbones. Fig. 4 displays the land cover classification results by different methods. Noticeably, our methods (both CLCFormer and CLCFormer-S) produced more accurate segmentations than a CNN-based method (i.e., ResUNet), transformer-based methods (i.e., SSformer and SegViT), and hybrid CNN and transformer methods (ST-UNet, CATS, and TransFuse). This figure shows that our methods can be applied to different scenarios, even with shadow occlusion and a high interclass spectral similarity. We further assessed the computational efficiency and complexity of the proposed model and compared it with existing models (Table III). This table shows that the proposed CLCFormer and its variant (CLCFormer-S) have a smaller number of parameters and less computation time than the existing methods.

2) Ablation study: To evaluate the effectiveness of the proposed modules and strategy (i.e., BiFFM, ATR, ATG modules, and a strategy ASS), we conducted ablation experiments on the WHU dataset. The experimental results are shown in Table IV.

We can find that adding these modules and the ASS further improved semantic segmentation accuracies for all evaluation measures. The results further show that the BiFFM, ATR, and ATG modules have a robust ability to capture discriminative semantic cues at different levels, enhancing fine-grained spatial details, and improving semantic representation, respectively.

IV. Conclusion

In this letter, we proposed a cross-learning network (CLCFormer) based upon the transformer and CNN, achieving fine-grained spatial details with long-range global context semantics, for semantic segmentation of VHR images. We developed a BiFFM, an ATG block, and an ATR module to enhance meaningful semantic feature extraction. Moreover, an ASS is developed to speed up model training. We tested our method on the WHU, Inria, and Potsdam datasets and compare it with existing methods. Experimental results show that our method performed better than the existing methods and achieved state-of-the-art performance on all datasets, demonstrating the effectiveness of our method for land cover semantic segmentation using VHR images. Future studies can be conducted to verify the potential of CLCFormer for more tasks like object detection and instance segmentation.

References

[1] G. Liu, L. Li, L. Jiao, Y. Dong, and X. Li, "Stacked Fisher autoencoder for SAR change detection," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106971.

[2] L. Ding, J. Zhang, and L. Bruzzone, "Semantic segmentation of large-size VHR remote sensing images using a two-stage multiscale training architecture," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5367–5376, Aug. 2020.

[3] X. Zheng, T. Gong, X. Li, and X. Lu, "Generalized scene classification from small-scale datasets with multitask learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609311.

[4] X. Zheng, X. Chen, X. Lu, and B. Sun, "Unsupervised change detection by cross-resolution difference learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5606616.

[5] Y. Zhou et al., "BOMSC-Net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5618617.

[6] S. Saha, M. Shahzad, L. Mou, Q. Song, and X. X. Zhu, "Unsupervised single-scene semantic segmentation for Earth observation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5228011.

[7] R. Zhou et al., "Weakly supervised semantic segmentation in aerial imagery via explicit pixel-level constraints," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5634517.

[8] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 240–252, Jan. 2022.

[9] M. Turker and D. Koc-San, "Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 34, pp. 58–69, Feb. 2015.

[10] Y. Wang, X. Zeng, X. Liao, and D. Zhuang, "B-FGC-Net: A building extraction network from high resolution remote sensing imagery," *Remote Sens.*, vol. 14, no. 2, p. 269, Jan. 2022.

[11] W. Kang, Y. Xiang, F. Wang, and H. You, "EU-Net: An efficient fully convolutional network for building extraction from optical remote sensing images," *Remote Sens.*, vol. 11, no. 23, p. 2813, Nov. 2019.

[12] X. Zheng, H. Sun, X. Lu, and W. Xie, "Rotation-invariant attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 4251–4265, 2022.

[13] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.

[14] Z. Wan, Q. Zhang, and G. Zhang, "Low-level feature enhancement network for semantic segmentation of buildings," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[15] J. Chen, Y. Jiang, L. Luo, and W. Gong, "ASF-Net: Adaptive screening feature network for building footprint extraction from remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4706413.

[16] X. Zheng, B. Wang, X. Du, and X. Lu, "Mutual attention inception network for remote sensing visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5606514.

[17] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[18] K. Chen, Z. Zou, and Z. Shi, "Building extraction from remote sensing images with sparse token transformers," *Remote Sens.*, vol. 13, no. 21, p. 4441, Nov. 2021.

[19] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[20] C. Qiu et al., "Transferring transformer-based models for cross-area building extraction from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4104–4116, 2022.

[21] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625711.

[22] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding UNet for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408715.

[23] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2021, pp. 14–24.

[24] H. Li, D. Hu, H. Liu, J. Wang, and I. Oguz, "Cats: Complementary CNN and transformer encoders for segmentation," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.

[25] Z. Liu et al., "Swin transformer V2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12009–12019.

[26] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[27] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2018, pp. 421–429.

[28] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[29] O. Oktay et al., "Attention U-net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[30] X. Zheng, X. Chen, and X. Lu, "Visible-infrared person re-identification via partially interactive collaboration," *IEEE Trans. Image Process.*, vol. 31, pp. 6951–6963, 2022.

[31] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[32] H. Zhang, Y. Liao, H. Yang, G. Yang, and L. Zhang, "A local–global dual-stream network for building extraction from very-high-resolution remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 1269–1283, Mar. 2022.

[33] B. Zhang et al., "SegViT: Semantic segmentation with plain vision transformers," 2022, *arXiv:2210.05844*.

[34] W. Shi, J. Xu, and P. Gao, "SSformer: A lightweight transformer for semantic segmentation," in *Proc. IEEE 24th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2022, pp. 1–5.