buildings

MDPI

Article

Enhanced Hybrid U-Net Framework for Sophisticated Building Automation Extraction Utilizing Decay Matrix

Ting Wang [1,†], Zhuyi Gong [2,†], Anqi Tang [2], Qian Zhang [2] and Yun Ge [3,*]

[1] BIM Research Institute, Nanchang Hangkong University, Nanchang 330036, China; 70545@nchu.edu.cn
[2] School of Civil Engineering and Architecture, Nanchang Hangkong University, Nanchang 330063, China; 2211085900023@stu.nchu.edu.cn (Z.G.); 2211085900218@stu.nchu.edu.cn (A.T.); 2211081400003@stu.nchu.edu.cn (Q.Z.)
[3] School of Software, Nanchang Hangkong University, Nanchang 330063, China
* Correspondence: geyun@nchu.edu.cn; Tel.: +86-136-7791-6702
† These authors contributed equally to this work.

Abstract: Automatically extracting buildings from remote sensing imagery using deep learning techniques has become essential for various real-world applications. However, mainstream methods often encounter difficulties in accurately extracting and reconstructing fine-grained features due to the heterogeneity and scale variations in building appearances. To address these challenges, we propose LDFormer, an advanced building segmentation model based on linear decay. LDFormer introduces a multi-scale detail fusion bridge (MDFB), which dynamically integrates shallow features to enhance the representation of local details and capture fine-grained local features effectively. To improve global feature extraction, the model incorporates linear decay self-attention (LDSA) and depthwise large separable kernel multi-layer perceptron (DWLSK-MLP) optimizations in the decoder. Specifically, LDSA employs a linear decay matrix within the self-attention mechanism to address long-distance dependency issues, while DWLSK-MLP utilizes step-wise convolutions to achieve a large receptive field. The proposed method has been evaluated on the Massachusetts, Inria, and WHU building datasets, achieving IoU scores of 76.10%, 82.87%, and 91.86%, respectively. LDFormer demonstrates superior performance compared to existing state-of-the-art methods in building segmentation tasks, showcasing its significant potential for building automation extraction.

Keywords: building segmentation; convolutional neural network; vision transformer

check for updates

Citation: Wang, T.; Gong, Z.; Tang, A.; Zhang, Q.; Ge, Y. Enhanced Hybrid U-Net Framework for Sophisticated Building Automation Extraction Utilizing Decay Matrix. *Buildings* **2024**, *14*, 3353. https://doi.org/10.3390/buildings14113353

Academic Editor: Osama Abudayyeh

Received: 23 September 2024
Revised: 13 October 2024
Accepted: 21 October 2024
Published: 23 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Accurate and effective automatic extraction of buildings from high-resolution remote sensing images is of great importance for a wide range of application scenarios, such as urban planning [1], population density estimation [2], and disaster management [3]. Accurate and effective building extraction can improve the efficiency and accuracy of urban planning and provide reliable data support for population density estimation. Additionally, it can quickly provide information on building damage in disaster management, thereby accelerating emergency response and resource allocation.

Traditional building extraction methods mainly use machine learning methods (e.g., support vector machines and random forests) to artificially design features for building segmentation based on characteristics such as spectra, shape, texture, shadows, and height [4–7]. However, buildings in remote sensing images have diverse sizes, shapes, colors, and textures, and buildings in images are often obscured by various types of shadows, making it difficult to identify the boundary contours [8], as shown in Figure 1. Relying only on human-designed features cannot adequately cover all the changing situations, and the model's generalization ability is insufficient, so there is an urgent need for more accurate, effective, and automated extraction techniques.



(a) **(b)**

Figure 1. Some challenges in remote sensing images. (**a**) Buildings vary in size, shape, texture, and color. (**b**) Shadows obscure the buildings in remote sensing images.

In recent years, convolutional neural networks have attracted much attention for their powerful feature learning ability [9]. In the field of remote sensing, a convolutional neural network (CNN) [10] has been applied to many geospatial tasks [11,12], especially automated building extraction [13–15]. Compared with traditional machine learning segmentation methods, CNNs can capture finer-grained local contextual information without manually designing the extraction features, especially when Shelhamer et al. proposed a fully convolutional neural network (FCN) [16] in 2015, which achieves end-to-end pixel-level segmentation. It also laid the foundation for convolutional neural networks to be used to solve the image semantic segmentation problem. The UNet network, which is innovative and improved based on FCN, further solves the problems of rough segmentation and fuzzy segmentation boundary [17,18]. For example, Xu et al. [14] proposed a UNet building extraction model guided by deep residual and filtering, which showed excellent performance in extracting buildings from a variety of targets in urban areas. Li et al. [15] proposed pyramid feature extraction to construct a multi-scale representation of buildings to alleviate edge inaccuracies caused by scale differences between low-level and high-level features. However, the above networks all have a fixed-size local sense field through the convolutional operation of CNN [19], and thus they tend to capture local features and exhibit a weak understanding of the global context when dealing with remote sensing building segmentation tasks, which implies limitations when dealing with buildings of different scales or complex backgrounds.

Recent research suggests that obtaining sufficient contextual architectural training data is one of the keys to effective model training [20]. To enhance the network's ability to capture global contextual information, ViT has been introduced for remote sensing image segmentation [21]. Different from the operation of convolution, ViT captures global dependencies in remote sensing images through a self-attention mechanism, which effectively makes up for the deficiency of CNNs, which can only extract local features but cannot model global contextual information, and which has made a breakthrough in semantic segmentation of remote sensing images [22,23].

However, relying only on ViT ignores spatial detail information, although it effectively captures global context information, but both detail information and global information are essential for achieving the fine segmentation of remote sensing buildings. Therefore, choosing a hybrid model based on CNN–Transformers can allow the full play of the advantages of the two in retaining local details and global long-range dependency [24–26]. However, due to the different computational paradigms of CNN and Transformers, how to effectively integrate the feature information of the two and strengthen the ability of the network to capture both local and global features is still a problem to be solved.

To address the challenge of inadequate local and global feature extraction in remote sensing building segmentation, we propose a novel segmentation framework named LDFormer. Specifically, we introduce a multi-scale detail fusion bridge (MDFB) within the encoder–decoder skip connections to fuse multi-scale shallow features adaptively.

Additionally, we optimize the transformer-based decoder by incorporating linear decay self-attention (LDSA) to mitigate the long-distance dependency issues present in sequence models. Furthermore, we propose the depthwise large separable kernel multi-layer perceptron (DWLSK-MLP), which integrates large kernel convolution with depth-wise separable convolution to effectively capture global boundary positional features. The main contributions of this paper are as follows:

1. We propose a multi-scale detail fusion bridge (MDFB) module that dynamically integrates and optimizes multi-scale shallow features to enhance the network's segmentation of buildings across various scales. Inspired by graph attention fusion, this module dynamically merges multi-scale feature maps and enhances attention to local detail features through detail attention mechanisms.
2. We propose a linear decay self-attention (LDSA) strategy to effectively address long-distance dependency issues in large-scale remote sensing images. This approach introduces a mathematical constraint (weight matrix) based on linear decay within the transformer self-attention computation paradigm. It ensures that the score matrix linearly decays concerning the distance from the central point, without introducing any additional model parameters.
3. We propose a depthwise large separable kernel multi-layer perceptron (DWLSK-MLP) module to improve the segmentation accuracy of complex building boundaries. This method integrates the concepts of large kernel convolution and depth-wise separable convolution to further compensate for positional boundary features, without introducing excessive computational or parameter overhead.
4. The method achieves a significant improvement in building segmentation accuracy compared to other state-of-the-art methods. On three publicly available datasets, the proposed method improves over BuildFormer in terms of IoU metrics by 1.07%, 1.63%, and 1.13%.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive overview of the relevant technologies and the development history involved in the automated extraction of buildings from remote sensing imagery. Section 3 presents a detailed description of the LDFormer model architecture, with a focus on the design of its key components: MDFB, LDSA, and DWLSK-MLP. Section 4 outlines the experimental processes of the model, including the experimental setup, results on the datasets, and investigations of ablation studies. Section 5 discusses relevant information concerning the model. Finally, Section 6 concludes the paper.

2. Related Work

2.1. Building Extraction Based on U-Net Network

Full Convolutional Neural Network (FCN) [16] Since the first realization of semantic segmentation of images with end-to-end training, FCN-based model variants have become the dominant approach to building extraction in remote sensing [27,28]. However, the fully convolutional structure of FCN cannot fully utilize the rich local and detailed features (e.g., texture, shape, etc.) of remotely sensed buildings, which can affect the accuracy of segmentation, and the convolution-inverse convolution structure of FCN can lead to blurring or inaccuracy of the boundary contours of remotely sensed buildings, which ultimately affects the quality of segmentation [8,29].

In order to solve the above problems, an encoder–decoder network called U-Net has been proposed for the semantic segmentation of remotely sensed buildings as an improvement and development of FCN [16]. The encoder part is responsible for receiving the input remote sensing image and gradually reducing its spatial dimensions to extract high-level semantic features of an image, while the decoder's task is to map the features extracted via the encoder back to the dimensions of the original input image and to gradually recover the spatial information to generate the segmentation results at the pixel level. In addition, the uniqueness of the U-Net network is that there is a jump connection between its encoder and decoder, which allows the decoder to realize the combination of

upper-level semantic information and shallow spatial location information, which helps to realize the balanced segmentation process to preserve and convey rich detail information, as well as to ensure the quality of the boundary contour of the remotely sensed building to improve the accuracy and clarity of the segmentation results. For example, Qiu et al. [30] designed the jump connection to introduce the ASPP module and the in-depth separable convolution (IDSC) module to capture more levels of features, which further improved the accuracy of building segmentation. Li et al. [31] developed a dual-branching module at the decoder side to capture multi-scale information and enhance high-level semantic detail features in order to further improve the accuracy of building segmentation. You et al. [32] embedded spatial and channel squeezing excitation (scSE) at the decoder side for forward correction of features to improve the accuracy and speed of small building extraction.

Although the U-Net network with an encoder–decoder structure has made great progress and development in the semantic segmentation of remote sensing buildings, in remote sensing images, buildings may exist in different environments and backgrounds, and global context information is also very important for the correct understanding and segmentation of building locations and boundaries, and the CNN-based U-Net network mainly focuses on the local feature extraction and processing. The CNN-based U-Net network mainly focuses on the extraction and processing of local features, and it has a weak ability to understand the global context of the overall large-scale remote sensing image. On the other hand, the U-Net network relies on the transfer mechanism of local connections and hopping connections between the encoder and the decoder, which is not able to take full advantage of the long-distance dependence of pixels in the image, thus leading to an insufficient understanding of the correlation relationship between the building and its surroundings and segmentation of the complex remote sensing image [33].

2.2. Building Extraction Based on Vision Transformer

Vit-based methods have made significant progress and development in the field of the semantic segmentation of remote-sensing images [22,34]. Different from the U-Net network mentioned above, vit converts 2D images into 1D ordered sequences based on its unique self-attention mechanism to fully capture the global information in remote sensing images, which is no longer limited to local features, and the network, based on the self-attention mechanism, can capture the long-distance dependency between any two positions in the input sequence of remote sensing images, which means that the vision transformer based model can better understand the relationship between buildings and their surroundings and improve segmentation accuracy. Driven by this, many researchers in the field of remote sensing have introduced the visual transformer to segmentation-related tasks, such as road extraction [35,36], surface change detection [37], land cover classification [38], and urban scene segmentation [39,40], especially building extraction [41–43]. For example, Wang et al. [22] designed global context paths and window-based linear multi-head self-attention to capture global dependencies, thus achieving state-of-the-art performance on the Massachusetts building dataset. Yuan et al. [42] decoded different levels of feature maps obtained based on swin transformer coding, and to adaptively add weights to each level of features, they utilized convolution for fusion in building extraction, which allowed the network to obtain richer multi-scale features and global information, further improving the performance of building segmentation.

Although, compared with the U-Net network-based method, the visual transformer-based method can fully extract the global information and establish the long-distance dependency, at the same time, it cannot effectively capture the local detail information of the image. But for the task of building segmentation for high-resolution remote sensing images, the local detail information is essential to realizing the fine segmentation of remote sensing buildings.

2.3. Graph Attention Mechanisms

To focus on both local and global features during remote sensing building extraction, we chose a CNN-based encoder and a transformer-based decoder. On this basis, to obtain richer detail features for the model, we chose to introduce the graph attention fusion strategy for multi-scale detail fusion at jump connection [44]. Unlike the traditional jump-connection feature fusion strategy, the graph attention mechanism can be used to learn relationships and feature representations between nodes without the limitation of fixed spatial convolution kernel sizes and shapes, and it can flexibly learn multi-scale shallow detail features. On the other hand, the graph attention mechanism can enable the model to adaptively focus on important features by dynamically assigning the attention weights between nodes, which can effectively enhance the fusion of local and global information in remote sensing images and ultimately perform building segmentation more accurately [45–47]. For example, Zi et al. [48], after constructing a pixel dependency graph from the extracted remote sensing image feature maps, designed a graph attention network to obtain the correlation of remote sensing image pixels as a way to improve the distant dependency between pixels, which in turn improves the performance of the semantic segmentation of remote sensing images. Deng et al. [49] proposed a spatial graph inference module and a dual spatial graph refinement module for constructing the neighborhood relationship of dense pixels and the correlation of building structural information, achieving state-of-the-art performance on high-resolution building datasets.

3. Methodology

This section presents the proposed LDFormer architecture and discusses and analyses its key designs. These key designs include multi-scale detail fusion bridge (MDFB), linear decay self-attention (LDSA), and depthwise large separable kernel multi-layer perceptron (DWLSK-MLP).

3.1. CNN-Based Encoder

The LDFormer network structure is shown in Figure 2. We adopted ResNet-18 as the shallow semantic extractor on the Encoder side to significantly reduce the model size and computational cost. This choice has been demonstrated in a wide range of semantic segmentation tasks, showing efficient properties. Specifically, ResNet-18 consists of four residual Blocks to extract shallow semantic features efficiently. During the feature compression process, each Block is downsampled by shrinking the spatial resolution by a factor of two, and the number of channels is expanded by a factor of two to mitigate the loss of detailed features due to downsampling. On each Block, a jump connection is used to output to the multi-scale detail fusion bridge (MDFB) for feature fusion and further refinement of feature extraction to bridge the semantic gap between shallow and deep features.

3.2. Transformer-Based Decoder

For the Decoder expansion phase, four LDBlocks are used for deep feature extraction and expansion. The LDBlock structure contains linear decay self-attention (LDSA) and a depthwise large separable kernel multi-layer perceptron structure (DWLSK-MLP), as shown in Figure 3. The input of each LDBlock comes from the multiscale detail fusion bridge (MDFB) and the output of the previous LDBLock.

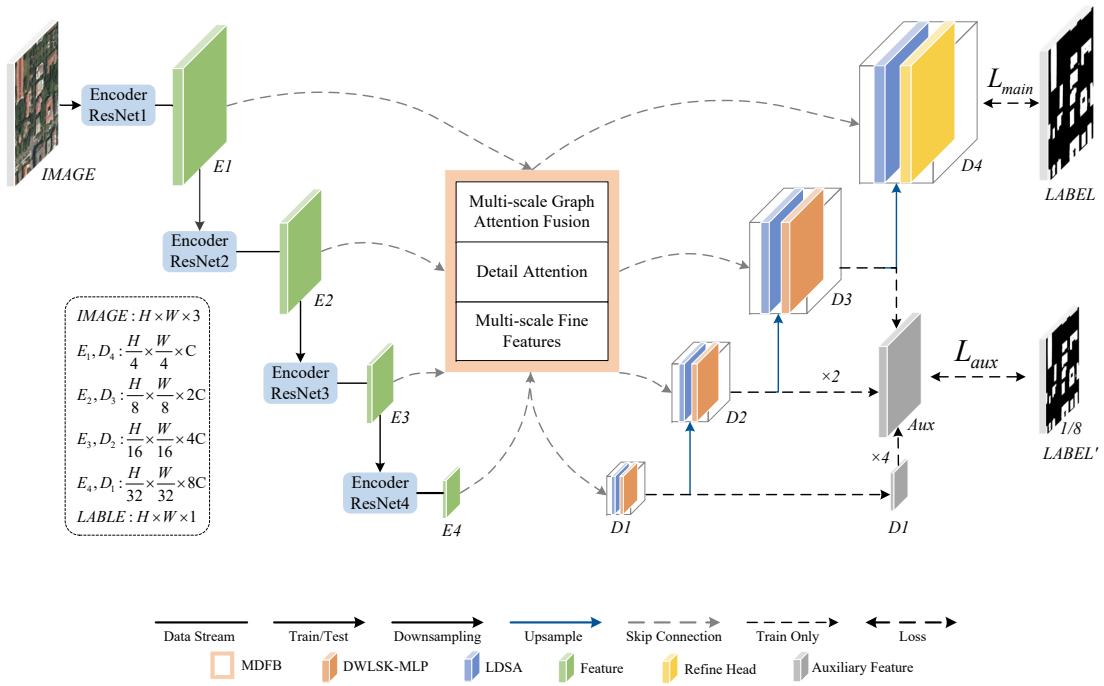


Figure 2. An overview of the LDFormer model.

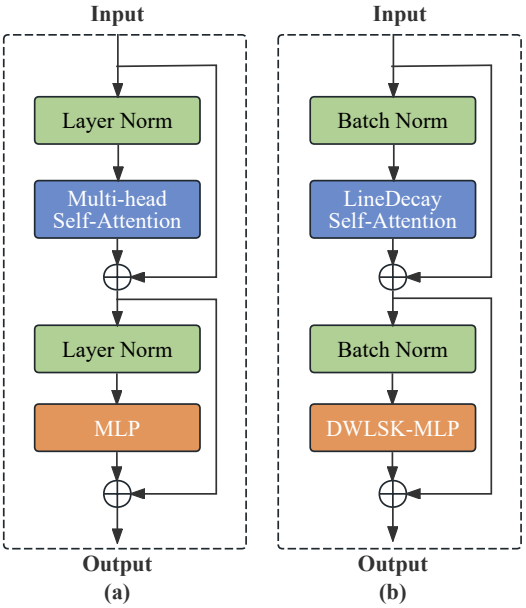


Figure 3. The structure of LDBlock. (**a**) Block in Swin-Transformer. (**b**) LDBlock in LDFormer.

3.2.1. Multi-Scale Detail Fusion Bridge (MDFB)

Shallow features usually contain rich detail and texture information, which is important for high-quality image recovery [50]. Unlike other static fusion methods, a multi-scale detail fusion bridge (MDFB) can dynamically fuse shallow features according to the network demand by adaptively fusing multi-scale shallow features at the decoder side and then expanding the attention to important features by using the attention mechanism to improve the quality and accuracy of image synthesis. MDFB consists of three parts: multiscale graph attention fusion, detail attention, and multiscale fine features, as shown in Figure 4, and each part is independent of the others, possessing great flexibility and maintainability.

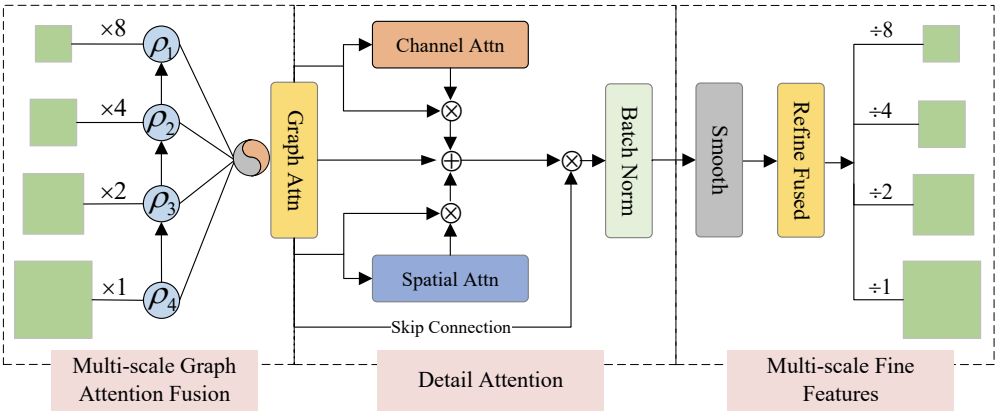


Figure 4. The structure of Multi-scale Detail Fusion Bridge (MDFB).

The Multi-scale Graph Attention (MSGAT) module is inspired by the weighted fusion of Graph Attention. The module assigns each shallow feature at each scale a weight value that can be learned, and finally, the up-sampling is scaled down to the same size for adaptive selection fusion, so that the model can dynamically learn to monitor the more important fine-grained features. Let the encoder have shallow features f_1, f_2, f_3, f_4; the feature map sizes are $\frac{H}{32} \times \frac{W}{32} \times 8c$, $\frac{H}{16} \times \frac{W}{16} \times 4c$, $\frac{H}{8} \times \frac{W}{8} \times 2c$, $\frac{H}{4} \times \frac{W}{4} \times c$, and the corresponding weight values of the feature maps are ρ_1, ρ_2, ρ_3, ρ_4. Then, after the Multi-scale Map Attention Fusion to obtain the output features F_g, the number of channels is C_{out}, which can be expressed as the following Equation:

$$\begin{aligned} F_g = \mathrm{MSGAT}(f_1', f_2', f_3', f_4') &= \rho_1 * f_1' + \rho_2 * f_2' + \rho_3 * f_3' + \rho_4 * f_4' \\ f_1' &= CONV_{1x1}(UP_8(f_1)) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_{out}} \\ f_2' &= CONV_{1x1}(UP_4(f_2)) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_{out}} \\ f_3' &= CONV_{1x1}(UP_2(f_3)) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_{out}} \\ f_4' &= CONV_{1x1}(UP_1(f_4)) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_{out}} \end{aligned} \tag{1}$$

where $CONV_{1x1}$ is the convolution with kernel size 1, and the number of output channels is C_{out}; $*$ denotes the number multiplication operation, UP_8, UP_4, UP_2, UP_1 represent the up-sampling with sizes of 8, 4, 2, and 1, respectively, and the final output of F_g is of the dimensions $\frac{H}{4} \times \frac{W}{4} \times C_{out}$.

The Detail Attention module, by explicitly utilizing the channel attention and spatial attention mechanisms, enables the model to expand the level of attention to important fine-grained features in order to improve the accuracy and robustness of the model segmentation. In this strategy, the use is based on the conventional convolutional implementation of channel attention and spatial attention. Unlike other attention methods, we use the feature values weighted by the combination of spatial and channel convolution as the attention score map again, and we perform matrix multiplication re-weighting with the initial features again to expand the channels and local detailed features attended via the attention again. Specifically, the channel attention is compressed into a global context descriptor for each channel of the feature map through the global average pooling operation to obtain the compressed feature representing the global information of each channel, and then the compressed feature is learned and excited by the FC fully connected layer to obtain the channel attention feature with the same dimensions as the initial feature channel; the spatial attention feature is computed through the CNN convolution of 3 × 3 size. The spatial attention feature is calculated by a CNN convolution of size 3 × 3, the size of which is unchanged before and after the convolution, and the spatial attention map with channel

number 1 is output after Sigmoid activation. Let the input feature $F_g \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_{out}}$; then, the output feature F_d can be expressed as follows:

$$\begin{aligned} F_d &= BN\left(F_g \otimes Attn_{cb}\right) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times c_{out}} && (2) \\ Attn_{cb} &= \left(Attn_s \otimes F_g\right) \oplus \left(Attn_c \otimes F_g\right) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times c_{out}} && (3) \\ Attn_s &= CONV_{3\times3}\left(Sig\left(F_g\right)\right) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 1} && (4) \\ Attn_c &= EXCITATION(FC\Big(AVG_{pool}\left(F_g\right)\Big) \in \mathbb{R}^{1\times1\times C_{out}} && (5) \end{aligned}$$

where BN is the batch normal regularization operation, $\otimes$ denotes a matrix multiplication operation, $\oplus$ denotes the matrix addition operation, and AVG_{pool} is the channel global average pooling. Sig denotes the sigmoid activation operation to perform some nonlinear feature mapping. $CONV_{3x3}$ denotes a convolution operation with a convolution kernel size of 3×3 and an output channel number of 1. FC is the fully connected feature layer to learn and compress the channel features. $EXCITATION$ is an activation reconstruction layer to recover the compressed channel features to the same number of channels as the initial feature map.

The Multi-scale Fine Feature module mainly resamples the features after smoothing attention to the same size as the original encoder feature map to obtain a multi-scale fine shallow feature map. Let the input features $F_d \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_{out}}$; then, the final output features are $f_1'', f_2'', f_3'', f_4''$, which is expressed in the following Equation:

$$\begin{aligned} F_{sm} &= SMOOTH(F_d) \\ f_1'' &= CONV_{1x1}(DOWN_8(F_{sm})) \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times 8C} \\ f_2'' &= CONV_{1x1}(DOWN_4(F_{sm})) \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 4C} \\ f_3'' &= CONV_{1x1}(DOWN_2(F_{sm})) \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 2C} \\ f_4'' &= CONV_{1x1}(DOWN_1(F_{sm})) \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C} \end{aligned} \quad (6)$$

where $SMOOTH$ contains a set of combined operations of 3x3 convolution, BN regularization, and Relu activation. $CONV_{1x1}$ is still a scheduling convolution with a constant size, changing only the number of channels. $DOWN_8, DOWN_4, DOWN_2, DOWN_1$ denote downsampling operations with downsampling of 8, 4, 2, and 1, respectively.

3.2.2. Linear Decay Self-Attention (LDSA)

The introduction of a linear decay matrix (commonly referred to as a "decay matrix" or "decay weights") in transformer models is primarily aimed at enhancing the model's focus on relevant (nearby) information, thereby mitigating the self-attention mechanism's tendency to be influenced by information that is farther away from the current token. This adjustment helps address the long-range dependency problem, which refers to the model's difficulty in accurately capturing information that appears earlier in a sequence when processing long-sequence tasks (global context information). In the context of 2D images [51], this issue manifests as challenges in recognizing long-distance features. The underlying reason is that, in the process of updating states over time (or space), information may gradually dissipate or become blurred [21]. By incorporating a weight decay matrix into the self-attention mechanism, the model assigns certain decay weights to earlier information, allowing it to accumulate and update this information over a defined number of time steps, rather than relying solely on the most recent inputs. This approach effectively captures the dynamic changes within the sequence. It enhances the model's memory capacity and representational learning ability while alleviating some of the computational complexity and memory overhead associated with increasing sequence lengths.

LDSA reduces the weights linearly the further away from the token center by introducing a linearly decaying weight matrix in the self-attention score matrix, the effect of

which is shown in Figure 5; the algorithm flow is shown in Algorithm 1. The method does not introduce new learning parameters so that the changes in the sequence can be monitored dynamically, which effectively enhances the accuracy and effectiveness of the model for long-distance modeling. In a two-dimensional image, the picture will be cut into many small pieces; each piece is represented as a token, and unlike one-dimensional sequences, two-dimensional tokens exist not only in the sequence of dimensions but also in the spatial distance of the information relationship. For each token center, it can be represented by a unique coordinate in the plane. Based on the location of this center element, we can establish the spatial mapping relationship of all pixels in this plane about the token centroid using the point-to-point Hamiltonian distance, which is finally transformed into the weight coefficients in the decay matrix, and this decay coefficient is introduced in the self-attention score matrix. Let the decay matrix be $D^{2d} \in \mathbb{R}^{N\times N}$; N is the length of the token (generally set to the window size), the coordinates of the token centroid are (x_c, y_c), and the coordinates of any point in the plane are (x_c, y_c). Then, the linear decay matrix (x_c, y_c) can be expressed as the following Equation:

$$D^{2d}_{(n,c)} = 1 - \frac{\mathrm{Min}(|x_n - x_c|, |y_n - y_c|)}{N-1} \in \mathbb{R}^{N\times N} \tag{7}$$

where Min is the function that takes the minimum value, and N is the size of the partition window. Based on the introduction of the linear decay matrix, the linear decay self-attention (LDSA) mechanism can be expressed as follows:

$$\begin{aligned} LDSA(Q,K,V) &= LDAT \otimes V \in \mathbb{R}^{N\times d} \\ LDAT(Q,K,D^{2d}) &= Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right) \odot D^{2d} \in \mathbb{R}^{N\times N} \\ Q &= X_m W_q \in \mathbb{R}^{N\times d} \\ K &= X_m W_k \in \mathbb{R}^{N\times d} \\ V &= X_m W_v \in \mathbb{R}^{N\times d} \end{aligned} \tag{8}$$

where X_m is the input vector of the mth header, Q, K, and V are the three projection matrices, respectively, W_q, W_k, W_v are the query features, key features, and value features generated by the mapping, d denotes the dimension of the token, and d_k denotes the dimension of the feature header. $\otimes$ denotes performing the matrix multiplication operation, and $\odot$ denotes the Hamada product dot product operation. Since the introduced D^{2d} participation in the dot product operation, the amount of computation introduced is negligible.

Algorithm 1 Linear decay self-attention (LDSA)

Require: X_m, W_q, W_k, W_v, N
Ensure: LDSA output matrix $LDSA_out$
Initialize decay matrix $D^{2d} \leftarrow \mathrm{zeros}(N, N)$
Calculate token centroid coordinates (x_c, y_c)
for $n \in \{1, \ldots, N\}$ **do**
    for $c \in \{1, \ldots, N\}$ **do**
        $D^{2d}[n, c] \leftarrow 1 - \left(\frac{\min(|x_n - x_c|, |y_n - y_c|)}{N-1}\right)$
    end for
end for
$Q \leftarrow X_m W_q \quad K \leftarrow X_m W_k \quad V \leftarrow X_m W_v$
$A \leftarrow \frac{QK^T}{\sqrt{d_k}}$
$P \leftarrow \mathrm{Softmax}(A)$
$LDAT \leftarrow P \odot D^{2d}$
$LDSA_out \leftarrow LDAT \times V$ **return** $LDSA_out$

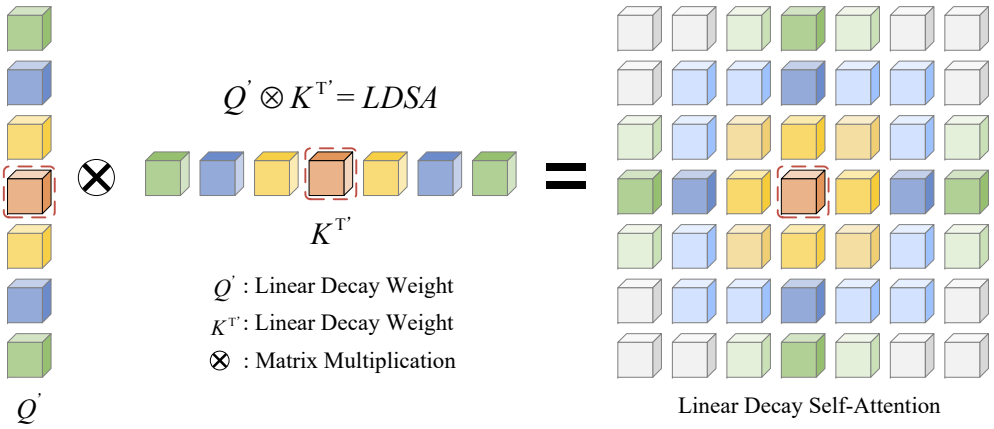


Figure 5. Illustration of the LDSA strategy. Different colors represent different weight values for the data; the closer to the token center, the greater the data weight value.

3.2.3. Depthwise Large Separable Kernel Multi-Layer Perceptron (DWLSK-MLP)

The latest research on transformer architectures reveals that incorporating depthwise (DW) convolutions within the multi-layer perceptron (MLP) structure can effectively facilitate the integration of features extracted via self-attention and convolutions, enhancing the model's ability to learn abstract features [52]. Guo et al. [53] proposed an inverted residual DW-MLP, composed of an expansion layer, a depthwise convolution, and a projection layer, which achieves better performance by altering the position of shortcut connections and introducing 3 × 3 depthwise convolutions for enhanced aggregation. Long et al. [54] noted that single-scale token aggregation fails to exploit rich channel representations fully and, therefore, proposed multi-scale token aggregation (MS-MLP) to capture rich multiscale features. However, relying solely on MLP-based feedforward neural networks for multiscale token aggregation to enrich the model's multiscale features has its limitations, as MLPs are also tasked with extracting higher-level abstract features. Consequently, a novel approach, the depthwise large separable kernel multi-layer perceptron (DWLSK-MLP) module, is proposed to further enhance the model's recognition accuracy, as detailed in Figure 6.

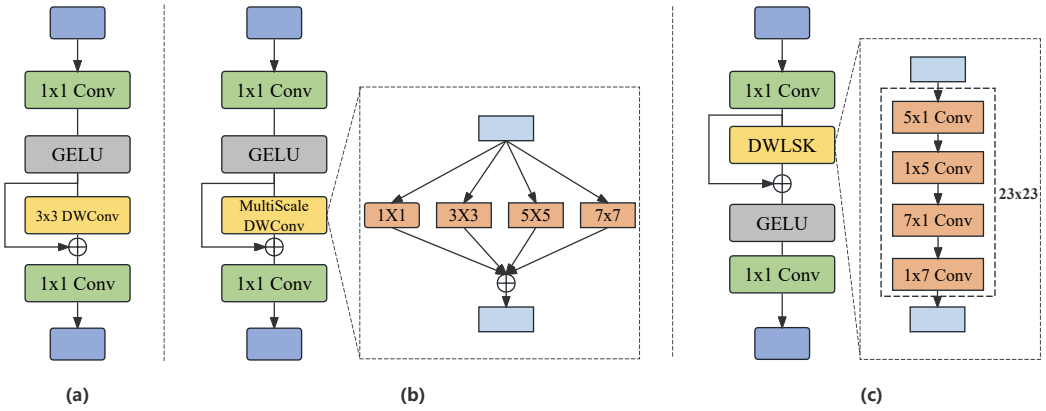


Figure 6. (**a**) DW-MLP. (**b**) MS-MLP. (**c**) Our DWLSK-MLP.

DWLSK-MLP effectively enhances the capture of boundary and large-scale features through a clever integration of large kernel convolution and depthwise separable convolution. Specifically, the model incorporates a 23 × 23 large convolution kernel, which enables the capture of global features across a wide area, thereby improving the precise recognition of complex boundary information. To ensure the integration of convolutional and self-attention features, we designed this kernel in a separable form. This innovative design consists of sequentially combined kernels of varying scales (including 5 × 1, 1 × 5, 7 × 1, and 1 × 7), which not only facilitates the capture of large-scale features but also

maintains the model's lightweight architecture while significantly enhancing its feature extraction capabilities across multiple orientations. These diverse convolution kernels are particularly adept at processing high-frequency information such as edges and corners, greatly enhancing the model's sensitivity to shape and boundary details. In summary, the design of DWLSK-MLP retains the functional characteristics of the fusion of convolutional and self-attention features while achieving a large receptive field without introducing excessive computational or parameter overhead, thereby significantly improving the model's ability to recognize features of small objects.

3.3. Loss Function

The model adopts Cross-entropy Loss L_{ce}, Dice Loss L_{dice}, and joint segmentation main loss L_{main}. In the training process, to improve the segmentation effect of multi-scale targets, we introduce the auxiliary loss constraints based on the Cross-entropy Loss function L_{aux}. According to Cross-entropy and Dice loss, L_{main}, L_{aux} can be expressed as follows:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{n=1}^{N}\sum_k = 1^{K} y_k^{(n)} \log\left(\hat{y}_k^{(n)}\right) \tag{9}$$

$$\mathcal{L}_{dice} = 1 - \frac{2}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{y_k^{(n)}\hat{y}_k^{(n)}}{y_k^{(n)} + \hat{y}_k^{(n)}} \tag{10}$$

$$\mathcal{L}_{main} = \mathcal{L}_{ce} + \mathcal{L}_{dice}, \quad \mathcal{L}_{aux} = \mathcal{L}_{ce} \tag{11}$$

where N denotes the number of samples, and K denotes the number of categories. $y_k^{(n)}$ denotes the uniquely hot coding of semantic labels of category k in that nth sample, and $\hat{y}_k^{(n)}$ denotes the confidence level of the prediction into category k in that nth sample. With the above elaboration, the total loss L of the model is L_{main}, L_{aux}, the sum of the two losses, and the formula is expressed as follows:

$$\mathcal{L} = \mathcal{L}_{main} + \mathcal{L}_{aux}. \tag{12}$$

4. Experiments

To validate the effectiveness of the proposed method, extensive experiments were conducted on three publicly available building datasets: Inria, Massachusetts, and WHU. The model was quantitatively and qualitatively compared with other state-of-the-art models. In addition, we set up multiple sets of ablation experiments on components, the choice of large kernel convolution, the choice of recession strategy, the number of multiple self-attentive heads, and the size of window divisions to fully demonstrate the superiority of the model performance and strategy.

4.1. Experiment Settings

4.1.1. Dataset

In our experiment, we utilized the Inira, Massachusetts, and WHU public building datasets to comprehensively evaluate the effectiveness of the model. The general overview of these datasets is presented in Table 1. These datasets are commonly employed in tasks such as building semantic segmentation, change detection, and object recognition, and they serve as typical representative benchmarks.

Table 1. Introduction to the dataset in the experiment.

Dataset	Resolution	Quantity	Origin Size	Train Size	Partitioning		
					Train	Value	Test
Inria	0.3 m/pixel	180	5000 × 5000	512 × 512	8756	1799	-
Massachusetts	1 m/pixel	151	1500 × 1500	1024 × 1024	137	4	10
WHU	0.3 m/pixel	8189	512 × 512	512 × 512	4736	1036	2416

(1) Inria: The Inria Aerial Image Labelling Dataset (IAILDS) comprises 360 high-resolution aerial images from five distinct cities: Austin, Chicago, Kitsap County, West Tyrol, and Vienna. These images, characterized by fine details and rich visual information, serve as an ideal testbed for building extraction and semantic segmentation tasks. Specifically, between one to five images from each city were selected as the validation set, while the remaining images were utilized for the model's training process. To ensure consistency and validity in image processing, images with an original size of 5000 × 5000 pixels were first padded to 5120 × 5120 pixels and subsequently cropped into smaller 512 × 512 pixel blocks. Noteably, the dataset originally contained numerous non-building areas; however, to focus on building-specific training and mitigate the risk of overfitting, we excluded image blocks that did not contain buildings, enhancing the model's focus and computational efficiency. Following this filtering process, a total of 8756 image blocks were used for training, and 1799 image blocks were used for validation. Such organization and preprocessing steps of the dataset not only optimize the quality of the training data but also provide a robust foundation for the training and evaluation of deep learning models.

(2) Massachusetts: The Massachusetts Buildings dataset is a collection of 151 aerial images of the Boston area, each with a spatial resolution of 1 m, covering an area of approximately 340 square kilometers. This dataset was carefully designed to contain two different types of scenes, urban and suburban, and all images are 1500 × 1500 pixels in size. The organization of the dataset consists of 137 images for training, 4 images for validation, and 10 images for testing. To enhance the generalization ability of the model, we used a data enhancement technique to expand the original training set to 411 images. In the model training phase, these images and their corresponding labels were randomly cropped to 1024 × 1024 pixels in order to serve as inputs to the model. Whereas, in the validation and testing phase, the images and labels were populated to 1536 × 1536 pixels to ensure that the image size could be divisible by 32 for subsequent computational processing. It is worth noting that we chose to ignore these padding parts added for computational convenience when calculating the model performance evaluation metrics. This meticulous processing helps more accurately assess the real-world performance of the model and ensures the fairness and reliability of the assessment results.

(3) WHU: This dataset was carefully constructed by Wuhan University (WHU), and it contains two main categories of image data: satellite images and aerial images. In this study, we focus on in-depth analysis and experiments on aerial images. Specifically, the aerial image dataset consists of 8189 image blocks, of which 4736 image blocks are used in the training phase, 1036 image blocks are used in the validation process, and the remaining 2416 image blocks are classified as the test set. It is worth mentioning that these images have a spatial resolution of 0.3 m, which means that every square meter of the area can be captured and rendered in fine detail. The entire dataset covers 22,000 buildings, which together form a vast area of more than 450 square kilometers, providing rich data support and an experimental basis for research on deep learning and semantic segmentation of buildings.

4.1.2. Implementation Details

In the experiments, all models were implemented using the PyTorch framework on a single NVIDIA GTX 3090 GPU (24 GB RAM). The encoder utilized pre-trained weights from ResNet-18 available in the timm library. We initially referenced the hyperparameter settings from BuildFormer and fine-tuned them based on extensive cross-validation to determine the optimal hyperparameters for LDFormer. The specific working environment is detailed in Table 2, while the foundational training parameters are provided in Table 3.

Table 2. Working environment details.

Component	Specification
Framework	PyTorch 1.10
GPU	NVIDIA GTX 3090
GPU memory	24 GB RAM
Encoder architecture	ResNet-18 (pre-trained weights from timm library)
Optimizer	AdamW
Data augmentation strategy	Horizontal and vertical random flips

Table 3. Training parameters.

Phase	Dataset	Epochs	Learning Rate	Weight Decay	Batch Size
Initial training	WHU	105	1×10^{-3}	0.0025	16
Initial training	Massachusetts	105	6×10^{-4}	0.0025	8
Initial training	Inria	105	6×10^{-4}	0.0025	12
Fine-tuning	All datasets	-	5×10^{-5}	0.0025	8

For training, we employed the AdamW optimizer, combined with a cosine annealing strategy, and we applied horizontal and vertical random flipping as data augmentation techniques. During the initial training phase, we uniformly set the number of training epochs to 105 for the WHU, Massachusetts, and Inria building datasets. The learning rates were set to 1×10^{-3} for WHU and 6×10^{-4} for Massachusetts and Inria, with a weight decay of 0.0025 and batch sizes of 16, 8, and 12, respectively.

In the fine-tuning phase, we adopted a progressive learning strategy, reducing the learning rate for all datasets to 5×10^{-5} while maintaining the weight decay at 0.0025 and using a batch size of 8 to ensure more precise adjustments as the model approaches the optimal solution. During the testing phase, we applied the same data augmentation techniques, including horizontal and vertical flips, as in the training process to ensure the validity of all experiments. Throughout the comparative and ablation experiments, the data augmentation and training strategies remained consistent.

4.1.3. Evaluation Indicators

In order to evaluate the effectiveness of the model, we used four metrics: IoU, precision, recall, and F1-score, which are often applied in the field of building extraction. IoU is the most common and important evaluation metric that can intuitively reflect the model's ability to separate the foreground and the background. When the value of IoU is higher, it indicates that the model's segmentation results are more accurate. Precision measures the accuracy of the model in predicting the true examples. Recall reacts to the model's ability to recognize the true examples. The F1-score is the reconciled average of accuracy and recall, which is used to evaluate the model's accuracy and recall performance comprehensively.

Here, we define TP as true positive, TN as true negative, FP as false positive, and FN as false negative. According to the above definitions, the relevant formulas can be expressed as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \quad \text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{14}$$

Additionally, to quantify the relative performance improvement, we introduce the percentage increase (PI) formula, which denotes the performance improvement between different methods:

$$\text{Percentage Increase} = \left(\frac{\text{New Metric} - \text{Baseline Metric}}{\text{Baseline Metric}} \right) \times 100\% \tag{15}$$

where "New Metric" represents the evaluation metric value of the new model, and "Baseline Metric" represents the evaluation metric value of the baseline model.

4.2. Experimental Result

4.2.1. Results on Massachusetts Dataset

Experiments were conducted to compare the U-Net, DeepLabv3+, MA-FCN, BRRNet, Bias-UNet, CBRNet, MANet, BANet, DC-Swin, FD-Net, BuildFormer, and BCTNet network models. The model contains a classical semantic segmentation network, a transformer architecture semantic segmentation network, a CNN–transformer hybrid architecture-based semantic segmentation model, and an advanced SOTA network model in the field of building extraction to adequately compare the performance difference between the model and various backbone networks.

Quantitative analysis: The results of the quantitative comparison experiments are shown in Table 4. Compared with other SOTA methods, LDFormer obtained excellent performance of 76.10 IoU, 87.75 in precision, 85.34 in recall, and an 86.12 F1-score. Specifically, LDFormer outperformed its SOTA method BCTNet by 1.06% in IoU, 0.18% in precision, 1.35% in recall, and 0.38% in F1. This significant comparison result fully proves the effectiveness of our proposed model and further argues for the effectiveness of the linear decay weighting strategy that we introduced.

Table 4. A quantitative comparison with existing methods was conducted on the Massachusetts dataset, with the best values in each column highlighted in bold. The percentage increase (PI) in IoU is presented relative to U-Net. # denotes results cited from the paper.

Method	IoU	Precision	Recall	F1-Score	PI (IoU)
U-Net [17]	67.61	79.13	82.29	80.68	-
DeepLabv3+ [55] #	69.23	84.73	79.10	81.82	+2.40%
MANet [56] #	70.76	82.00	83.77	82.88	+4.66%
BANet [40]	72.20	83.07	84.66	83.86	+6.79%
DC-Swin [57]	72.59	83.07	85.19	84.12	+7.37%
BRRNet [58] #	73.25	-	-	84.56	+8.34%
Bias-UNet [59] #	73.49	83.34	86.15	84.72	+8.70%
MA-FCN [28] #	73.80	87.07	82.89	84.93	+9.16%
FD-Net [60] #	74.54	**87.95**	83.02	85.42	+10.25%
CBRNet [61] #	74.55	86.50	84.36	85.42	+10.26%
BuildFormer [22]	75.03	86.69	84.79	85.73	+10.97%
BCTNet [25] #	75.04	87.57	83.99	85.74	+10.99%
LDFormer(ours)	**76.10**	87.75	**85.34**	**86.12**	**+12.56%**

Qualitative analysis: The local visualization comparison diagram is shown in Figure 7. We selected three typical image cases to compare the segmentation effect of our proposed LDFormer and other SOTA methods. In the first figure, we purposely selected the dense small-target prediction map in order to test the model's ability to predict dense small targets. As shown in the figure, there is a building on the left of the red rectangular box that is extremely similar to the background, which is difficult to distinguish with the naked eye, and there are two small buildings with clear outlines on the right, which are easily distinguishable between the foreground and the background. Only our proposed LDFormer model clearly and accurately predicted the three buildings' outlines completely, with the best overall recognition effect, while all other SOTA methods show the phenomenon of sticking and blurring outlines. This result further demonstrates the effectiveness of our proposed DWLSK-MLP method. Enhancing the effective sensing field via large kernel convolution effectively enhances the recognition of small targets in dense prediction tasks. In the second figure, we selected the prediction images with a mixture of large and small targets to detect the detail recognition ability of multi-scale mixed targets. As shown in the figure, two buildings are facing each other and spaced apart in the red rectangular box, the color, and texture of the buildings are still easy to distinguish relative to the background,

and the spacing bands between the buildings and the shadows of the buildings are very similar and difficult to distinguish. The visualization results indicate that only our model succeeds in identifying the spacing band between the two buildings clearly, while all other methods suffer from certain recognition errors and unclear boundaries. This result fully proves that our proposed method can fully learn the detail texture features at various scales and can effectively identify the differences between classes, and further proves the effectiveness of our introduced multi-scale detail fusion bridge (MDFB) method. In the third image, we chose images with large-scale architectural targets to test the model's ability to model against a global long-range context. As shown in the image, the red rectangular box contains a circular pool wrapped around the irregular boundary of the large building area. The prediction results show that our proposed LDFormer method, which predicts the circular pool most clearly and completely, is the best for large-scale irregular boundaries. The other methods, on the other hand, have some fragmentation for the irregular regions of large-scale buildings, and cannot completely identify the clear contours. This result fully proves the effectiveness of our proposed LDSA strategy, which significantly enhances the ability to learn the representation of global features, and it can accurately identify the super-large scale targets. In summary, our proposed LDFormer achieves superior performance in the recognition of inter-class differences, intra-class mutations, and multi-scale detailed features.

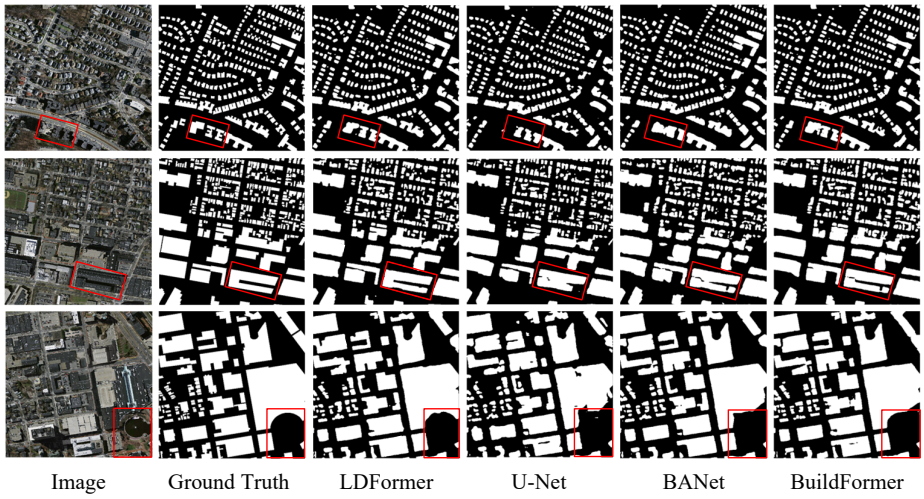


Figure 7. Qualitative comparison under Massachusetts test sets. We added some red boxes to highlight the differences in order to facilitate model comparison.

Big picture inference: the big picture inference visualization results are shown in Figure 8. We selected two typical representative images for global visualization to fully demonstrate the recognition effect of our proposed method from a global perspective view. The upper image in the inference mainly contains many dense, small-scale targets and a large, irregular background region, which mainly tests the model's ability to predict small-scale, dense targets and discriminate large-scale background targets. The lower panel mainly contains medium- and large-scale building targets, involving multiple building types with clear boundaries and irregularly shaped boundaries, which tests the model's ability to recognize global features and discriminate inter-class differences. The results show that the segmentation recognition effect of our proposed LDFormer method is close to the label, and the overall effect is within a tolerable normal error range.

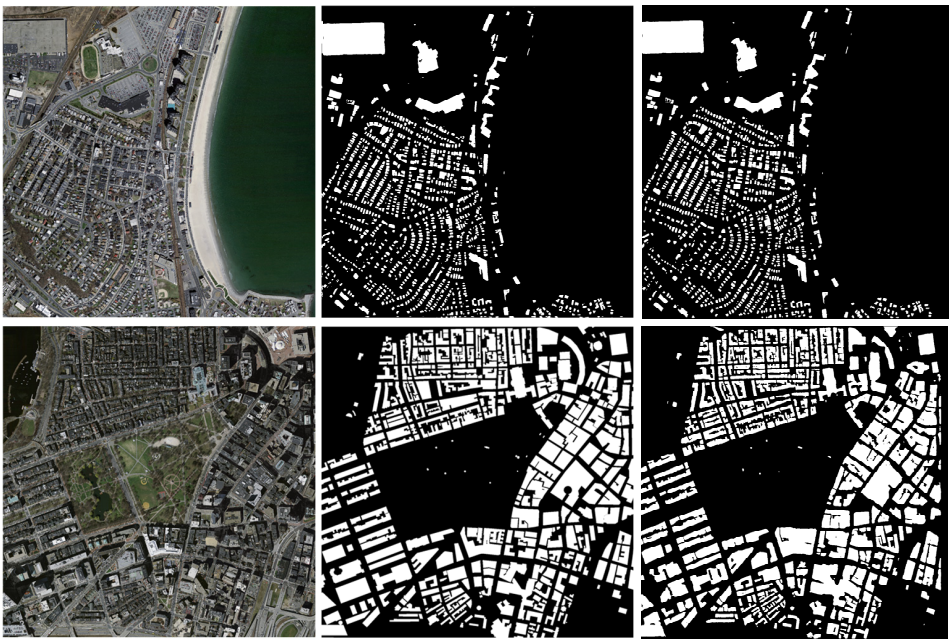


Figure 8. Visualization of large image inference on the Massachusetts dataset.

4.2.2. Results on Inria Dataset

The experiments compare the U-Net, SiU-Net, HRNet, Swin-UNet, SRI-Net, STT, SwinUperNet, Segmenter, Trans-UNet, CBRNet, and BuildFormer network models. A variety of backbone network models based on classical segmentation networks, pure transformer architecture-based networks, and CNN–transformer hybrid architecture-based networks are likewise addressed.

Quantitative analysis: The results of the quantitative comparison experiments are shown in Table 5. Compared with other SOTA methods, LDFormer obtains an excellent performance of 82.87 IoU, 91.79 in precision, 89.50 in recall, and 90.63 F1-scores, achieving the best performance in all metrics except recall. Specifically, LDFormer outperforms its SOTA method, BuildFormer, by 1.63% in IoU, 1.14% in precision, 0.72% in recall, and 0.92% in F1. This significant comparison result proves once again that our proposed method, LDFormer, is effective, and its introduction of the LDSA, MDFB, and DWLSK-MLP methods plays a crucial role.

Table 5. A quantitative comparison with existing methods was conducted on the Inria dataset, with the best values in each column highlighted in bold. The percentage increase (PI) in IoU is presented relative to U-Net. # denotes results cited from the paper.

Method	IoU	Precision	Recall	F1-Score	PI (IoU)
U-Net [17]	70.78	85.18	80.72	82.89	-
SiU-Net [27] #	71.40	84.60	82.10	83.33	+0.88%
HRNet [62]	75.03	86.56	84.92	85.73	+6.00%
Swin-UNet [63] #	76.27	86.87	86.21	86.14	+7.76%
SRI-Net [64] #	76.84	-	-	86.32	+8.56%
STT [41] #	79.42	-	-	87.99	+12.21%
SwinUperNet [51] #	79.53	87.55	**89.67**	88.60	+12.36%
Segmenter [65] #	80.72	90.38	88.30	89.33	+14.04%
Trans-UNet [66] #	80.77	90.44	88.32	89.36	+14.11%
CBRNet [61] #	81.10	89.93	89.20	89.56	+14.58%
BuildFormer [22]	81.24	90.65	88.78	89.71	+14.78%
LDFormer (ours)	**82.87**	**91.79**	89.50	**90.63**	**+17.08%**

Qualitative analysis: The local visualization comparison diagram is shown on the right side in Figure 9. We selected four representative image cases to compare the segmentation effect of our proposed LDFormer and other SOTA methods. As shown in the figure, the first image mainly tests the model's ability to recognize small-scale dense targets, the second image examines the model's ability to reconstruct the details of buildings with unclear boundaries and irregular shapes, the third tests the model's ability to model and learn large-scale targets with long-distance contextual relationships, and the fourth examines the model's ability to learn the interclass differences and interclass mutation in the environment of complex backgrounds and unbalanced class. Under the comprehensive test, the results show that our proposed method achieves excellent performance, and the results prove the effectiveness of our targeted improvement on the difficulties in the field of building recognition.

4.2.3. Results on WHU Dataset

To further assess the validity of the model, we conducted relevant experiments on the WHU dataset. Comparisons were made with the state-of-the-art SOTA model, which was thoroughly evaluated in both quantitative and qualitative aspects. The quantitative results are shown in Table 6, while the qualitative results are exhibited as shown in the left part of Figure 9. In the quantitative experiments, we propose that LDFormer achieves an excellent performance of 91.86 IoU, 95.47 in precision, 95.72 in recall, and a 95.73 F1. Compared with the SOTA method BCTNet, it achieves 0.71% IoU, 0.20% in precision, 0.45% in recall, and a 0.36% F1. In the quantitative experiments, we also carried out the same qualitative setup with the Inria dataset to test the model's discriminative ability from four perspectives in order to maximally and comprehensively accurately detect the feature recognition effect of the model. The results show that our proposed method achieves satisfactory segmentation results, but there are still some errors in the detail features of some multi-scale irregular buildings. It shows the multi-scale detail fusion bridge (MDFB) method proposed by us for details; although it moderates the loss of detail features to a certain extent, there is still a loss of detail features for an irregular and easy background mixing of building details, which can be further optimized.

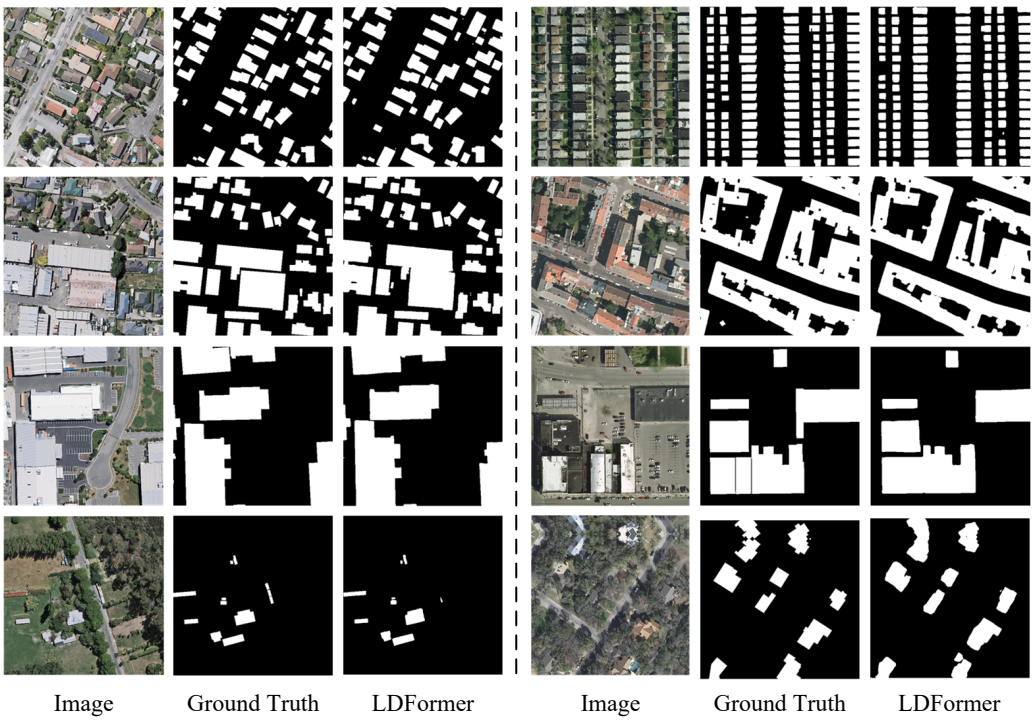


Figure 9. Qualitative comparison under WHU (**left**) and Inria (**right**) test sets.

Table 6. A quantitative comparison with existing methods was conducted on the WHU dataset, with the best values in each column highlighted in bold. The percentage increase (PI) in IoU is presented relative to U-Net. # denotes results cited from the paper.

Method	IoU	Precision	Recall	F1-Score	PI (IoU)
U-Net [17]	85.92	91.87	93.01	92.39	-
CU-Net [67] #	87.10	94.60	93.70	93.13	+1.37%
MSST-Net [42] #	88.00	-	-	88.20	+2.42%
SiU-Net [27] #	88.40	93.80	93.90	93.85	+2.89%
SRI-Net [64] #	89.23	**95.67**	93.69	94.51	+3.85%
DE-Net [68] #	90.12	95.00	94.60	94.08	+4.89%
STT [41] #	90.48	-	-	94.97	+5.31%
EU-Net [69] #	90.56	94.98	95.10	95.04	+5.40%
MA-FCN [28] #	90.70	95.20	95.10	95.15	+5.56%
BuildFormer [22]	90.73	95.15	95.14	95.13	+5.60%
MAP-Net [70] #	90.86	95.62	94.81	95.21	+5.75%
BCTNet [25] #	91.15	95.27	95.27	95.37	+6.09%
LDFormer(ours)	**91.86**	95.47	**95.72**	**95.73**	**+6.91%**

4.3. Ablation Experiment

4.3.1. Each Component of LDFormer

To further verify the effectiveness of each module, we conducted a large number of ablation experiments on the Massachusetts and Inria building datasets, and we recorded the ablation results in Table 7, where the best score for this column metric is highlighted in bold. It is worth noting that, for a fair comparison, we uniformly adopt the same random augmentation strategy and test-time augmentation strategy, and the relevant hyperparameters remain the same as the baseline model by default.

Table 7. Ablation study of LDFormer for each component of the Massachusetts and Inria building dataset. ↑ indicates the higher score, the better, and vice versa.

BASE	MDFB	LDSA	DWLSK-MLP	Massachusetts (%)				Inria (%)			
				IoU ↑	*P* ↑	*R* ↑	*F1* ↑	*IoU* ↑	*P* ↑	*R* ↑	*F1* ↑
✓				73.26	84.47	83.29	83.68	79.78	88.18	85.72	87.89
✓	✓			74.45	85.61	84.83	84.71	81.23	90.32	87.07	89.25
✓		✓		74.61	86.74	84.90	84.86	81.07	90.27	86.85	89.03
✓			✓	75.13	86.95	85.05	85.34	81.50	90.77	87.31	89.60
✓	✓	✓		75.69	87.46	85.19	85.58	82.22	91.26	88.02	90.04
✓		✓	✓	75.51	87.40	85.13	85.26	82.52	91.53	88.68	90.28
✓	✓		✓	75.86	87.53	85.22	85.72	82.46	91.55	88.83	90.23
✓	✓	✓	✓	76.10	87.75	85.34	86.12	82.87	91.79	89.50	90.63

The experimental results show that each module added to the baseline model has a certain effect and achieves a significant performance improvement. The MDFB module alone has the smallest improvement effect on the Massachusetts dataset, with an increase of 1.19% IoU, and the DWLSK-MLP module has the best recognition effect, with an increase of 1.87% IoU compared with the baseline model. Among the pairwise combinations, MDFB and DWLSK-MLP achieved the best results, and the other pairwise combinations achieved high and low levels on the Massachusetts and Inria datasets with similar effects. In the above experiments, the results of the two data sets are the best under the three module combinations, which fully verifies the effectiveness of our proposed strategy.

4.3.2. Choice of Large Kernel Convolution

DWLSK-MLP aims to effectively facilitate the decoupled fusion of convolution and self-attention while enhancing the effective sensory field and abstract feature extraction

capability of the model. To further explore the effect of using different large kernel convolutions on the model, we conducted an experimental discussion using the WHU building dataset, and the results are shown in Table 8. The results show that using a convolutional kernel of size 23 resulted in the best overall model, obtaining the best score of 91.83% IoU and only a slight increase in flops and parameters over a convolutional kernel of size 11. Compared to the 0.13% performance improvement, the improvement at almost no computational and parameter cost is tolerable. At the same time, the experiment shows that the model performance does not get better as soon as the convolutional kernel size increases, but it tends to be on an upward and then a downward curve. The reason for this is attributed to the characteristics of architectural images themselves. There are often irregular buildings of different sizes in the images, but due to the uniform existence of a fixed spatial resolution scale in the image dataset, the overall building size is biased towards a mean scale, in which the adapted convolution kernel can adequately fit the building features. Based on the experimental results presented, we set the large kernel in DWLSK-MLP as the best-performing 23, and then we sequentially depth-decomposed it into four deep convolutions of (1,5) → (5,1) → (1,7) → (7,1) to reduce model computation and complexity.

Table 8. Ablation study of different large-kernel convolutions on the WHU dataset. Depthwise separable and asymmetric convolutions are used for sequential decomposition.

Kernel Size	Convolution Sequence (DW)	Flops (G)	Params (M)	IoU (%)
11	(1,3) → (3,1) →(1,5) → (5,1)	66.83	27.10	91.73
23	(1,5) → (5,1) →(1,7) → (7,1)	66.87	27.12	91.86
35	(1,5) → (5,1) →(1,11) → (11,1)	66.92	27.15	91.64
41	(1,5) → (5,1) →(1,13) → (13,1)	66.94	27.16	91.60
53	(1,5) → (5,1) →(1,17) → (17,1)	66.98	27.19	91.51

4.3.3. Choice of Decay Strategy

In order to justify the choice of the optimal decay strategy, we conducted an ablation comparison experiment with different decay strategies on the Inria building dataset, and the results are shown in Table 9. The experiments compare the logarithmic, linear, Gaussian, and exponential decay strategies based on distance variation, and the relevant parameters are set uniformly during the process. The results show that the distance-based linear recession change of the weights gives the best results, obtaining the best 82.87% IoU; the distance-based exponential weight recession of the weights gives the worst results, obtaining 82.19% IoU. It is worth noting that, although the linear-based decay strategy gives the best overall results, obtaining the best scores on IoU, recall, and F1, it has the best results in the precision one metric, which is slightly lower than the Gaussian decay-based strategy. It shows that the pixel characteristics of this Inria building number image are overall Gaussian distributed, and the modeling constraints with linear variation can be closest to the real value. Based on the results of the experimental demonstration, we chose the weighting strategy based on linear decay to guarantee the optimal performance of the model.

Table 9. Ablation study of different decay strategies on the Inria dataset. All experiments were conducted on a single NVIDIA GTX 3090 GPU, with test inputs at a resolution of 1024 × 1024.

Decay Strategy	IoU	Precision	Recall	F1-Score
Logarithmic decay	82.55	91.63	89.23	90.44
Linear decay (ours)	82.87	91.79	89.50	90.63
Gaussian decay	82.64	91.89	89.14	90.50
Exponential decay	82.19	91.72	88.78	90.22

4.3.4. Number of Multi-Heads and Window Wize

In LDSA, affected by the self-attention paradigm, two important parameters, the size of the division window and the number of polytopes, will directly affect the performance of the module. To further explore the optimal parameter settings of the model, we conducted relevant ablation experiments, the results of which are shown in Figure 10. In the study of the ablation of the number of attention heads, the results show that the model achieves the best results when the number of heads is set to 16, and the model performance does not change linearly with the change of the number of heads, and setting it too large or too small will affect the model performance. In the study of the window size setting, the results showed that the best results were achieved when the window size was set to 8. The occurrence of this phenomenon indicates that the model relies more heavily on medium and small local detail features and less heavily on large-scale global features.

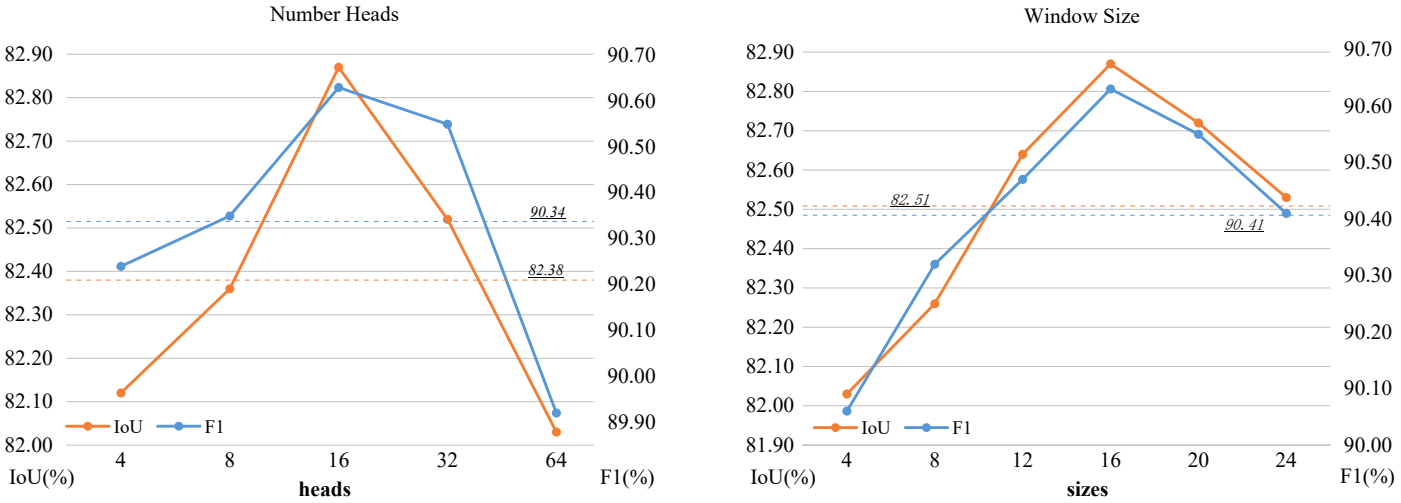


Figure 10. Ablation analysis of the impact of the number of model heads and window size using the Inria building dataset.

5. Discussion

5.1. About Encoder Selection

Considering that encoders with different feature extraction capabilities will affect the decoding fusion with the deep decoder, which in turn affects the performance of the model. We further explored the effect of different encoder backbones on the recognition accuracy of the model, and the specific results are shown in Table 10. The results show that ResNet18 achieves the best score of 82.87% IoU due to its simplicity and efficiency in extracting shallow features, leading to better overall fusion. Using ResNet50, ResNet101, and ResNet152 as the Encoder backbone, the overall performance of the model is 1–2% IoU higher in accuracy compared to other convolution-based backbones, such as Vgg16, Hrnet-w32, and EffIcientNet. Transformer backbone, with a worthwhile 0.59% IoU improvement in performance. It is worth noting that ResNet18 not only obtains performance beyond other complex-based architectures in terms of accuracy but also maintains a leading edge in terms of model parameter complexity, number of parameters, and computation. The overall number of parameters of the model is only 27.12 M, the complexity is 144.81 G, and the computational volume is 66.87 G. Compared with the equivalent lightweight Vgg16 backbone network, the overall accuracy is yet 1.33% IoU higher, and the number of parameters of the model is reduced by about 16.27 M compared to than that of the similarly-performing ResNet50 backbone network. Therefore, it was proven from the experimental results that ResNet18 has the best integrated recognition performance. So, in this paper, ResNet18 was used as the encoder-side backbone network to maintain the best performance.

Table 10. Ablation study with different encoders on the Inria building dataset.

Encoder	Params (M)	Complexity (G)	Flops (G)	IoU (%)
ResNet18 (ours)	27.12	144.81	66.87	82.87
ResNet50	43.39	199.05	86.35	82.67
ResNet101	62.48	275.02	105.85	82.60
ResNet152	78.13	337.59	125.38	82.50
Vgg16	34.07	160.60	335.09	81.54
Hrnet-w32	50.37	225.79	106.66	80.47
EfficientNet	81.04	349.42	89.48	80.53
Swin-Transformer	106.17	433.25	148.98	82.28

5.2. Selection of Fusion Strategies

To further demonstrate the advanced performance of the Graph Attention fusion strategy in the MDFB method, we conducted experiments with commonly used fusion strategies, and the experimental results are shown in Table 11. In the experiments, we compared some of the commonly used fusion strategies for deep learning, such as the feature pixel-level Add operation, feature channel-level Concat stacking fusion operation, and feature-based Hamada product element multiplication operation. The experimental results show that the Graph Attention-based fusion strategy is the most effective, obtaining 91.86% IoU value on the WHU building set, which is a 0.65% higher IoU value than that of the Hamada product multiplication-based operation. The results after fusion based on the Feature Add operation and Concat operation are almost the same, with a 0.3% lower IoU value compared to the Graph Attention fusion strategy. This result is attributed to the adaptive feature fusion strategy of Graph Attention, such that the output of the node comes from the weighted input of all the previous nodes. The weight values of the node's feature graph are adaptively and dynamically adjusted along with the needs of the network model to achieve optimal feature fusion.

Table 11. Discussion of the fusion mode of MDFB on the WHU dataset.

Fusion Method	WHU			
	IoU	Precision	Recall	F1-Score
Add	91.55	95.52	95.37	95.59
Concat	91.58	95.89	95.32	95.60
Product of Hamada	91.21	95.70	95.09	95.40
Grapth Attention (ours)	91.86	95.47	95.72	95.73

5.3. Model Complexity Analysis

To validate the excellent performance of our proposed LDFormer network from multiple perspectives, we performed model parametric counts and IoU comparisons with the current SOTA approach on the Inria construction dataset, and the results are shown in Figure 11. Specifically, we compared UNet, DeepLabV3+, HRNet, Swin-UNet, STT, Segmenter, Trans-UNet, and BuildFormer networks. The results show that our proposed LDFormer based on the best accuracy of 82.87% IoU has fewer overall parameters than the advanced Segmenter, Trans-UNet, and BuildFormer in the equivalent transformer structure. Compared with the lightweight convolutional UNet, DeeplabV3+, and STT methods, LDFormer brings a substantial enhancement in accuracy and carries a certain number of tolerable parameters.

In the proposed LDFormer, the major computational overhead is brought about by the transformer computational paradigm in LDSA and the large kernel convolution in DWLSK-MLP. The computational complexity of the LDSA is increasing in square steps due to the presence of the multiscale QKV index weighting, which introduces a larger number of parametric and computational complexity to the model. The DWLSK-MLP, although the separating convolution is used to decompose the large kernel convolution step by step,

still introduces a certain amount of trainable parameter sizes to the model compared to the non-introduction of large kernel convolution. Overall, although the two methods introduce a certain amount of parameters and computational complexity, the number of parameters introduced is acceptable compared to the accuracy gain it enables.

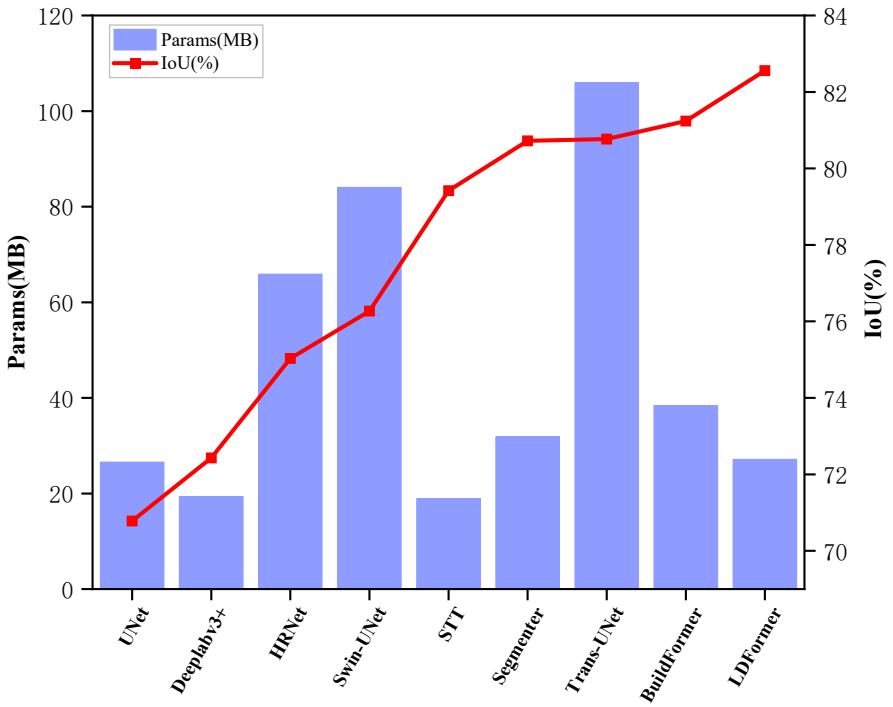


Figure 11. Model complexity comparison of LDFormer on Inria dataset.

5.4. Limitations and Future Work

While the linear decay-based CNN–transformer model, LDFormer, achieves promising results in building extraction, several limitations remain. Firstly, the model's adaptability to the heterogeneity and complexity of building appearances and backgrounds needs improvement. Factors such as image quality, lighting variations, and occlusions can affect segmentation performance, necessitating a further enhancement of the model's robustness across diverse environments. Secondly, despite controlling the computational load, LDFormer may encounter challenges in processing high-resolution images in real time. Future research should focus on optimizing the model architecture to utilize lighter network structures, enhancing its suitability for real-time applications. Lastly, as this study primarily relies on specific datasets, future work should assess the model's effectiveness on a wider variety of remote sensing imagery and scenes to evaluate its generalization capabilities for practical use.

6. Conclusions

In this paper, we have proposed a linear decay-based remote sensing building extraction model, LDFormer. The model initially addresses the critical roles of local and global contextual information in remote sensing building extraction through the design of the MDFB module. MDFB effectively extracts rich local detail features and adaptively fuses multi-scale information, significantly enhancing the model's feature representation capabilities. In order to further improve the model's ability to capture global features, we introduced LDSA and DWLSK-MLP in the transformer-based decoder. These innovations not only strengthen the model's long-range dependencies but also expand its receptive field while effectively controlling computational demands and parameter burdens. Extensive experiments were conducted on the Massachusetts, Inria, and WHU public building datasets, achieving IoU scores of 76.10%, 82.87%, and 91.86%, respectively. The experimental results

demonstrate that LDFormer effectively enhances segmentation accuracy for buildings of varying scales in complex backgrounds, proving its superior performance and advanced capabilities in the task of automated building extraction.

Author Contributions: Conceptualization, Z.G. and T.W.; methodology, Z.G., Y.G. and T.W.; software, Z.G.; validation, A.T. and Q.Z.; formal analysis, A.T. and Q.Z; investigation, Z.G., A.T. and Q.Z.; resources, Q.Z.; writing—original draft preparation, Z.G. and A.T.; writing—review and editing, T.W., A.T. and Q.Z.; visualization, A.T.; supervision, T.W.; project administration, T.W. and Q.Z.; funding acquisition, T.W. and Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under the title of "Theoretical Research on Convolutional Neural Networks for Automatic Generation of BIM Models from Point Clouds" (Project No. 51968051). National Natural Science Foundation of China, 42261070.

Data Availability Statement: The source code is available at https://github.com/Gongzhuyi/LDFormer, (accessed on 20 September 2024).

Acknowledgments: The authors thank the anonymous reviewers for the helpful comments that improved this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Qian, Z.; Chen, M.; Sun, Z.; Zhang, F.; Xu, Q.; Guo, J.; Xie, Z.; Zhang, Z. Simultaneous Extraction of Spatial and Attributional Building Information across Large-Scale Urban Landscapes from High-Resolution Satellite Imagery. *Sustain. Cities Soc.* **2024**, *106*, 105393. [CrossRef]
2. Liu, K.; Huang, J.; Xu, M.; Perc, M.; Li, X. Density Saliency for Clustered Building Detection and Population Capacity Estimation. *Neurocomputing* **2021**, *458*, 127–140. [CrossRef]
3. Dong, L.; Shan, J. A Comprehensive Review of Earthquake-Induced Building Damage Detection with Remote Sensing Techniques. *ISPRS J. Photogramm. Remote Sens.* **2013**, *84*, 85–99. [CrossRef]
4. Dornaika, F.; Moujahid, A.; El Merabet, Y.; Ruichek, Y. Building Detection from Orthophotos Using a Machine Learning Approach: An Empirical Study on Image Segmentation and Descriptors. *Expert Syst. Appl.* **2016**, *58*, 130–142. [CrossRef]
5. Turker, M.; Koc-San, D. 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 Obs. Geoinf.* **2015**, *34*, 58–69. [CrossRef]
6. Mountrakis, G.; Im, J.; Ogole, C. Support Vector Machines in Remote Sensing: A Review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259. [CrossRef]
7. Huang, X.; Chen, H.; Gong, J. Angular Difference Feature Extraction for Urban Scene Classification Using ZY-3 Multi-Angle High-Resolution Satellite Imagery. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 127–141. [CrossRef]
8. Srivastava, V.; Avudaiammal, R.; V George, S. Investigations on Extraction of Buildings from RS Imagery Using Deep Learning Models. *Int. J. Remote Sens.* **2024**, *45*, 68–100. [CrossRef]
9. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
10. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
11. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
12. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
13. Chen, Y.; Tang, L.; Yang, X.; Bilal, M.; Li, Q. Object-Based Multi-Modal Convolution Neural Networks for Building Extraction Using Panchromatic and Multispectral Imagery. *Neurocomputing* **2020**, *386*, 136–146. [CrossRef]
14. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
15. Li, W.; Sun, K.; Zhao, H.; Li, W.; Wei, J.; Gao, S. Extracting buildings from high-resolution remote sensing images by deep ConvNets equipped with structural-cue-guided feature alignment. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102970. [CrossRef]
16. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
17. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer: Cham, Switzerland, 2015; pp. 234–241.

18. Yang, J.; Matsushita, B.; Zhang, H. Improving Building Rooftop Segmentation Accuracy through the Optimization of UNet Basic Elements and Image Foreground-Background Balance. *ISPRS J. Photogramm. Remote Sens.* **2023**, *201*, 123–137. [CrossRef]
19. ABCNet: Attentive Bilateral Contextual Network for Efficient Semantic Segmentation of Fine-Resolution Remotely Sensed Imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 84–98. [CrossRef]
20. Tejeswari, B.; Sharma, S.K.; Kumar, M.; Gupta, K. Building footprint extraction from space-borne imagery using deep neural networks. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *43*, 641–647. [CrossRef]
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: New York, NY, USA, 2017; Volume 30.
22. Wang, L.; Fang, S.; Meng, X.; Li, R. Building Extraction With Vision Transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. [CrossRef]
23. Hu, Y.; Wang, Z.; Huang, Z.; Liu, Y. PolyBuilding: Polygon Transformer for Building Extraction. *ISPRS J. Photogramm. Remote Sens.* **2023**, *199*, 15–27. [CrossRef]
24. Chen, Z.; Luo, Y.; Wang, J.; Li, J.; Wang, C.; Li, D. DPENet: Dual-path Extraction Network Based on CNN and Transformer for Accurate Building and Road Extraction. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103510. [CrossRef]
25. Xu, L.; Li, Y.; Xu, J.; Zhang, Y.; Guo, L. BCTNet: Bi-branch cross-fusion transformer for building footprint extraction. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [CrossRef]
26. Li, Z.; Wang, Y.; Tian, F.; Zhang, J.; Chen, Y.; Li, K. BAFormer: A Novel Boundary-Aware Compensation UNet-like Transformer for High-Resolution Cropland Extraction. *Remote Sens.* **2024**, *16*, 2526. [CrossRef]
27. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 574–586. [CrossRef]
28. Wei, S.; Ji, S.; Lu, M. Toward Automatic Building Footprint Delineation From Aerial Images Using CNN and Regularization. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2178–2189. [CrossRef]
29. Chen, S.; Shi, W.; Zhou, M.; Zhang, M.; Xuan, Z. CGSANet: A Contour-Guided and Local Structure-Aware Encoder–Decoder Network for Accurate Building Extraction From Very High-Resolution Remote Sensing Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1526–1542. [CrossRef]
30. Qiu, W.; Gu, L.; Gao, F.; Jiang, T. Building Extraction From Very High-Resolution Remote Sensing Images Using Refine-UNet. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
31. Li, Y.; Lu, H.; Liu, Q.; Zhang, Y.; Liu, X. SSDBN: A Single-Side Dual-Branch Network with Encoder–Decoder for Building Extraction. *Remote Sens.* **2022**, *14*, 768. [CrossRef]
32. You, D.; Wang, S.; Wang, F.; Zhou, Y.; Wang, Z.; Wang, J.; Xiong, Y. EfficientUNet+: A Building Extraction Method for Emergency Shelters Based on Deep Learning. *Remote Sens.* **2022**, *14*, 2207. [CrossRef]
33. He, X.; Chen, Y.; Lin, Z. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 498. [CrossRef]
34. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure transformer network for remote sensing image change detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. [CrossRef]
35. Jiang, X.; Li, Y.; Jiang, T.; Xie, J.; Wu, Y.; Cai, Q.; Jiang, J.; Xu, J.; Zhang, H. RoadFormer: Pyramidal Deformable Vision Transformers for Road Network Extraction with Remote Sensing Images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 102987. [CrossRef]
36. Deng, F.; Luo, W.; Ni, Y.; Wang, X.; Wang, Y.; Zhang, G. UMiT-Net: A U-Shaped Mix-Transformer Network for Extracting Precise Roads Using Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–13. [CrossRef]
37. Li, Q.; Zhong, R.; Du, X.; Du, Y. TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–19. [CrossRef]
38. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A. Vision Transformers for Remote Sensing Image Classification. *Remote Sens.* **2021**, *13*, 516. [CrossRef]
39. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like Transformer for Efficient Semantic Segmentation of Remote Sensing Urban Scene Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
40. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer Meets Convolution: A Bilateral Awareness Network for Semantic Segmentation of Very Fine Resolution Urban Scene Images. *Remote Sens.* **2021**, *13*, 3065. [CrossRef]
41. Chen, K.; Zou, Z.; Shi, Z. Building Extraction from Remote Sensing Images with Sparse Token Transformers. *Remote Sens.* **2021**, *13*, 4441. [CrossRef]
42. Yuan, W.; Xu, W. MSST-Net: A multi-scale adaptive network for building extraction from remote sensing images based on swin transformer. *Remote Sens.* **2021**, *13*, 4743. [CrossRef]
43. Yuan, W.; Zhang, X.; Shi, J.; Wang, J. LiteST-Net: A Hybrid Model of Lite Swin Transformer and Convolution for Building Extraction from Remote Sensing Image. *Remote Sens.* **2023**, *15*, 1996. [CrossRef]
44. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4–24. [CrossRef] [PubMed]
45. Diao, Q.; Dai, Y.; Zhang, C.; Wu, Y.; Feng, X.; Pan, F. Superpixel-Based Attention Graph Neural Network for Semantic Segmentation in Aerial Images. *Remote Sens.* **2022**, *14*, 305. [CrossRef]

46. Cui, W.; He, X.; Yao, M.; Wang, Z.; Hao, Y.; Li, J.; Wu, W.; Zhao, H.; Xia, C.; Li, J.; et al. Knowledge and Spatial Pyramid Distance-Based Gated Graph Attention Network for Remote Sensing Semantic Segmentation. *Remote Sens.* **2021**, *13*, 1312. [CrossRef]
47. Jia, M.; Zhang, C.; Zhao, Z.; Wang, L. Bipartite Graph Attention Autoencoders for Unsupervised Change Detection Using VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
48. Zi, W.; Xiong, W.; Chen, H.; Li, J.; Jing, N. SGA-Net: Self-Constructing Graph Attention Neural Network for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2021**, *13*, 4201. [CrossRef]
49. Deng, R.; Guo, Z.; Chen, Q.; Sun, X.; Chen, Q.; Wang, H.; Liu, X. A Dual Spatial-Graph Refinement Network for Building Extraction From Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–20. [CrossRef]
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
51. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
52. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. Coatnet: Marrying convolution and attention for all data sizes. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–14 December 2021; Volume 34, pp. 3965–3977.
53. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12175–12185.
54. Lou, M.; Zhou, H.Y.; Yang, S.; Yu, Y. TransXNet: Learning both global and local dynamics with a dual dynamic token mixer for visual recognition. *arXiv* **2023**, arXiv:2310.19380.
55. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
56. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
57. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
58. Shao, Z.; Tang, P.; Wang, Z.; Saleem, N.; Yam, S.; Sommai, C. BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. *Remote Sens.* **2020**, *12*, 1050. [CrossRef]
59. Chen, Z.; Li, D.; Fan, W.; Guan, H.; Wang, C.; Li, J. Self-attention in reconstruction bias U-Net for semantic segmentation of building rooftops in optical remote sensing images. *Remote Sens.* **2021**, *13*, 2524. [CrossRef]
60. Guo, H.; Su, X.; Wu, C.; Du, B.; Zhang, L. Decoupling semantic and edge representations for building footprint extraction from remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–16. [CrossRef]
61. Guo, H.; Du, B.; Zhang, L.; Su, X. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *183*, 240–252. [CrossRef]
62. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef] [PubMed]
63. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 205–218.
64. Liu, P.; Liu, X.; Liu, M.; Shi, Q.; Yang, J.; Xu, X.; Zhang, Y. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. *Remote Sens.* **2019**, *11*, 830. [CrossRef]
65. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7262–7272.
66. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
67. Wu, G.; Shao, X.; Guo, Z.; Chen, Q.; Yuan, W.; Shi, X.; Xu, Y.; Shibasaki, R. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. *Remote Sens.* **2018**, *10*, 407. [CrossRef]
68. Liu, H.; Luo, J.; Huang, B.; Hu, X.; Sun, Y.; Yang, Y.; Xu, N.; Zhou, N. DE-Net: Deep encoding network for building extraction from high-resolution remote sensing imagery. *Remote Sens.* **2019**, *11*, 2380. [CrossRef]
69. Kang, W.; Xiang, Y.; Wang, F.; You, H. EU-Net: An efficient fully convolutional network for building extraction from optical remote sensing images. *Remote Sens.* **2019**, *11*, 2813. [CrossRef]
70. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6169–6181. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.