Building extraction with vision transformer

Libo Wang, Shenghui Fang, Rui Li and Xiaoliang Meng

Abstract—As an important carrier of human productive activities, the extraction of buildings is not only essential for urban dynamic monitoring but also necessary for suburban construction inspection. Nowadays, accurate building extraction from remote sensing images remains a challenge due to the complex background and diverse appearances of buildings. The convolutional neural network (CNN) based building extraction methods, although increased the accuracy significantly, are criticized for their inability for modelling global dependencies. Thus, this paper applies the Vision Transformer for building extraction. However, the actual utilization of the Vision Transformer often comes with two limitations. First, the Vision Transformer requires more GPU memory and computational costs compared to CNNs. This limitation is further magnified when encountering large-sized inputs like fine-resolution remote sensing images. Second, spatial details are not sufficiently preserved during the feature extraction of the Vision Transformer, resulting in the inability for fine-grained building segmentation. To handle these issues, we propose a novel Vision Transformer (BuildFormer), with a dual-path structure. Specifically, we design a spatial-detailed context path to encode rich spatial details and a global context path to capture global dependencies. Besides, we develop a window-based linear multi-head self-attention to make the complexity of the multi-head self-attention linear with the window size, which strengthens the global context extraction by using large windows and greatly improves the potential of the Vision Transformer in processing large-sized remote sensing images. The proposed method yields state-of-the-art performance (75.74% IoU) on the Massachusetts building dataset. Code will be available.

Index Terms—Vision Transformer, building extraction, remote sensing, attention mechanism.

I. INTRODUCTION

Building extraction using fine-resolution remote sensing images, i.e., the task of identifying building and non-building pixels in an image [1], plays a crucial role in a wide range of application scenarios such as urban planning, population statistic, economic assessment and disaster management [2-6].

Conventional methods for building extraction commonly extract hand-craft features (e.g., spectral, spatial, textural) and apply traditional machine learning methods (e.g., Support Vector Machine and Random Forest) to recognize buildings [7-9]. However, the empirically designed hand-craft features restrict the generalization ability of these traditional methods.

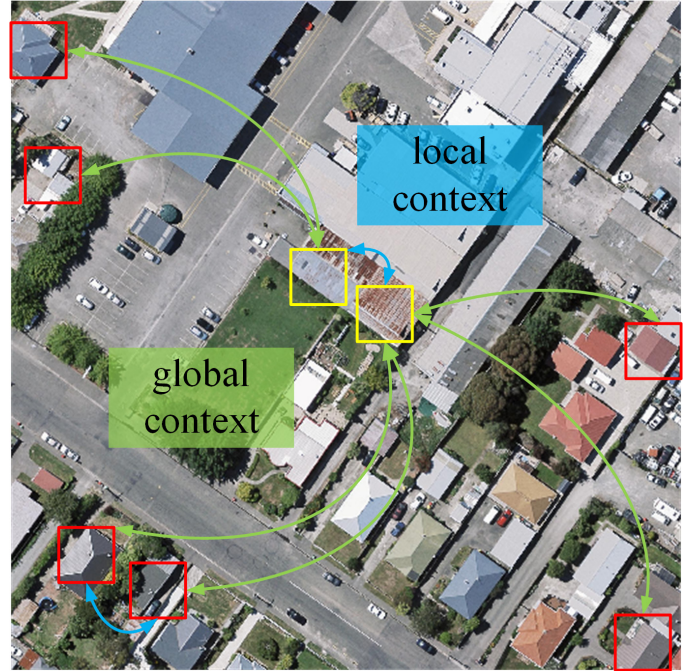In the past few years, Deep Learning (DL) has become a

Fig. 1. Illustration of the global context and local context. The squares represent the receptive view of the convolution. The yellow regions represent the blurry building pixels where the local context is indistinguishable.

popular approach for automatic feature learning [10] and achieved great breakthroughs in the computer vision (CV) domain [11]. In the field of remote sensing, DL methods, especially the convolutional neural network (CNN) [12], have been introduced and implemented in many geospatial tasks [13-15], especially for building extraction [16]. In comparison with conventional methods, CNN-based methods can capture various kinds of information including textures, spectrums, spatial context, and the interactions among geo-objects.

Since the pioneer CNN structure, i.e., Fully Convolutional Neural Network (FCN), was proposed for pixel-level dense prediction, a series of researches were carried out on automatic building extraction from remote sensing images [17-20]. Subsequently, the encoder-decoder structure was proposed to address the coarse-resolution segmentation of FCN-based networks by constructing a symmetrical decoder. Typical methods like UNet and SegNet restored the spatial resolution of

extracted features progressively for fine-resolution feature representation [21, 22]. The results of these CNN-based methods, although encouraging, encounter bottlenecks in building extraction. To be specific, the CNN is designed to extract the local context and thus lacks the ability to model global context in its nature. However, the local context is often ambiguous for identifying building pixels, while the extraction will become much simpler if the global context from the whole remote sensing image is available, as illustrated in Fig.1.

For capturing the global context, the most popular way is to incorporate attention mechanisms into networks. For example, the non-local module [23], the dual attention module [24], the criss-cross attention block [25] and the object context block [26], obtained great improvements in semantic segmentation thanks to their ability in modelling global dependencies by attention mechanisms. In the field of building extraction, several attempts were made to introduce attention mechanisms for stronger feature representation, which differentiates heterogeneous buildings from complex backgrounds in fine-resolution remote sensing images [6, 27]. However, these methods still follow the CNN structure, restricting the global feature representation.

Recently, the Transformer [28], originally designed for natural language processing (NLP) tasks, comprises a hot topic in the computer vision domain, namely Vision Transformer (ViT) [29]. Different from the CNN structure, the ViT translates 2D image-based tasks into 1D sequence-based tasks. Due to the strong sequence-to-sequence modelling ability, the ViT demonstrates superior characterization of extracting global context than attention-based CNNs, obtaining numerous breakthroughs on fundamental vision tasks, such as image classification [29] and object detection [30] as well as semantic segmentation [31].

However, the actual utilization of ViTs often comes with huge memory requirements and computational costs [32, 33], which seriously affects its potential for downstream tasks like building extraction. Even though the Swin Transformer adopts the hierarchical structure and designs a window-based multi-head self-attention mechanism to improve efficiency, its complexity still increases quadratically along with the increasing size of the window [34]. Furthermore, ViTs mainly focus on capturing the global context while ignoring preserving the spatial-detailed context, but spatial details are also essential for fine-grained building segmentation in fine-resolution remote sensing images [35].
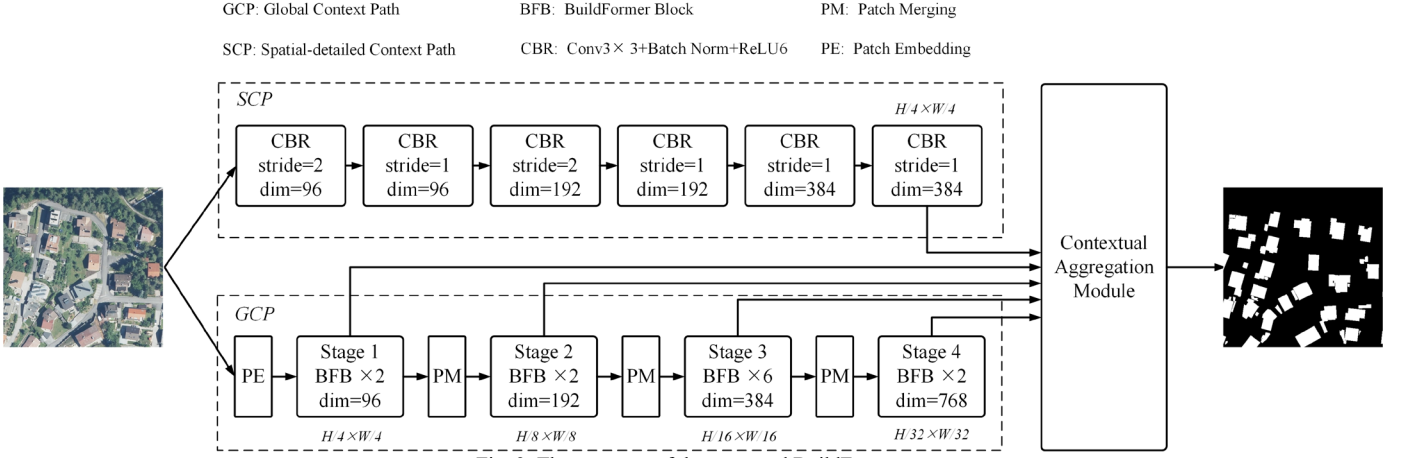
In this paper, we propose a novel Vision Transformer, namely BuildFormer, for building extraction from fine-resolution remote sensing images to address the existing issues of ViTs. Specifically, we adopt a dual-path structure to construct the BuildFormer, i.e. a global context path and a spatial-detailed context path. In the global context path, we develop a novel Transformer block to construct a Vision Transformer backbone, enhancing the ability for global context extraction. In the spatial-detailed context path, we utilize stacked convolutional layers to preserve rich spatial details. The major contributions of this paper are as follows:

1) We propose a novel Vision Transformer (BuildFormer) based on the dual-path structure, which can capture the global context while preserving spatial-detailed features.
2) We present a novel Transformer block to construct the global context path, namely BuildFormer Block (BFB), which is mainly composed of a window-based linear multi-head self-attention (W-LMHSA) and a convolutional multilayer perceptron (C-MLP).
3) The W-LMHSA reduces the complexity of the window-based multi-head self-attention (W-MHSA) [34] to linear complexity. Benefiting from this, the BuildFormer can apply larger windows to extract global features from large inputs without resulting in high computations, which is more suitable for large-scale fine-resolution remote sensing images. The C-MLP strengthens the cross-window interactions, which further enhances the ability of the BuildFormer for global information modelling.

II. RELATED WORK

A. CNN-based Building Extraction Methods

With the rapid development of Deep Learning, the convolutional neural network (CNN) has become the mainstream method for the automatic remote sensing building extraction task. In comparison with the conventional methods that design hand-crafted feature operators (colour, texture, shallow, etc.) [37-42] or those using active remote sensing data (LiDAR and SAR) [5, 43-46], the CNN-based methods have advantages in hierarchical feature extraction and efficiency [9, 47-50]. Although the CNN-based methods achieve many breakthroughs, their weaknesses in global information modelling limit further improvements in accuracy, as global information is crucial for detecting buildings from low-interclass and high-intraclass remote sensing images [51-53]. To address it, several studies have introduced attention mechanisms to strengthen the global feature representation for building extraction [54-57]. For example, Deng et al. [58] developed a grid-based attention gate module to extract semantic features with a global receptive field, further boosting the accuracy. Guo et al. [6] introduced the parallel attention to capturing global scene information, which further improved the accuracy of building segmentation. Pan et al. [59] combined spatial and channel attention mechanisms into the generative adversarial network and achieved advanced results. Cai et al [60] proposed a multipath hybrid attention network to enhance the performance of extracting small buildings. Since these attention-based methods relied too much on convolution operations, they failed to liberate the network from the CNN structure and have certain limitations in global information modelling.

GCP: Global Context Path   BFB: BuildFormer Block   PM: Patch Merging

SCP: Spatial-detailed Context Path   CBR: Conv3×3+Batch Norm+ReLU6   PE: Patch Embedding

Fig. 2. The structure of the proposed BuildFormer.

B. ViT-based Building Extraction Methods

ViT-based methods have brought tremendous progress and evolution for semantic segmentation [31, 61, 62]. The structure of the ViT is completely different from the CNN, which treats the 2D image as the 1D ordered sequence and applies the self-attention mechanism for global dependency modelling, demonstrating stronger global feature extraction. Driven by this, many researchers in the field of remote sensing introduced ViTs for segmentation-related tasks, such as land cover classification [63-68], urban scene parsing [69-74], change detection [75, 76], road extraction [77] and especially building extraction [78]. For example, Chen et al. [79] proposed a sparse token Transformer to learn the global dependency of tokens in both spatial and channel dimensions, achieving state-of-the-art accuracy on benchmark building extraction datasets. Yuan et al [80] introduced the widely used Swin Transformer [34] as the encoder and design a scale-adaptive decoder for multi-scale feature representation. Compared with the CNN-based methods, the global information is fully extracted by ViT-based methods. However, the spatial detailed context, meanwhile, is ignored.

III. METHODOLOGY

A. Overview

The structure of the proposed BuildFormer is illustrated in Fig. 2 with a Global Context Path (GCP) and a Spatial-detailed Context Path (SCP). In GCP, four BuildFormer Blocks are designed to extract four global feature maps at different scales. Meanwhile, the high-resolution spatial-detailed feature map will be generated by SCP. Finally, the four global feature maps and the spatial-detailed feature map are fed into the contextual aggregation module to generate the final semantic feature.

B. Spatial-detailed Context Path

It is very challenging to reconcile the demand for spatial-detailed features with global dependencies simultaneously in the Vision Transformer. However, both of them are essential for obtaining high accuracy of building segmentation. To address this issue, in the proposed BuildFormer, we adopt a dual-path structure [36], which introduces a spatial-detailed context path to produce a high-resolution feature map for preserving spatial details. Concretely, we apply six (Convolution-BatchNorm-ReLU6) CBR blocks to construct this path and expand their channel dimensions progressively to encode sufficient spatial-detailed information, as shown in Fig. 2. Specifically, six standard 3×3 convolutional layers are employed and each layer is equipped with a batch normalization operation and a ReLU6 activation function. To ensure sufficient spatial details, the size of the output feature map is designed as 1/4 of the original input image.

C. Global Context Path

The global context path is a novel self-designed Vision Transformer. The main basic modules of this path include the BuildFormer Block, Patch Embedding, and Patch Merging, as shown in Fig. 2. Due to its linear complexity, this path is more suitable for capturing global context from large-scale remote sensing images.

Fig. 3. (a) the Patch Embedding module, (b) the Patch Merging module.

Patch Embedding: The original ViT [29] utilizes linear projections to split the input image into non-overlapping patches directly. However, this scheme has limitations in modelling the structure information within patches. To overcome it, we apply convolutional layers to split the input image into overlapping patches. As shown in Fig. 3 (a), we use two 3×3 convolutional layers with a stride of 2 and a padding value of 1, while each layer is followed by a batch

normalization operation and a ReLU6 activation function. Proceed by the two convolutional layers, the channel dimension of patches is expanded to 96 and the resolution is reduced to 1/4. In addition, a standard 3×3 depth-wise convolution and a residual connection are employed to enhance the relative location priors of patches.

Patch Merging: To obtain the hierarchical feature representation, four Patch Merging modules are employed and each module reduces the resolution of intermediate patches and expands the channel dimension. As shown in Fig. 3. (b), we first use the batch normalization operation to normalize the patches then apply a 2×2 convolutional layer to down-sampling it to 1/2 and expand its channel dimension to 2 times. Similar to the Patch Embedding module, we utilize a standard 3×3 depth-wise convolution and a residual connection to strengthen the location information extraction.

BuildFormer Block: Each BuildFormer Block is composed of a Window-based Linear Multi-Head Self-Attention module (W-LMHSA), a convolutional multilayer perceptron, two batch normalization operations and two residual connections, as illustrated in Fig. 4.
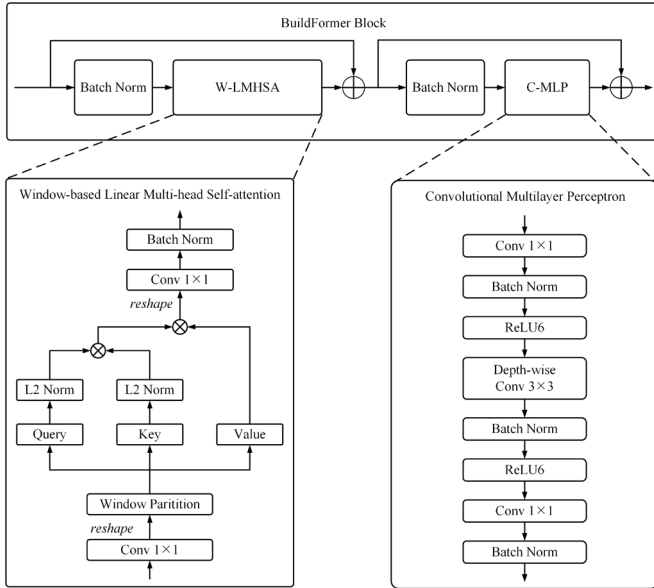

Fig. 4. Details of the BuildFormer Block.

In Swin Transformer [34], the Window-based Multi-Head Self-Attention (W-MHSA) splits the input into non-overlapping windows and performs the standard Multi-Head Self-Attention (MHSA) [28] in each local window. Benefiting from the window partition operation, the W-MHSA saves much computational burden compared to the MHSA. Even though, the computational complexity of each local window is still $O(N^2)$ due to the application of the MHSA. N is the square of the window size. Thus, the W-MHSA comes with huge computations and memory requirements if using large windows.

By contrast, the proposed W-LMHSA further eliminate the high demand of W-MHSA in computations and memory based on our previous work on the linear attention mechanism [52], which makes the computational complexity linear with the window size. For each local window, the multi-head self-attention can be defined as:

$$\text{MHSA}(\boldsymbol{X}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)\boldsymbol{W_o} \tag{1}$$

Here, $\boldsymbol{X}$ is the input vector and h is the number of heads. $\boldsymbol{W_o} \in \mathbb{R}^{N \times D}$ is a projected matrix, where D is the dimension of the input vector. Each head denotes a self-attention operation which can be defined as:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}_{\text{row}}\left(\frac{\boldsymbol{QK^T}}{s}\right)\boldsymbol{V} \tag{2}$$

$$\boldsymbol{Q} = \boldsymbol{X_m W_q} \in \mathbb{R}^{N \times d} \tag{3}$$

$$\boldsymbol{K} = \boldsymbol{X_m W_k} \in \mathbb{R}^{N \times d} \tag{4}$$

$$\boldsymbol{V} = \boldsymbol{X_m W_v} \in \mathbb{R}^{N \times d} \tag{5}$$

where $\boldsymbol{X_m}$ is the input vector of the m-th head. $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ are the *query* feature, *key* feature and *value* feature, which are generated by the three projected matrixs $\boldsymbol{W_q}$, $\boldsymbol{W_k}$ and $\boldsymbol{W_v}$, respectively. d denotes the dimension of the m-th head and $d = D/h$. s represents the scale factor and s is set to 1 by default. $\text{Softmax}_{\text{row}}(\boldsymbol{QK^T})$ computes the similarities between each pair of pixels of the input vector and applies the softmax normalization function along each row of the similarity matrix $\boldsymbol{QK^T}$, which is the key step to model global dependencies. However, the product between $\boldsymbol{Q} \in \mathbb{R}^{N \times d}$ and $\boldsymbol{K^T} \in \mathbb{R}^{d \times N}$ belongs to $\mathbb{R}^{N \times N}$, which leads to the $O(N^2)$ computational costs and memory requirements. As N is the square of the window size, the resource-demanding of the W-MHSA can increase significantly when using large windows. To address this, we simplify Eq. (2) by replacing the softmax normalization function with the first-order approximation of the Taylor expansion. Specifically, when using the softmax normalization function, the i-th row of the result matrix generated by Eq. (2) can be written as:

$$\text{Attention}_i(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \frac{\sum_{j=1}^{N} e^{\boldsymbol{q}_i^T \boldsymbol{k}_j} \boldsymbol{v}_j}{\sum_{j=1}^{N} e^{\boldsymbol{q}_i^T \boldsymbol{k}_j}} \tag{6}$$

Here, $\boldsymbol{q}_i^T \in \mathbb{R}^d$ is the i-th *query* feature. $\boldsymbol{k}_j$ and $\boldsymbol{v}_j$ are the j-th *key* feature and *value* feature, respectively. Please note that the vectors in this research are column vectors by default. Actually, Eq. (6) can be generalized to any normalization function as:

$$\text{Attention}_i(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \frac{\sum_{j=1}^{N} sim(\boldsymbol{q}_i, \boldsymbol{k}_j) \boldsymbol{v}_j}{\sum_{j=1}^{N} sim(\boldsymbol{q}_i, \boldsymbol{k}_j)} \tag{7}$$

$$sim(\boldsymbol{q}_i, \boldsymbol{k}_j) = \phi(\boldsymbol{q}_i)^T \varphi(\boldsymbol{k}_j) \tag{8}$$

$sim(\boldsymbol{q}_i, \boldsymbol{k}_j)$ can measure the similarity between $\boldsymbol{q}_i$ and $\boldsymbol{k}_j$. The normalization functions $\phi(\cdot)$ and $\varphi(\cdot)$ are used to ensure $sim(\boldsymbol{q}_i, \boldsymbol{k}_j) \geq 0$. According to the first-order approximation of the Taylor expansion:

$$e^{\boldsymbol{q}_i^T \boldsymbol{k}_j} \approx 1 + \boldsymbol{q}_i^T \boldsymbol{k}_j \tag{9}$$

We set $\phi(\cdot)$ and $\varphi(\cdot)$ as the L2 normalization function to guarantee $\boldsymbol{q}_i^T \boldsymbol{k}_j \geq -1$:

$$sim(\boldsymbol{q}_i, \boldsymbol{k}_j) = 1 + \left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2}\right)^T \left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right) \tag{10}$$

Thus, Eq. (6) can be rewritten as Eq. (11), simplified as Eq. (12) and further turned into Eq. (13):

$$\text{Attention}_i(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \frac{\sum_{j=1}^{N}\left(1+\left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2}\right)^T\left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right)\right)\boldsymbol{v}_j}{\sum_{j=1}^{N}\left(1+\left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2}\right)^T\left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right)\right)} \tag{11}$$

$$\text{Attention}_i(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \frac{\sum_{j=1}^{N}\boldsymbol{v}_j+\left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2}\right)^T\sum_{j=1}^{N}\left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right)\boldsymbol{v}_j^T}{N+\left(\frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2}\right)^T\sum_{j=1}^{N}\left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right)} \tag{12}$$

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \frac{\sum_j \boldsymbol{V}_{i,j}+\left(\frac{\boldsymbol{Q}}{\|\boldsymbol{Q}\|_2}\right)\left(\left(\frac{\boldsymbol{K}}{\|\boldsymbol{K}\|_2}\right)^T\boldsymbol{V}\right)}{N+\left(\frac{\boldsymbol{Q}}{\|\boldsymbol{Q}\|_2}\right)\sum_j\left(\frac{\boldsymbol{K}}{\|\boldsymbol{K}\|_2}\right)^T_{i,j}} \tag{13}$$

Since $\sum_{j=1}^{N}\left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right)\boldsymbol{v}_j^T$ and $\sum_{j=1}^{N}\left(\frac{\boldsymbol{k}_j}{\|\boldsymbol{k}_j\|_2}\right)$ can be calculated and reused for each *query*, the time and memory complexity of the proposed attention based on Eq. (13) is the $O(dN)$ linear complexity.

The cross-window interaction is crucial for global dependencies modelling when using the W-MHSA. The Swin Transformer [34] introduces a shifted window operation to strengthen the cross-window interaction. This scheme, although very effective, increases the complexity of the network due to adding another shifted-window Transformer block. In this paper, we provide a convolutional multilayer perceptron (C-MLP) to strengthen the interaction within windows. In comparison with the Swin Transformer, the employment of the C-MLP can maintain competitive accuracy while improving efficiency. The detailed components of the C-MLP are illustrated in Fig. 4.
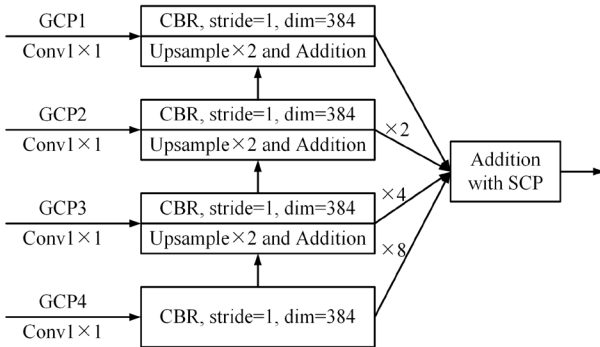
D. Context Aggregation Module

Fig. 5. Details of the Context Aggregation Module.

The output features from GCP and SCP are complementary. The feature from the SCP mainly encodes rich detailed information, while the four features generated by the GCP provide high-level global semantic information. To better fuse them, we adopt the feature fusion strategy like the feature pyramid feature (FPN) [81], as shown in Fig. 5. Specifically, the four global feature maps from the GCP are first proceeded by four 1×1 convolution layers to unify the channel dimension to 384. Then, we apply four CBR blocks as well as upsampling and addition operations to perform multi-level feature fusion. Finally, the fused global feature is further aggregated with the

spatial-detailed feature from the SCP to generate the final fused feature.

E. Loss Function

Improving the accuracy of building boundaries is vital for high-precision building extraction [35, 82-84]. Thus, we introduce the boundary supervision technology and adopt a joint loss to train the BuildFormer. The joint loss function L can be defined as:

$$L = L_{ce}(Y,\hat{Y}) + L_{dice}(Y,\hat{Y}) + L_{bce}\big(\mathcal{L}(Y),\mathcal{L}(\hat{Y})\big) \tag{14}$$

where Y and $\hat{Y}$ denote the predicted label and the true label, respectively. L_{ce} is the cross-entropy loss. L_{dice} is the dice loss. $\mathcal{L}$ represents the Laplacian convolution [85] with a kernel of $\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$ that extracts the building boundaries of the predicted label and the true label. The binary cross-entropy loss (denoted by L_{bce}) is employed on the extracted building boundaries.

IV. EXPERIMENTAL SETTINGS AND DATASETS

A. Datasets

To evaluate the performance of the proposed BuildFormer, three publicly available building datasets are considered comprehensively for conducting experiments, including the Massachusetts building dataset, WHU building dataset and Inria Aerial Image Labeling dataset. The details are as follows.

1) Massachusetts: The Massachusetts building dataset is composed of 151 aerial images of the Boston area with a size of 1500×1500 pixels and a ground sampling distance of 1 m. The dataset involves urban and suburban scenes, where the buildings are varied in sizes, shapes, textures and colours. Thus, this dataset is very challenging and suitable to verify the effectiveness of modules. We follow the official partition provided by the dataset and use data augmentation technologies like vertical and horizontal flip to further expand the training set. As a result, we use 411 images for training, 4 images for validation, and 10 images for testing. In the training phase, we randomly crop the images and labels into 1024×1024 pixels as the input. In the validation and testing phase, the images and labels are padded to a size of 1536×1536 pixels to ensure it is divisible by 32 (the downsampling factor of the BuildFormer). The padded parts are ignored when computing evaluation metrics.

2) WHU: The WHU building dataset [18] includes two types of images, i.e. satellite imagery and aerial imagery. We only use aerial images in our experiments. The aerial imagery subset covers over 450 km² and includes 22000 buildings. The spatial resolution of the RGB aerial images is 0.3 m and the size of each image is 512×512 pixels. There are 8189 image tiles in this dataset, where 4736 tiles for training, 1036 tiles for validation and 2416 tiles for testing. We follow the official partition in our experiments.

3) Inria: The Inria Aerial Image Labeling Dataset [86] contains 360 fine-resolution aerial images collected from five cities (Austin, Chicago, Kitsap, Tyrol and Vienna). Since the labels of the test set are publicly available, we only use the original training set in our experiments. Suggested by the

official partition, the 1 to 5 tiles of each city are selected for validation and the rest for training. We first pad the original 5000×5000 images to 5120×5120 pixels, then crop them into 512×512 pixels image tiles. The image tiles, which do not contain buildings, are removed for efficient training. As a result, 9737 and 1942 image tiles are used for training and validation, respectively.

B. Evaluation Metrics

We use the intersection over union (IoU), F1 score, precision and recall to evaluate the performance of models. These metrics are widely used in the field of building extraction [27, 35], which can be defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{17}$$

$$\text{IoU} = \frac{TP}{TP + FN + FP} \tag{18}$$

TP, FP, and FN represent the true positive, the false positive, and the false negative, respectively.

C. Experimental Setting

All models in the experiments were implemented with the PyTorch framework on a single NVIDIA GTX 3090 GPU with 24GB RAM. The AdamW optimizer and the cosine strategy were employed to train all models in the experiments. The random horizontal and vertical flipping were selected as data augmentation strategies. For the WHU building dataset, we trained the BuildFormer from scratch for 105 epochs. The base learning rate was set to 1e-3 and the batch size was set to 8. For the Massachachusets building dataset and the Inria Aerial Image Labelling dataset, we used BuildFormer's weight trained on the WHU building dataset, then fine-tuned it for 105 epochs with a learning rate of 5e-4. In the testing phase, we applied the data augmentation technologies like horizontal and vertical flipping, which is also known as test-time augmentation (TTA).

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Ablation Study

To verify the effectiveness of the proposed modules, we conducted ablation experiments on the Massachachusets building dataset.

TABLE I
THE ABLATION EXPERIMENTAL RESULTS OF SPATIAL-DETAILED CONTEXT PATH ON THE MASSACHUSETTS BUILDING DATASET.

Method	IoU	F1
BuidFormer without SCP	74.38	85.30
BuildFormer with SCP	75.74	86.19

1) The effectiveness of the spatial-detailed context path (SCP): In the proposed BuildFormer, the spatial-detailed context path aims to encode rich spatial-detailed information for fine-grained building segmentation. To test its effectiveness, we remove it from BuildFormer. As listed in Table I, the utilization of the spatial-detailed context path provides an increase of 1.36% in IoU, which demonstrates its effectiveness and necessity.

Furthermore, this result also illustrates the superiority of the dual-path structure over than single-path structure for fine-grained building extraction.

2) The superiority of the global context path (GCP): The global context path in the proposed BuildFormer is a Vision Transformer backbone. To demonstrate its superiority in building extraction, we replace it with other backbones for comparison. The results show that our method yields an improvement of 2.04% in IoU compared to the Swin-Small [34] and surpassed the classical convolutional backbone ResNet101 [87] by 5.01% in IoU (Table II).

TABLE II
THE ABLATION EXPERIMENTAL RESULTS OF THE GLOBAL CONTEXT PATH ON THE MASSACHUSETTS BUILDING DATASET.

Method	IoU	Parameter (M)
ResNet101	70.73	49.35
Swin-Small	73.70	61.54
ours	75.74	40.52

3) The effectiveness of the convolutional multilayer perceptron (C-MLP): The C-MLP aims to strengthen the cross-window interaction, improving the ability of the BuildFormer Block for capturing global context. To demonstrate its contribution to accuracy, we replace it with the standard multilayer perceptron (MLP) for ablation experiments. As illustrated in Table III, the employment of the C-MLP increases the IoU metric and the F1 score by 5.16% and 3.36%, respectively, demonstrating its effectiveness and essential.

TABLE III
THE ABLATION EXPERIMENTAL RESULTS OF SPATIAL-DETAILED CONTEXT PATH ON THE MASSACHUSETTS BUILDING DATASET.

Method	IoU	F1
BuidFormer with MLP	70.58	82.75
BuildFormer with C-MLP	75.74	86.19

4) The advantages of the window-based linear multi-head self-attention (W-LMHSA): To better demonstrate the improvements of the proposed W-LMHSA, we conduct comprehensive experiments in comparison with the window-based multi-head self-attention (W-MHSA). We apply the W-LMHSA and W-MHSA to construct the BuildFormer, respectively. As shown in Table IV, the computational complexities of the W-MHSA and W-LMHSA under different window sizes are measured by the floating-point operation count (Flops) in M. The speed of the network (FPS) is measured by a 1024×1024 pixels image tile on a single NVIDIA GTX 3090 GPU. The results reveal that the proposed W-LMHSA has advantages in both accuracy and efficiency compared to the W-MHSA. Specifically, the proposed W-LMHSA can provide an improvement of 2% IoU while saving about 25% computational complexity. Besides, the W-LMHSA maintains the GPU memory requirement and the speed stable even with a large window, while the W-MHSA increases memory requirements and reduces the speed significantly.

TABLE IV
THE ABLATION STUDY OF THE W-LMHSA WITH DIFFERENT WINDOW SIZES. * MEANS THE NETWORK RUNS OUT OF MEMORY.

Global contextual path	Window Size	Complexity (M)	Memory (MB)	Parameter (M)	Speed (FPS)	IoU
W-MHSA	8	2.39	7477.36		16.73	72.70
	16	9.56	9301.36	40.52	14.98	73.56
	32	38.24	16597.36		10.28	*
	64	152.96	*		*	*
W-LMHSA (ours)	8	1.77	7060.89		17.03	74.83
	16	7.08	7032.99	40.52	17.17	75.74
	32	28.31	7024.14		17.18	75.59
	64	113.25	7022.16		17.04	75.36

B. Comparison of State-of-the-art Methods

To further verify the effectiveness of the proposed method, we compare it with state-of-the-art methods on three publicly available datasets, i.e. the Massachusetts building dataset, WHU building dataset and Inria Aerial Image Labeling dataset. The selected methods include convolutional networks, such as U-Net [21], Deeplabv3+ [88], SRI-Net [16], DS-Net [49], BRRNet [20], SiU-Net [18], CU-Net [19], EU-Net [89], DE-Net [90], MA-FCN [48], MANet [53], MAP-Net [27], Bias-UNet [57], CBRNet [35], and ViT-based networks like SwinUperNet [34], Sparse Token Transformer (STT) [79], MSST-Net [80], BANet [72], DC-Swin [69].

For the Massachusetts building dataset, the proposed method yields a 75.74% IoU and outperforms the recent method CBRNet by 1.19% (Table V). To our knowledge base, this score is state-of-the-art on this dataset. Notably, our method achieves the highest Recall (87.52%) and surpasses other networks by a significant gap (more than 2.33%). Higher Recall means fewer building pixels missed. As shown in Fig. 6, our approach outperforms other networks in recognizing hard building pixels and maintaining the integrity of buildings, which benefits from the dual-path structure and the aggregation of the global context and spatial-detailed context.

TABLE V
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE MASSACHUSETTS BUILDING DATASET.

Method	IoU	Precision	Recall	F1
U-Net	67.61	79.13	82.29	80.68
DeepLab V3+	69.23	84.73	79.10	81.82
MA-FCN	73.80	**87.07**	82.89	84.93
BRRNet	73.25	-	-	84.56
Bias-UNet	73.49	83.34	86.15	84.72
CBRNet	74.55	86.50	84.36	85.42
MANet	70.76	82.00	83.77	82.88
BANet	72.20	83.07	84.66	83.86
DC-Swin	72.59	83.07	85.19	84.12
BuildFormer	**75.74**	84.90	**87.52**	**86.19**

TABLE VI
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE WHU BUILDING DATASET.

Method	IoU	Precision	Recall	F1
CU-Net	87.10	94.60	91.70	93.13
SiU-Net	88.40	93.80	93.90	93.85
SRI-Net	89.23	**95.67**	93.69	94.51
DE-Net	90.12	95.00	94.60	94.08
EU-Net	90.56	94.98	95.10	95.04
MA-FCN	90.70	95.20	95.10	95.15
MAP-Net	90.86	95.62	94.81	95.21
MSST-Net	88.00	-	-	88.20
STT	90.48	-	-	94.97
BuildFormer	**91.44**	95.40	**95.65**	**95.53**

TABLE VII
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE INRIA AERIAL IMAGE LABELING DATASET.

Method	IoU	Precision	Recall	F1
U-Net	70.78	85.18	80.72	82.89
SRI-Net	76.84	-	-	86.32
DS-Net	80.73	-	-	-
BRRNet	77.05	-	-	86.61
SiU-Net	71.40	84.60	82.10	83.33
CBRNet	81.10	**89.93**	89.20	89.56
STT	79.42	-	-	87.99
SwinUperNet	79.53	87.55	89.67	88.60
BuildFormer	**81.44**	88.81	**90.75**	**89.77**

For the WHU building dataset, the proposed method yields the best IoU (91.44%), which not only exceeds the advanced CNN-based building extraction methods by more than 0.58% but also outperforms the recent Sparse Token Transformer (STT) by 0.96% (Table VI). For the Inria Aerial Image Labeling dataset, our approach still maintains the most advanced performance with 81.44% IoU and 89.77% F1 score (Table VII). The predicted results on these two datasets are shown in Fig.7. All results reveal the importance of global context for building extraction and the superiority of dual-path structure for Vision Transformer.

VI. CONCLUSION

In this paper, we proposed a novel Vision Transformer for building extraction from fine-resolution remote sensing images, namely the BuildFormer. Since both global context and spatial-detailed context were crucial for precise building segmentation, we designed the BuildFormer based on the dual-path structure which could capture the global information and spatial details simultaneously. Furthermore, we proposed a window-based linear multi-head self-attention to reduce the complexity of the window-based multi-head self-attention into $O(N)$. Benefiting from this, the BuildFormer could apply large windows to enhance the global context modelling without resulting in high computation. An extensive ablation study evaluated the impact of each component of the BuildFormer and experimental results on the Massachusetts, WHU, and Inria building datasets demonstrated the superiority of the proposed method in comparison with state-of-the-art methods.
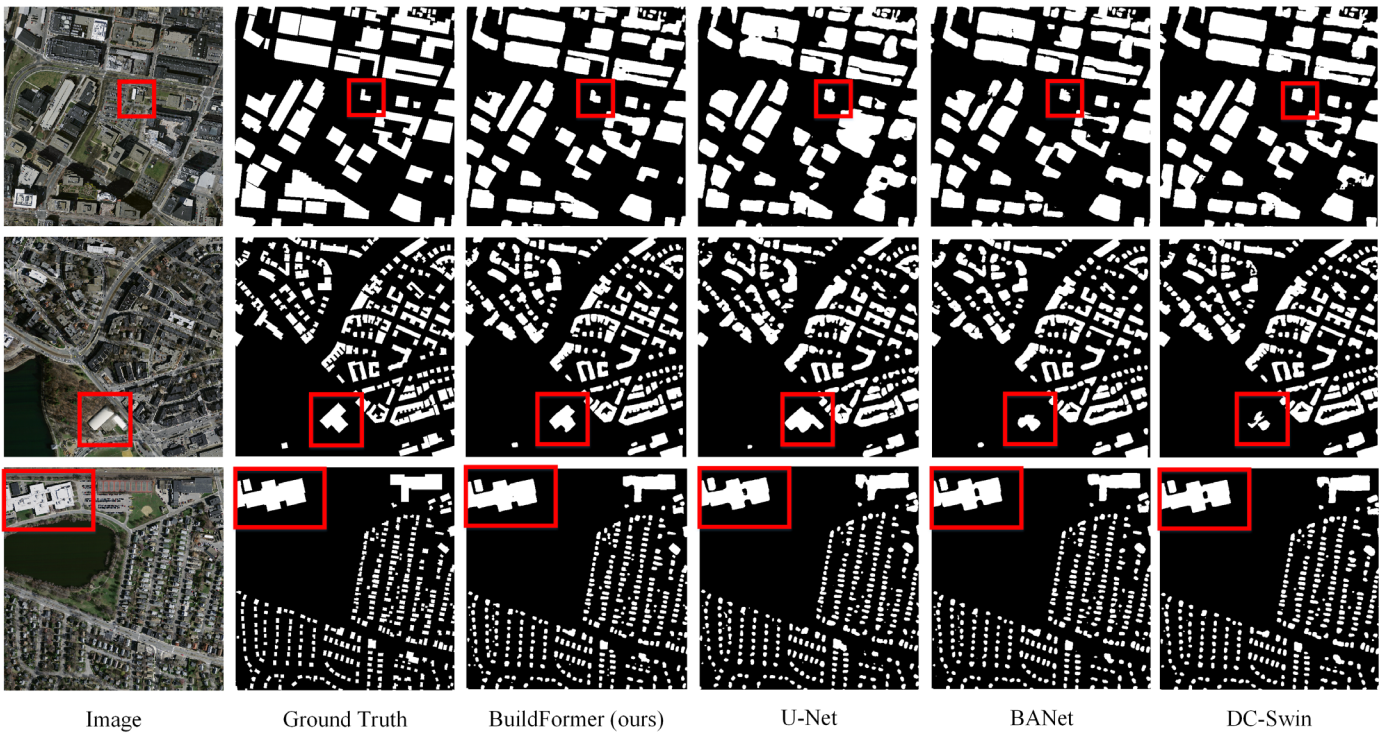
Image	Ground Truth	BuildFormer (ours)	U-Net	BANet	DC-Swin

Fig. 6. Visualized results of the U-Net, BANet, DC-Swin and BuildFormer (ours) on the Massachusetts Building dataset.

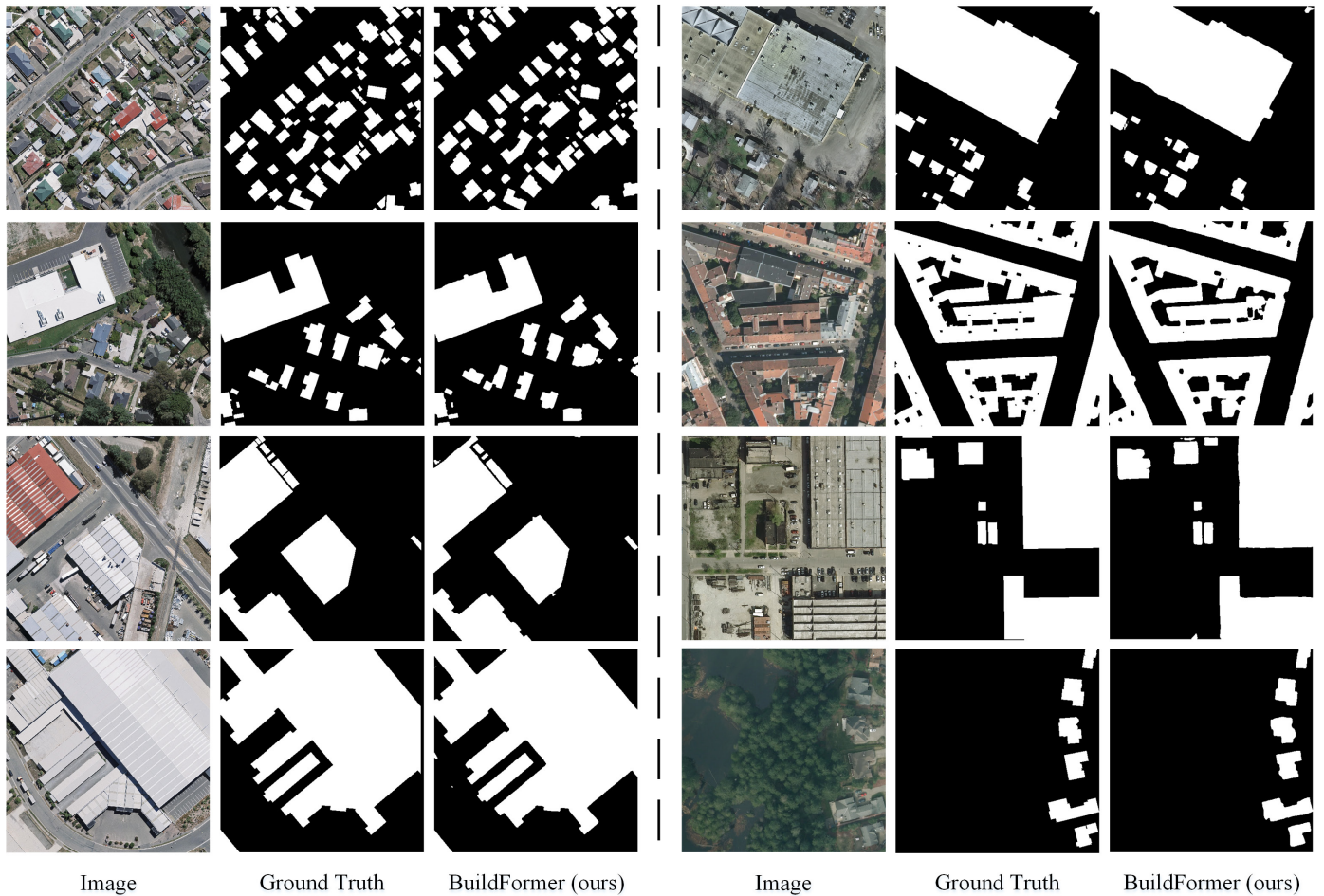|					
Image	Ground Truth	BuildFormer (ours)	Image	Ground Truth	BuildFormer (ours)

Fig. 7. Predicted results of the BuildFormer on the WHU Building dataset (left) and the Inria Aerial Image Labeling dataset (right).

REFERENCES

[1] W. Li, C. He, J. Fang, and H. Fu, "Semantic segmentation based building extraction method using multi-source gis map datasets and satellite imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 238-241.

[2] L. Dong and J. Shan, "A comprehensive review of earthquake-induced building damage detection with remote sensing techniques," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 84, pp. 85-99, 2013.

[3] M. Belgiu and L. Drăguţ, "Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 96, pp. 67-75, 2014/10/01/ 2014, doi: https://doi.org/10.1016/j.isprsjprs.2014.07.002.

[4] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Building detection in very high resolution multispectral data with deep learning features," in *2015 IEEE international geoscience and remote sensing symposium (IGARSS)*, 2015: IEEE, pp. 1873-1876.

[5] D. Griffiths and J. Boehm, "Improving public data for building segmentation from Convolutional Neural Networks (CNNs) for fused airborne lidar and image data using active contours," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 154, pp. 70-83, 2019.

[6] 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 Transactions on Geoscience and Remote Sensing,* vol. 59, no. 5, pp. 4287-4306, 2020.

[7] Y. Zhang, "Optimisation of building detection in satellite images by combining multispectral classification and texture filtering," *ISPRS journal of photogrammetry and remote sensing,* vol. 54, no. 1, pp. 50-60, 1999.

[8] 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," *International Journal of Applied Earth Observation and Geoinformation,* vol. 34, pp. 58-69, 2015.

[9] F. Dornaika, A. Moujahid, Y. El Merabet, and Y. Ruichek, "Building detection from orthophotos using a machine learning approach: An empirical study on image segmentation and descriptors," *Expert Systems with Applications,* vol. 58, pp. 130-142, 2016.

[10] Y. LeCun, Y. Bengio, and G. J. n. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-444, 2015.

[11] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730-3738.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, pp. 1097-1105, 2012.

[13] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks," *IEEE Geoscience and Remote Sensing Letters,* vol. 13, no. 1, pp. 105-109, 2016, doi: 10.1109/LGRS.2015.2499239.

[14] 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.

[15] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 152, pp. 166-177, 2019/06/01/ 2019, doi: https://doi.org/10.1016/j.isprsjprs.2019.04.015.

[16] P. Liu *et al.*, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sensing,* vol. 11, no. 7, p. 830, 2019.

[17] 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.

[18] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 57, no. 1, pp. 574-586, 2018.

[19] G. Wu *et al.*, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sensing,* vol. 10, no. 3, p. 407, 2018.

[20] Z. Shao, P. Tang, Z. Wang, N. Saleem, S. Yam, and C. Sommai, "BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images," *Remote Sensing,* vol. 12, no. 6, p. 1050, 2020.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 2020: Springer, pp. 173-190.

[27] 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 Transactions on Geoscience and Remote Sensing,* 2020.

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

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

[30] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," *arXiv preprint arXiv:2010.04159,* 2020.

[31] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881-6890.

[32] K. Han *et al.*, "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2022.

[33] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys (CSUR),* 2021.

[34] 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.

[35] 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 Journal of Photogrammetry and Remote Sensing,* vol. 183, pp. 240-252, 2022.

[36] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325-341.

[37] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS journal of photogrammetry and remote sensing,* vol. 86, pp. 21-40, 2013.

[38] M. Awrangjeb, C. Zhang, and C. S. Fraser, "Improved building detection using texture information," *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* vol. 38, pp. 143-148, 2011.

[39] X. Huang and L. Zhang, "A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery," *Photogrammetric Engineering & Remote Sensing,* vol. 77, no. 7, pp. 721-732, 2011.

[40] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 5, no. 1, pp. 161-172, 2011.

[41] Z. Li, W. Shi, Q. Wang, and Z. Miao, "Extracting man-made objects from high spatial resolution remote sensing images via fast level set evolutions," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 53, no. 2, pp. 883-899, 2014.

[42] T. Zhang, X. Huang, D. Wen, and J. Li, "Urban building density estimation from high-resolution imagery using multiple features and support vector regression," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 10, no. 7, pp. 3265-3280, 2017.

[43] G. Zhou and X. Zhou, "Seamless fusion of LiDAR and aerial imagery for building extraction," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 52, no. 11, pp. 7393-7407, 2014.

[44] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS journal of photogrammetry and remote sensing,* vol. 151, pp. 91-105, 2019.

[45] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Deep multisensor learning for missing-modality all-weather mapping," *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 174, pp. 254-264, 2021.

[46] Y. Sun, Y. Hua, L. Mou, and X. X. Zhu, "Cg-net: Conditional gis-aware network for individual building segmentation in vhr sar images," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 60, pp. 1-15, 2021.

[47] 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 Sensing,* vol. 10, no. 1, p. 144, 2018.

[48] S. Wei, S. Ji, and M. Lu, "Toward automatic building footprint delineation from aerial images using cnn and regularization," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 58, no. 3, pp. 2178-2189, 2019.

[49] 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 Transactions on Neural Networks and Learning Systems,* 2020.

[50] H. Hosseinpour, F. Samadzadegan, and F. D. Javan, "CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images," *ISPRS Journal of*

Photogrammetry and Remote Sensing, vol. 184, pp. 96-115, 2022.

[51]   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.

[52]   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.

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

[54]   M. Guo, H. Liu, Y. Xu, and Y. Huang, "Building extraction based on U-Net with an attention block and multiple losses," *Remote Sensing,* vol. 12, no. 9, p. 1400, 2020.

[55]   Q. Tian, Y. Zhao, Y. Li, J. Chen, X. Chen, and K. Qin, "Multiscale building extraction with refined attention pyramid networks," *IEEE Geoscience and Remote Sensing Letters,* vol. 19, pp. 1-5, 2021.

[56]   P. Das and S. Chand, "AttentionBuildNet for building extraction from aerial imagery," in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 2021: IEEE, pp. 576-580.

[57]   Z. Chen, D. Li, W. Fan, H. Guan, C. Wang, and J. Li, "Self-attention in reconstruction bias U-Net for semantic segmentation of building rooftops in optical remote sensing images," *Remote Sensing,* vol. 13, no. 13, p. 2524, 2021.

[58]   W. Deng, Q. Shi, and J. Li, "Attention-Gate-Based Encoder–Decoder Network for Automatical Building Extraction," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 14, pp. 2611-2620, 2021.

[59]   X. Pan *et al.*, "Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms," *Remote Sensing,* vol. 11, no. 8, p. 917, 2019.

[60]   J. Cai and Y. Chen, "MHA-Net: Multipath Hybrid Attention Network for building footprint extraction from high-resolution remote sensing imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 14, pp. 5807-5817, 2021.

[61]   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.

[62]   E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers,"

Advances in Neural Information Processing Systems, vol. 34, 2021.

[63]   Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral-spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Transactions on Geoscience and Remote Sensing,* 2021.

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

[65]   Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing,* vol. 13, no. 3, p. 516, 2021.

[66]   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.

[67]   X. He, Y. Chen, and Z. Lin, "Spatial-spectral transformer for hyperspectral image classification," *Remote Sensing,* vol. 13, no. 3, p. 498, 2021.

[68]   K. Xu, P. Deng, and H. Huang, "Vision Transformer: An Excellent Teacher for Guiding Small Networks in Remote Sensing Image Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing,* 2022.

[69]   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.

[70]   L. Gao *et al.*, "STransFuse: Fusing Swin Transformer and Convolutional Neural Network for Remote Sensing Image Semantic Segmentation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* vol. 14, pp. 10990-11003, 2021.

[71]   C. Zhang, W. S. Jiang, Y. Zhang, W. Wang, Q. Zhao, and C. J. Wang, "Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-high-resolution Remote Sensing Imagery," *IEEE Transactions on Geoscience and Remote Sensing,* 2022.

[72]   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.

[73]   X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation," *IEEE Transactions on Geoscience and Remote Sensing,* 2022.

[74]   Z. Xu, W. Zhang, T. Zhang, Z. Yang, and J. Li, "Efficient transformer for remote sensing image segmentation," *Remote Sensing,* vol. 13, no. 18, p. 3585, 2021.

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

[76] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection," *IEEE Transactions on Geoscience and Remote Sensing,* 2022.

[77] Z. Sun, W. Zhou, C. Ding, and M. Xia, "Multi-Resolution Transformer Network for Building and Road Segmentation of Remote Sensing Image," *ISPRS International Journal of Geo-Information,* vol. 11, no. 3, p. 165, 2022.

[78] 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 Geoscience and Remote Sensing Letters,* 2022.

[79] K. Chen, Z. Zou, and Z. Shi, "Building Extraction from Remote Sensing Images with Sparse Token Transformers," *Remote Sensing,* vol. 13, no. 21, p. 4441, 2021.

[80] W. Yuan and W. Xu, "MSST-Net: A Multi-Scale Adaptive Network for Building Extraction from Remote Sensing Images Based on Swin Transformer," *Remote Sensing,* vol. 13, no. 23, p. 4743, 2021.

[81] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6399-6408.

[82] G. Yang, Q. Zhang, and G. Zhang, "EANet: Edge-aware network for the extraction of buildings from aerial images," *Remote Sensing,* vol. 12, no. 13, p. 2161, 2020.

[83] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask R-CNN with building boundary regularization," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 247-251.

[84] 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.

[85] M. Fan *et al.*, "Rethinking bisenet for real-time semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9716-9725.

[86] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017: IEEE, pp. 3226-3229.

[87] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[88] 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.

[89] 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 Sensing,* vol. 11, no. 23, p. 2813, 2019.

[90] H. Liu *et al.*, "DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery," *Remote Sensing,* vol. 11, no. 20, p. 2380, 2019.