ConvBNet: A Convolutional Network for Building Footprint Extraction

Tong Yu, Panpan Tang, Bo Zhao, Shi Bai, Peng Gou, Jiachun Liao, and Caifeng Jin

***Abstract*— Building footprint is a key indicator of urban structures and economic development. Automatic extraction of building footprint from very-high-resolution (VHR) remote sensing imagery, which is of great practical interest for various geospatial-related applications, is still a challenging task for complex textures, varying scales and shapes, and other confusing artificial objects. To alleviate these problems and improve the extraction accuracy, this study proposed a novel pure convolutional neural network called ConvBNet, which integrates ConvNeXt-XL with a fusing decoder and adopts deep supervision in the training stage for the middle stage. Two-stage training strategy, using weighted cross-entropy (CE) loss and mask CE loss, respectively, ensures the stable convergence and makes the network focus on the boundary region. The proposed model was tested on the Wuhan University (WHU) building dataset (publicly available) and one private Zhejiang building dataset. Compared with other state-of-the-art (SOTA) methods, ConvBNet achieved the best intersection over Union (IoU), 91.22% and 77.90% for the two datasets, which proves its good performance in the task of extracting buildings from VHR images.**

***Index Terms*— Building extraction, convolutional neural network, deep learning, remote sensing.**

I. Introduction

AS IMPORTANT artificial objects, the buildings have always been one of the popular research targets of earth observation. Efficient automatic extraction of them has practical application significance in urban change monitoring and planning, digital city, and land use analysis. Nowadays, multisourced very-high-resolution (VHR) remote sensing images are getting more accessible, which makes them become an indispensable basis and reference for building extraction. However, due to the inherent characteristics of buildings, such as varying shapes and sizes, complex textures, and similarities with other artificial objects, this task is still challenging. Different imaging angles, lighting conditions, and background information, also create difficulties. To extract buildings accurately, many traditional and deep learning-based methods have been proposed.

Manuscript received 29 November 2022; revised 16 January 2023; accepted 18 February 2023. Date of publication 1 March 2023; date of current version 9 March 2023. This work was supported in part by the Nonprofit Research Project of Jiaxing City under Grant 2022AY30001, in part by the Key Project of "Prospering Mongolia with Science and Technology" under Grant KJXM-EEDS-2020006, and in part by the National Natural Science Foundation of China under Grant 41901360. *(Corresponding author: Panpan Tang.)*

Tong Yu, Panpan Tang, Bo Zhao, Shi Bai, Peng Gou, and Jiachun Liao are with the Big Data Technology Research Center, Nanhu Laboratory, Jiaxing 314002, China (e-mail: yutong@nanhulab.ac.cn; tangpp@nanhulab.ac.cn).

Caifeng Jin is with the School of Civil Engineering and Architecture, Jiaxing Nanhu University, Jiaxing 314001, China.

Digital Object Identifier 10.1109/LGRS.2023.3250091

Traditional methods mainly include geometric element-based and overall features-based methods. Corners [1] and edges [2] are the commonly used elements to reconstruct building shapes. With empirical knowledge, overall features designed and calculated from color, shape, texture, and spectral information are applied for distinguishing building and non-building [3]. The above methods can be improved by adding some auxiliary information, like digital surface model (DSM) [4], building shadows [5], light detection and ranging (LiDAR) [6], etc. With the development of deep learning methods in image analysis, the end-to-end convolutional neural networks have greatly increased the accuracy in many downstream tasks and reduced the importance of the subjective factors that need to be manually adjusted. Fully convolutional networks (FCNs) [7], UNet and its variants [8], Deeplab v3+ [9], high-resolution net (HRNet) [10], and other semantic segmentation networks can well meet the requirements of pixel-level building extraction in remote sensing. The building extraction networks rely heavily on scene classification networks and semantic segmentation networks. Multiband inputs enrich the priori knowledge and enhanced the distinguishing ability of Res-U-Net [11]. Combined with densely connected convolution and residual design, Deeplab v3+ achieved good accuracy with a small number of parameters [12]. Multiresolution feature propagation, fusion, and efficient attention module contribute to a competitive solution to building extraction [13]. In addition, the attention mechanism has been demonstrated to be helpful for feature extraction, which can selectively focus on some information. Spatial and channel attention [14], [15], attention gate [16] have been applied to fine-tune the outputs of encoders.

Recently, the transformer which was first proposed in natural language processing has been adapted to many computer vision (CV) tasks [17], [18] and achieved state-of-the-art (SOTA) performances. But its strong capability of modeling long-range dependencies between pixels is not enough to make up for the deficiency of local feature and position information. Some attempts for building extraction [19], [20], [21] did not fully outperform the existing methods, so the CNNs still have advantages in this field. ConvNeXt [22], a pure convolutional network, improves the residual neural network (ResNet) toward the design of the Swin Transformer [23] in structure and achieves better accuracy, showing its great potential for CV tasks. We chose it as the backbone, adding attention modules and a decoder to make it meet the requirements of semantic segmentation. The new proposed network is called

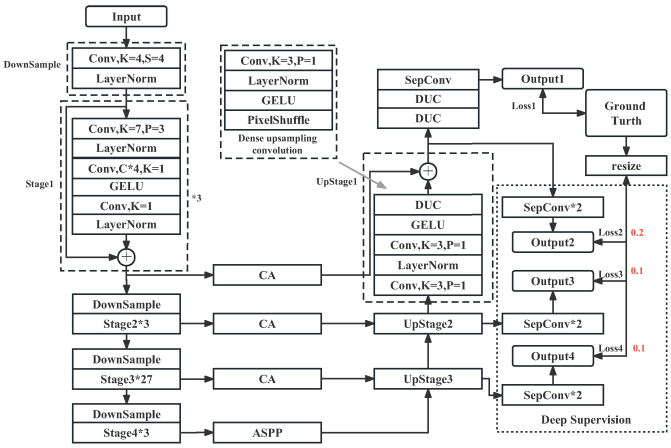

Fig. 1. Architecture of ConvBNet.

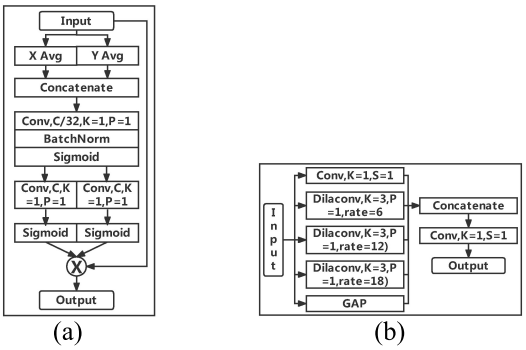

Fig. 2. Architectures of (a) CA and (b) ASPP.

ConvBNet, and it achieves a competitive accuracy. What is more, a multistage multiloss training strategy was adopted to achieve balanced convergence and boundary area optimization.

II. Methods

ConvBNet network combines ConvNeXt, attention module, and deep supervision, and the detailed architecture is shown in Fig. 1.

A. Encoder

ConvNeXt-XL, consisting of DownSample block and Stage block, is applied as an encoder to obtain features of four stages. Referring to the structural design of the Swin Transformer, it has few activation layers and normalization layers, and uses the gaussian error linear unit (GeLu) and layer normalization (LN) [24] function. Each encoding stage contains only one DownSample block, but the Stage block is repeated 3, 3, 27, and 3 times, respectively. DownSample block is achieved by a convolutional layer with a stride (S) and kernel size (K) of 2, while that of the first DownSample block is 4. Then the convolutional layer is followed by an LN layer. Stage block selects the convolution with a 7×7 large kernel and three padding pixels (P) as the first layer. The second convolutional layer with a 1×1 kernel increases the feature channels (C) by 4 times and the third convolutional layer with a 1×1 kernel restores it. Then the feature is added with input to obtain the stage output. The sizes of the four stages' output are 64×64, 32×32, 16×16, and 8×8. These feature maps will be fed into the intermediate stage. ConvNeXt-XL has a strong ability to extract multilevel abstract features of the inputs.

B. Intermediate Stage

For the feature optimization, coordinate attention (CA) [25] is used in the intermediate stages from 1 to 3, and atrous spatial pyramid pooling (ASPP) [26] is used in stage 4. Their architectures can be seen in Fig. 2. As a lightweight and semantic segmentation suitable module, CA captures interchannel relationships and long-range dependencies. It also contributes to locate and restore the features. The ASPP module performs one global average pooling (GAP) and four convolutional operations, including three atrous convolutional layers, on the final output of the encoder. It expands the receptive field of the convolutional kernel and extracts global information without losing resolution. The five output features in ASPP module are concatenated and then fused by a convolutional layer.

C. Decoder

The UpStage blocks in the decoder upsample and aggregate the corresponding outputs to restore their spatial dimensions step by step. Each UpStage block has two convolutional layers, followed by LN and GeLu, respectively, and dense upsampling convolution (DUC) module [27]. Each convolutional layer doubles the channels. The learnable DUC module can capture and restore feature details that are lost in down sampling without supplement the padding information. The structure of DUC module is shown in Fig. 1, and the core is the pixel shuffle layer. It reorganizes the pixels by reducing the number of channels and increasing the spatial size (e.g., $H \times W \times 4C$ to $2H \times 2W \times C$), while the total number of pixels remains unchanged. The upsampled feature is added with the output of the upper intermediate stage. Because of the $4\times$ down sampling convolution in the encoder at the beginning, the decoder also needs to connect two DUC modules on the top before the classification.

Deep supervision [28] is a training technique of adding an auxiliary classifier to some intermediate hidden layers of the deep neural network. It is a general practice and has been used in building extraction tasks [29], [30]. The channels of middle layers are much higher than the number of classes. Therefore, we add two layers of separable convolution (SepConv) [31] as buffer layers between the output of each UpStage and classifier to reduce the number of channels. Small weights are set to these auxiliary losses.

III. Experiments and Results

A. Datasets

To validate the performance of the methods, comparison and ablation experiments were conducted on one publicly available dataset named Wuhan University (WHU) [32], and one private dataset containing Zhejiang buildings. The details are described as follows.

WHU building dataset is composed of an aerial imagery dataset and two satellite imagery datasets. The aerial dataset,

created from the cropped red, green, and blue (RGB) aerial imagery of New Zealand, was used in our experiments for its clear images, accurate labels, and large number of samples. The samples have been split into a training set (4736 tiles), a validation set (1036 tiles), and a test set (2416 tiles). The size of each tile is 512 × 512 pixels, and the spatial resolution is 0.3 m. Following the original partition and resolution, we cropped the tiles into 256 × 256 pixels with no overlap and the number of tiles has quadrupled.

The original RGB images in the Zhejiang building dataset are obtained from the Worldview satellite. According to the existing building footprint products and visual interpretation, we removed the wrong labels and manually labeled some images. The updated dataset was also divided into a training set (8108 tiles), a validation set (1026 tiles), and a test set (2065 tiles). The size of each tile is 256 × 256 pixels, and the spatial resolution is about 0.5 m.

B. Training Implementation

The weighted cross-entropy (WCE) loss and the mask CE (MCE) loss were applied in training stages 1 and 2. Given the ubiquitous sample imbalance in datasets, a 2:1 calculation weight was set for buildings and the background. The loss function in stage 1 consists of a main loss and three auxiliary losses, and the weights of the auxiliary losses are 0.2, 0.1, and 0.1, respectively. The MCE loss generates a boundary mask first, which divides the images into the boundary area (5 pixels inside and outside the boundary) and the non-boundary area, and then applies a weight of 2:1 for these two areas. Learning rate was set to 1×10^{-4} and weight decay was set to 1×10^{-5}. The max epoch was set to 200, and the train can be early stopped based on the accuracy of validation set. No seed was used in this experiment.

Our experiments were performed on Pytorch version 1.7.1. AdamW is the optimizer used in the two stages. Referring to the transfer learning, the initial weight of encoder, ConvNeXt-XL, was obtained on ImageNet [22] and then the whole net was trained on these two datasets.

C. Comparison With SOTA Methods

ConvBNet was compared with four SOTA methods, including Deeplab v3+, HRNet, and Swin Transformer [33]. These segmentation models have achieved outstanding performance on different tasks. The Deeplab v3+ with a ResNet-101 backbone was used in the experiments. Among the HRNets, HRNet-W48 was selected. We chose the SwinV2-L for comparison and added a decoder for building extraction.

The prediction results of public dataset based on trained models are shown in Fig. 3. Three representative scenes, high-density small buildings, large buildings with unified roofs, and large buildings with various roofs, are presented. For all models, buildings are basically identified from other objects, however, obvious differences also exist. In the first scene, the color and texture of the roof vary greatly. The boundary of Deeplab v3+ is too smooth, and though the HRNet significantly improved it, some speckles appear. SwinV2-L identifies a large piece of ground as a building. Our model achieved

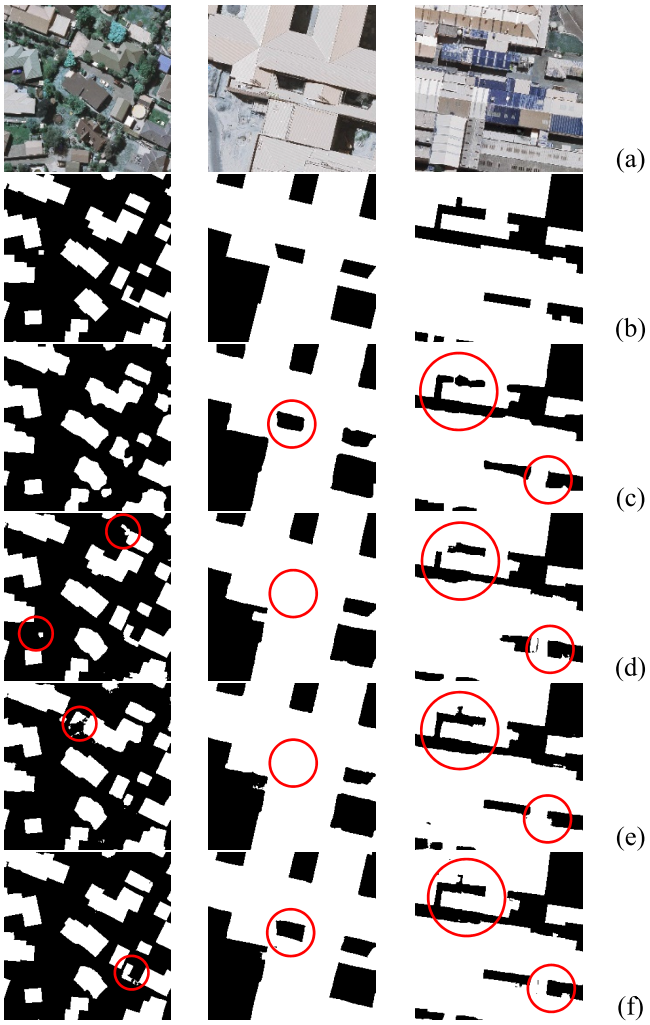

Fig. 3. Examples of building extraction results of models on the WHU dataset. (a) Image. (b) Ground truth. (c) Deeplab v3+. (d) HRNet. (e) SwinV2-L. (f) ConvBNet.

the best performance with the least misclassification and the most regular shapes. In the second scene, the cavity in the middle is the most obvious difference between the methods. HRNet and SwinV2-L fail to extract it. Although Deeplab v3+ has identified the cavity, the boundary is still curved. ConvBNet does well in both the outcomes of cavity and boundary, showing the highest similarity to the ground truth. In the third scene, the performance differences are mainly reflected in the narrow objects. As shown in the larger circle, Deeplab v3+ and HRNet got the approximate position while the discontinuous shape. SwinV2-L works better, but the small protrusion in the smaller circle is missing. Overall, our model is best and has relatively complete shape.

The prediction results of Zhejiang dataset are shown in Fig. 4. Three scenes, including large plants, rural agglomeration area, and urban residential area, are presented. In the first scene, the texture of large plants is uniform. Like the results in WHU dataset, Deeplab v3+ has smooth boundaries and obvious misclassification. As for the unlabeled buildings, shown in the upper circle, HRNet, SwinV2-L, and ConvBNet can partially extract that. But the HRNet does not distinguish

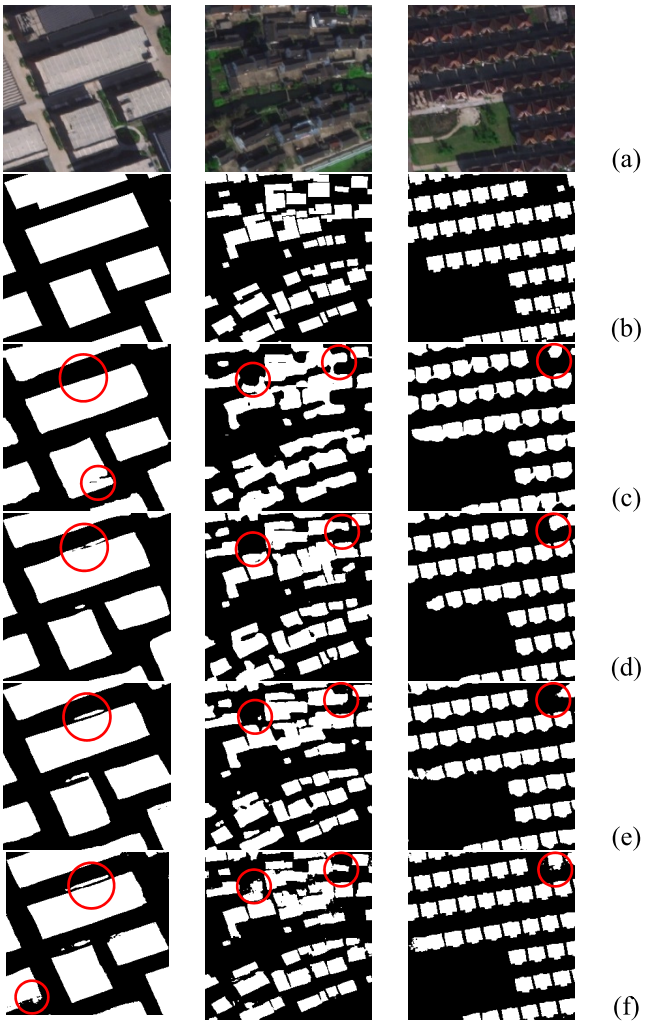

Fig. 4. Examples of building extraction results of models on the Zhejiang dataset. (a) Image. (b) Ground truth. (c) Deeplab v3+. (d) HRNet. (e) SwinV2-L. (f) ConvBNet.

the large and narrow buildings, causing them to be connected. SwinV2-L has the largest missing area of narrow buildings, while ConvBNet also suffers from the same missing problem in the lower left corner. In the second scene, the high density and strong aggregation of rural houses result in borders being covered by shadows and vegetation. Building adhesion occurs in every method, but ours has the most accurate and abundant details. In the third scene, buildings with uniform shapes and colors are relatively easy to extract, and all models have achieved good results. ConvBNet accurately depicts the protrusion of the buildings, which are not regular rectangles, and correctly extracts the unlabeled building, as shown in circle.

Four metrics were calculated on two datasets to quantitatively evaluate above methods, as shown in Table I. The intersection over Union (IoU) is the ratio of true positive pixels and all pixels without true negative pixels. Overall accuracy (OA) presents the proportion of correctly classified pixels in all pixels. ConvBNet achieved the highest IoU and OA for the WHU dataset, reaching 91.22% and 98.98%. Compared with others, ConvBNet increased the IoU by 0.88%–2.70%.

TABLE I
EVALUATION METRICS ON TEST DATASETS

Dataset	Model	IoU	OA	Precision	Recall
WHU	DeeplabV3+	88.52%	98.63%	92.92%	94.93%
	HRNet-W48	90.30%	98.85%	93.68%	96.17%
	SwinV2-L	90.34%	98.85%	93.61%	**96.28%**
	ConvBNet	**91.22%**	**98.98%**	95.17%	95.65%
	SiU-Net	88.40%	-	93.80%	93.90%
	BOMSC-Net	90.15%	98.20%	95.14%	94.50%
	DS-Net	90.40%	-	94.85%	95.06%
	CBRNet	90.97%	98.95%	**95.28%**	95.26%
Zhejiang	DeeplabV3+	76.24%	97.02%	86.16%	86.89%
	HRNet-W48	76.78%	97.10%	86.62%	87.12%
	SwinV2-L	76.85%	97.10%	86.30%	**87.52%**
	ConvBNet	**77.90%**	**97.27%**	**87.84%**	87.31%

TABLE II
ABLATION EXPERIMENTAL RESULTS

Dataset	Blocks	IoU	OA	Precision	Recall
WHU	-	90.82%	98.92%	94.24%	**96.16%**
	DS	90.88%	98.93%	94.49%	95.97%
	DS+CA	91.14%	98.96%	94.91%	95.82%
	DS+CA+ MASK	**91.22%**	**98.98%**	**95.17%**	95.65%
Zhejiang	-	77.51%	97.22%	87.66%	87.00%
	DS	77.69%	97.23%	87.34%	**87.55%**
	DS+CA	77.84%	97.27%	**87.93%**	87.15%
	DS+CA+ MASK	**77.90%**	97.27%	87.84%	87.31%

In addition to the above evaluation, we cited the latest building extraction studies, siamese U-Net (SiU-Net) [32], deep-supervision convolutional neural network (DS-Net) [34], boundary optimization and multi-scale context convolutional neural network (BOMSC-Net) [35], and coarse-to-fine boundary refinement network (CBRNet) [36], for further comparison, and the CBRNet was also trained on the WHU dataset. ConvBNet outperformed the four methods with an increase of 0.25%–2.82% in IoU. For the Zhejiang dataset, ConvBNet also achieved the highest IoU, OA, Precision, reaching 77.90%, 97.27%, and 87.84%. Compared with others, it increased the IoU by 1.05%–1.66%.

D. Ablation Experiments

ConvNeXt is a scene classification network, and we modified it to support semantic segmentation (pixel-level classification). To verify the effectiveness of each improved module, ablation experiments were conducted on both datasets, and contribution of each improvement to the results was quantitatively presented. As shown in Table II, we tested the deep supervision module, CA module, and MCE loss functions. IoU was taken as the main analysis indicator among the four metrics. The deep supervision module contributed an increase of 0.06% and 0.18%, respectively, indicating that adding weaker supervision to the middle layer can improve the target extraction ability. The multistage outputs of the encoder were optimized by the CA module and the ASPP module, which brought an improvement of 0.26% and 0.15%. Refining the key information of features in the intermediate stages is helpful for building extraction. At last, the MCE loss enhanced the distinction of the boundary area, contributing an improvement of 0.08% and 0.06%.

IV. Conclusion

In this letter, we proposed ConvBNet for building extraction based on the ConvNeXt, which outperforms the transformers. The CA and ASPP modules optimize the four-stage outputs of the encoder and input them into the decoder, which refers to the design of UNet and ConvNeXt. A deep supervision module is added so the ground truth can directly adjust the features of the middle layers. The multistage and multiloss training strategy alleviates class imbalance and improves the accuracy of boundaries. The analysis results on the public dataset and the private dataset demonstrate the excellent performance of our network, which is also highly competitive with recent research. Our network structure and training strategy are proved to be suited for building extraction tasks. Each of our designs is effective and contributes to improvement.

Although our algorithm has achieved high accuracy, there are still some areas need to be improved, such as the high memory usage and the difficulty of training. In addition, pretraining methods and transferability are not tested in this letter. In the future, we will conduct research on transferability and regularization to improve the practical value of the algorithm. ConvBNet is a general semantic segmentation network, and we will also expand its usability in other research areas.

References

[1] S. Cui, Q. Yan, and P. Reinartz, "Complex building description and extraction based on Hough transformation and cycle detection," *Remote Sens. Lett.*, vol. 3, no. 2, pp. 151–159, Mar. 2012.

[2] S. Cui, Q. Yan, and Z. Liu, "Right-angle building extraction based on graph-search algorithm," in *Proc. Int. Workshop Earth Observ. Remote Sens. Appl.*, Beijing, China, Jun. 2008, pp. 1–7.

[3] X. Huang and L. Zhang, "Morphological building/shadow index fobuilding extraction from high-resolution imagery over urban areas," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 161–172, Feb. 2012.

[4] X. Wang, F. Duan, X. Qu, D. Li, and P. Yu, "Building extraction based on UAV imagery data with the synergistic use of objected-based method and SVM classifier," *Remote Sens. Land Resour.*, vol. 29, no. 1, pp. 97–103, 2017.

[5] D. Chen, S. Shang, and C. Wu, "Shadow-based building detection and segmentation in high-resolution remote sensing image," *J. Multimedia*, vol. 9, no. 1, pp. 181–188, 2014.

[6] T. Hermosilla, L. A. Ruiz, J. A. Recio, and J. Estornell, "Evaluation of automatic building detection approaches combining high resolution images and LiDAR data," *Remote Sens.*, vol. 3, no. 6, pp. 1188–1210, 2011.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 3431–3440.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany, Oct. 2015, pp. 234–241.

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

[10] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5693–5703.

[11] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, p. 144, Jan. 2018.

[12] M. Chen et al., "DR-Net: An improved network for building extraction from high resolution remote sensing image," *Remote Sens.*, vol. 13, no. 2, p. 294, 2021.

[13] Y. Yu et al., "Building extraction from remote sensing imagery with a high-resolution capsule network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[14] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2021.

[15] Q. Tian, Y. Zhao, Y. Li, J. Chen, X. Chen, and K. Qin, "Multiscale building extraction with refined attention pyramid networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[16] W. Deng, Q. Shi, and J. Li, "Attention-gate-based encoder–decoder network for automatical building extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2611–2620, 2021.

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

[18] 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.

[19] 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.

[20] W. Yuan and W. Xu, "MSST-Net: A multi-scale adaptive network for building extraction from remote sensing images based on swin transformer," *Remote Sens.*, vol. 13, no. 23, p. 4743, Nov. 2021.

[21] X. Chen, C. Qiu, W. Guo, A. Yu, X. Tong, and M. Schmitt, "Multiscale feature learning by transformer for building extraction from satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[22] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," 2022, *arXiv:2201.03545*.

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

[24] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[25] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13713–13722.

[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[27] P. Wang et al., "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.

[28] C. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, San Diego, CA, USA, May 2015, pp. 562–570.

[29] S. Ran, X. Gao, Y. Yang, S. Li, G. Zhang, and P. Wang, "Building multi-feature fusion refined network for building extraction from high-resolution remote sensing images," *Remote Sens.*, vol. 13, no. 14, p. 2794, Jul. 2021.

[30] H. Jung, H.-S. Choi, and M. Kang, "Boundary enhancement semantic segmentation for building extraction from remote sensed image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5215512.

[31] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1251–1258.

[32] 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.

[33] Z. Liu et al., "Swin transformer v2: Scaling up capacity and resolution," 2021, *arXiv:2111.09883*.

[34] H. Guo, X. Su, S. Tang, B. Du, and L. Zhang, "Scale-robust deep-supervision network for mapping building footprints from high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10091–10100, 2021.

[35] 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. 21645395.

[36] 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.