electronics

MDPI

Article

MSTrans: Multi-Scale Transformer for Building Extraction from HR Remote Sensing Images

Fei Yang [1], Fenlong Jiang [2,*], Jianzhao Li [3,4] and Lei Lu [1]

[1] School of Information Engineering, Yulin University, Yulin 719000, China; yangfei@yulinu.edu.cn (F.Y.); lulei@yulinu.edu.cn (L.L.)
[2] Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, School of Computer Science and Technology, Xidian University, Xi'an 710071, China
[3] Guangzhou Institute of Technology, Xidian University, Guangzhou 510555, China; lijianzhao@xidian.edu.cn
[4] Key Laboratory of Collaborative Intelligence Systems, Ministry of Education, Xidian University, Xi'an 710071, China
* Correspondence: fljiang@xidian.edu.cn or jiangfenlong@outlook.com

Abstract: Buildings are one of the most important goals of human transformation of the Earth's surface. Therefore, building extraction (BE), such as in urban resource management and planning, is a task that is meaningful to actual production and life. Computational intelligence techniques based on convolutional neural networks (CNNs) and Transformers have begun to be of interest in BE, and have made some progress. However, the BE methods based on CNNs are limited by the difficulty in capturing global long-range relationships, while Transformer-based methods are often not detailed enough for pixel-level annotation tasks because they focus on global information. To conquer the limitations, a multi-scale Transformer (MSTrans) is proposed for BE from high-resolution remote sensing images. In the proposed MSTrans, we develop a plug-and-play multi-scale Transformer (MST) module based on atrous spatial pyramid pooling (ASPP). The MST module can effectively capture tokens of different scales through the Transformer encoder and Transformer decoder. This can enhance multi-scale feature extraction of buildings, thereby improving the BE performance. Experiments on three real and challenging BE datasets verify the effectiveness of the proposed MSTrans. While the proposed approach may not achieve the highest Precision and Recall accuracies compared with the seven benchmark methods, it improves the overall metrics F1 and mIoU by 0.4% and 1.67%, respectively.

Keywords: building extraction; computational intelligence; multi-scale Transformer; atrous convolution; remote sensing images

Citation: Yang, F.; Jiang, F.; Li, J.; Lu, L. MSTrans: Multi-Scale Transformer for Building Extraction from HR Remote Sensing Images. *Electronics* **2024**, *13*, 4610. https://doi.org/10.3390/electronics13234610

Academic Editor: Domenico Ursino

Received: 11 October 2024
Revised: 9 November 2024
Accepted: 20 November 2024
Published: 22 November 2024

1. Introduction

With the rapid development of remote sensing technology, high-resolution (HR) and very-high-resolution (VHR) remote sensing images have been popularized in practical applications. In particular, many scholars have focused on the study of building extraction (BE) and building change detection algorithms in urban scenes [1–3]. As one of the most valuable Earth surface targets, effective BE can be used for dynamic assessment and monitoring of urban development, urban disaster assessment and monitoring, and urban management and construction [4–6].

BE has received much attention in recent decades, and many methods have been proposed. Buildings usually have special characteristics such as spectrum, texture, context, and shape [7,8]. In this context, early BE methods focused on introducing building auxiliary information or designing artificially crafted building features for BE. The classical algorithms for BE algorithms based on building-assisted features are terrain and surface models [9], airborne laser scanning data [10], and digital elevation models [11,12]. In the early stage, the land cover classification result is adopted to acquire candidate building objects by

approximate location and shape in [13], thereby reaching BE. Afterwards, the representative algorithms for BE algorithms based on artificially crafted features are the pixel shape index [14], building index [15], morphological building/shadow index [16], building contours and color features [17], etc.

Over the years, BE methods have made great progress. Since the advent and application of deep learning, it has become the most mainstream method of BE. The existing BE approaches can be broadly summarized into two categories: CNN-based and Transformer-based BE methods.

In the last decade, the development and application of deep learning techniques have brought more new solutions for BE and have been greatly improved [18,19]. BE based on deep learning can utilize convolutional neural networks (CNNs) to extract buildings by segmenting their roofs. Specifically, BE can be viewed as a single-target semantic segmentation task. Hence, conventional end-to-end segmentation networks, such as U-Net [20], SegNet [21], Deeplab [22], etc., can be exploited to identify buildings. However, conventional segmentation networks still have low accuracy and completeness in extracting buildings of various scales due to their insufficient ability to perceive multi-scale features of buildings. Therefore, many networks dedicated to BE have been proposed in a steady stream. For example, Ji et al. released a BE dataset, and provided a Siamese U-Net (SiU-Net) for BE [23]. A new network was designed for BE, named Res2Unet, which aims to aims to enhance the network's ability to extract features of buildings of different scales, thereby improving BE performance [24]. Similar methods can also be found in [25–27]. These multi-scale techniques can significantly boost the accuracy of complex multi-scale BE. With the rise of attention mechanisms, BE networks based on attention mechanisms have attracted attention and promotion, and provide a new solution for multi-scale building extraction. Researchers have comprehensively developed a BE network based on hybrid pyramid and attention, which further strengthens multi-scale feature extraction and reduces the interference of background non-building targets, thereby significantly improving BE accuracy. Some methods combine different pyramid structures and attention mechanisms for BE to adapt to different multi-scale building targets, such as BRRNet [28] and AGPNet [29]. Other networks can be found in [30]. These approaches further improve the BE extraction performance by combining pyramid and attention modules.

Nevertheless, many studies that have presented the meaningful features of buildings, such as edges or outlines, have not been effectively utilized. To this end, some novel models target building edge and contour representations to improve BE capabilities, e.g., CFENet [31] and CBRNet [32]. In [33], multi-task loss is promoted, which focuses on the accurately detecting of pixels near building boundaries. The above edge- or contour-based methods can obtain more accurate BE results. Subsequently, some methods improve the accurate detection of building contours by enhancing edges. In [34], Chen et al. devised a contour-guided end-to-end network for BE through considering local structure. Zhu et al. designed an edge-detail-network in [35], which is composed of an edge subnetwork and a detail subnetwork for elevating the performance of building edges. There are also other related methods [36,37]. Some scholars regard the edges of buildings as high-frequency information in the frequency domain. To this end, they have enhanced the detail information of buildings by enhancing different frequency components in the frequency domain, thus improving the accuracy of pixel detection near the edges of buildings, such as [38–40]. In addition, some scholars also focus on using instance segmentation for effective building extraction [41,42]. For instance, Wen et al. devised a deep instance segmentation network based on Mask R-CNN to realize building instance extraction [43]. Wu et al. applied an improved anchor-free segmentation network in [44], which can achieve individual building instance extraction. Although these methods can extract building instance targets, they are unable to acquire the detailed shape, edge, and other information of the building well.

Recently, Transformer-based networks have been proposed to model long-range dependencies between global buildings, which can help improve BE performance. For instance,

in [45], a dual-path vision Transformer was designed to extract global spatial contextual features for fine-grained BE. Zhang et al. adopted a spatial attention Transformer network in [46], which fuses the local and global features by introducing local and global attention paths. In [8], a named Fusion-Former is proposed based on self-attention with depth-wise convolution for multi-scale feature extraction of BE. And Mask2Former and a shape-aware enhancement Transformer are proposed in [47,48]. The authors exploit Transformers with shifted windows and shape information to capture global information and detailed context information to improve building extraction performance. Some scholars have also proposed shape-based perception methods to extract buildings [47,49,50]. In addition, some studies have also introduced cross-modal images to assist traditional optical images to improve the accuracy of BE [51,52]. Notably, a multi-scale feature-based Transformer has been proposed to further improve the performance of BE. In [53], a hierarchical vision Transformer based on shifted windows is presented to capture multi-scale features for BE. Recently, Chang et al. developed a multi-scale attention network for BE in [54], which employs the multi-scale channel attention and spatial attention mechanism to extract multi-scale building features. Overall, the aforementioned approaches have greatly developed BE and further increased the practicality of the algorithm. Although the aforementioned research has made good progress, there are still some challenges that need to be continuously studied and resolved. Firstly, multi-source remote sensing images may show different spectral, texture, shape, and scale characteristics of buildings in remote sensing images from different sensors due to various imaging conditions such as altitude, weather, lighting, cloud thickness, season, etc. Therefore, in complex multi-sensor remote sensing images, the robustness and generalization of the current networks still need to be further explored and enhanced. Secondly, in BE tasks, the vision Transformer network often focuses on global features, which facilely brings the loss of spatial information and it cannot handle fine-grained pixel-level annotation tasks well. Thirdly, the existing BE methods based on vision Transformers mostly use Transformers to extract features at a single scale, while ignoring global features or long-range dependencies at different scales. In order to overcome these problems, some scholars recently proposed some CNN and Transformer hybrid networks to capture and aggregate local and global features for BE [55,56]. For instance, ref. [56] proposed an asymmetric network for BE, which can combine CNNs and Transformers to extract and fuse detailed information and long-dependencies relationships. Yuan et al. developed a CNN and Transformer hybrid network based a modified multi-head self-attention mechanism in [57], which can further explore multi-scale feature extraction of buildings. In summary, these recent hybrid CNN–Transformer networks are an effective solution for BE.

Facing the above challenges, this study proposes a multi-scale Transformer network (MSTrans) for BE. The motivation of our proposed MSTrans is summarized in the following two points. For one thing, although the pure vision Transformer can capture global building features, it is prone to missing spatial detail locations, which may lead to limited detail detection performance for pixel-level fine-grained annotation tasks. Therefore, introducing local features and effectively coupling local and global features are beneficial to improve BE detection performance. For another thing, the existing BE methods based on vision Transformer often employ single-scale tokens to capture global features, while ignoring the positive impact of tokens of different scales on modeling long-range relationships between buildings of various scales. Therefore, from the perspective of different scales, effectively using tokens of different scales to mine multi-scale global features is conducive to improving BE multi-scale feature extraction capabilities.

According to the above motivations, we propose MSTrans for BE from VHR remote sensing images. The major contributions of this study are summarized as follows:

(1) We propose a novel end-to-end multi-scale Transformer network for BE, which is a hybrid convolutional and Transformer network. The proposed MSTrans can effectively extract and aggregate local and global features by fusing CNNs and Transformers.
(2) In the proposed MSTrans, an elegant and plug-and-play multi-scale Transformer (MST) module is introduced, which is effectively integrated with ASPP. The proposed MST

module can effectively represent tokens of different scales through the Transformer encoder and Transformer decoder, and it can enhance multi-scale feature extraction of buildings.

(3) Experimental results show that the proposed MSTrans reaches a competitive performance on three public and challenging BE datasets.

The rest of this study is organized as follows. Section 2 introduces the proposed MSTrans in detail. In Sections 3 and 4, experimental results and analysis are given and discussed. Finally, the conclusion and future works are generalized in Section 5.

2. Methodology

2.1. Overview

Based on the wide spread of the CNN–Transformer hybrid architecture in the remote sensing field [58], we employ it in the proposed multi-scale Transformer network (MSTrans) for precise BE. In most cases of CNN–Transformer hybrid architectures, there is a CNN encoder to firstly extract deep features of the input remote sensing data. Then the Transformer is employed to improve the extracted features with its self-attention mechanisms. Finally, another CNN architecture is adopted as the decoder to fuse the features to obtain final predictions of the networks. Similar designs are utilized in the proposed method, as shown in Figure 1.

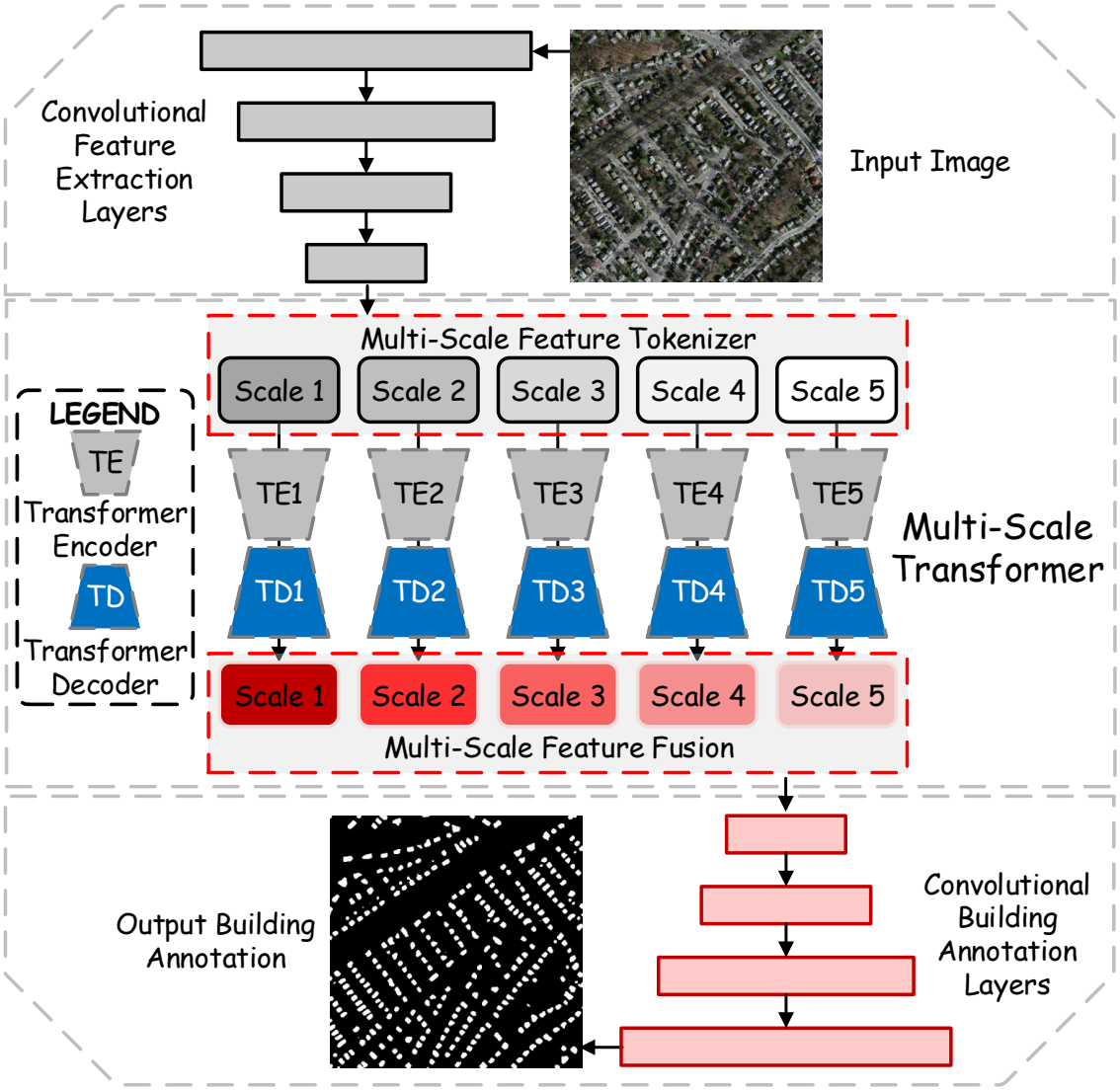

Figure 1. Overview of the proposed MSTrans BE framework.

We firstly use several convolutional feature extraction layers to extract building features from the input remotely sensed imagery. Then the feature maps from deep layers with rich contextual information are input into the proposed multi-scale Transformer module, which can improve the spatial recognition of building objects since the deep features usually lack location information. Different from other Transformer-based remote sensing image processing models, the proposed MSTrans utilizes several different branches of Transformers to acquire the multi-scale feature representation of buildings and refine the input deep features with both rich spatial information and contextual information, which has been validated as beneficial for the detection of buildings in the experiments. Given the refined deep features, the convolutional building annotation layers leverage the building information within the deep features to detect buildings and obtain the final binary building annotation map. Notably, the feature extraction layers and building annotation layers are connected by the skip connections proposed in [20], which can keep precise boundaries for detected buildings, thus further improving the BE performance.

As for detailed composition of the proposed model, the convolutional feature extraction layers are composed of convolutional layers with batch normalization (BN) and linear rectification functions. A max-pooling layer is also used in the feature extraction stage to build the multi-scale feature pyramid for better feature representation. The convolutional building annotation layers share a similar composition with the feature extraction layers. Different from feature extraction layers, the building annotation layers use bilinear upsampling layers instead of max-pooling layers. And we will subsequently introduce the detailed structure of the proposed MST module in the next sections.

2.2. Multi-Scale Transformer

Transformers have been commonly adopted in recent remote sensing applications. The multi-head self-attention mechanisms within can better uncover the contextual information of input features, which enrich the semantic information, thus further promoting the recognition of land cover objects with variant categories. BE tasks require better semantic segmentation performance, since the building objects share similar features with many other man-made land cover objects such as roads and squares. And the better semantic representation ability needed by BE can be offered by Transformer architectures. However, most existing BE Transformers uncover semantic information over a single scale of feature maps, which neglects the representation of multi-scale semantic information, thus diminishing the detection of multi-scale buildings. To address these problems, we firstly use different convolutional layers to extract multi-scale features and subsequently acquire the semantic tokens for different scales with the tokenizer. Then we utilize five non-Siamese Transformers with the same structure to enrich the semantic information of multi-scale feature maps. Finally, these multi-scale features are fused and input into building annotation layers, as shown in Figure 2, where the detailed procedure of the multi-scale feature tokenizer and multi-scale feature fusion can be seen.

2.2.1. Multi-Scale Feature Tokenizer

The proposed multi-scale feature tokenizer is conceived to obtain multi-scale feature tokens through extracted multi-scale features, thus enhancing the multi-scale semantic representation, which can be demonstrated in mathematical style as follows. Firstly, we use $I \in \mathbb{R}^{h \times w \times c}$ to denote the input feature maps of the proposed multi-scale feature tokenizer, where h, w, c denote the height, width, and channel size of the input features, respectively. Then, a set of modified convolutional layers derived from atrous pyramid pooling [22] are employed to capture multi-scale representation from the input feature maps, which elicits the feature maps of five different scales, as shown below:

$$I_1 = conv\_1x1(I) \tag{1}$$

$$I_2 = d\_conv\_6(I) \tag{2}$$

$$I_3 = d\_conv\_12(I) \tag{3}$$

$$I_4 = d\_conv\_18(I) \tag{4}$$

$$I_5 = g\_conv(I) \tag{5}$$

where $I_n \in \mathbb{R}^{h \times w \times (c/4)}$ $[n = 1, 2, 3, 4, 5]$ indicate the extracted feature maps containing representations of different scales. And $conv\_1 \times 1(\cdot)$ represents a convolutional layer with the kernel size of 1×1. The $d\_conv\_i(\cdot)[i = 6, 12, 18]$ denote the dilated convolution layer with the dilation rate of i. The $g\_conv(\cdot)$ comprises a global average pooling layer and a 1×1 convolutional layer. Notably, the aforementioned convolutional layers are followed by BN and rectified linear unit (ReLU). Then, the extracted multi-scale features are tokenized as follows:

$$T_n = Softmax(flatten(conv(I_n))) \odot flatten(I_n)^T \tag{6}$$

where $conv(\cdot)$ is an 1×1 convolution layer that changes the channel size of I_n to the token length of 4, the $flatten(\cdot)$ flattens the spatial dimension of the input feature maps, and $\odot$ represents the matrix multiplication.

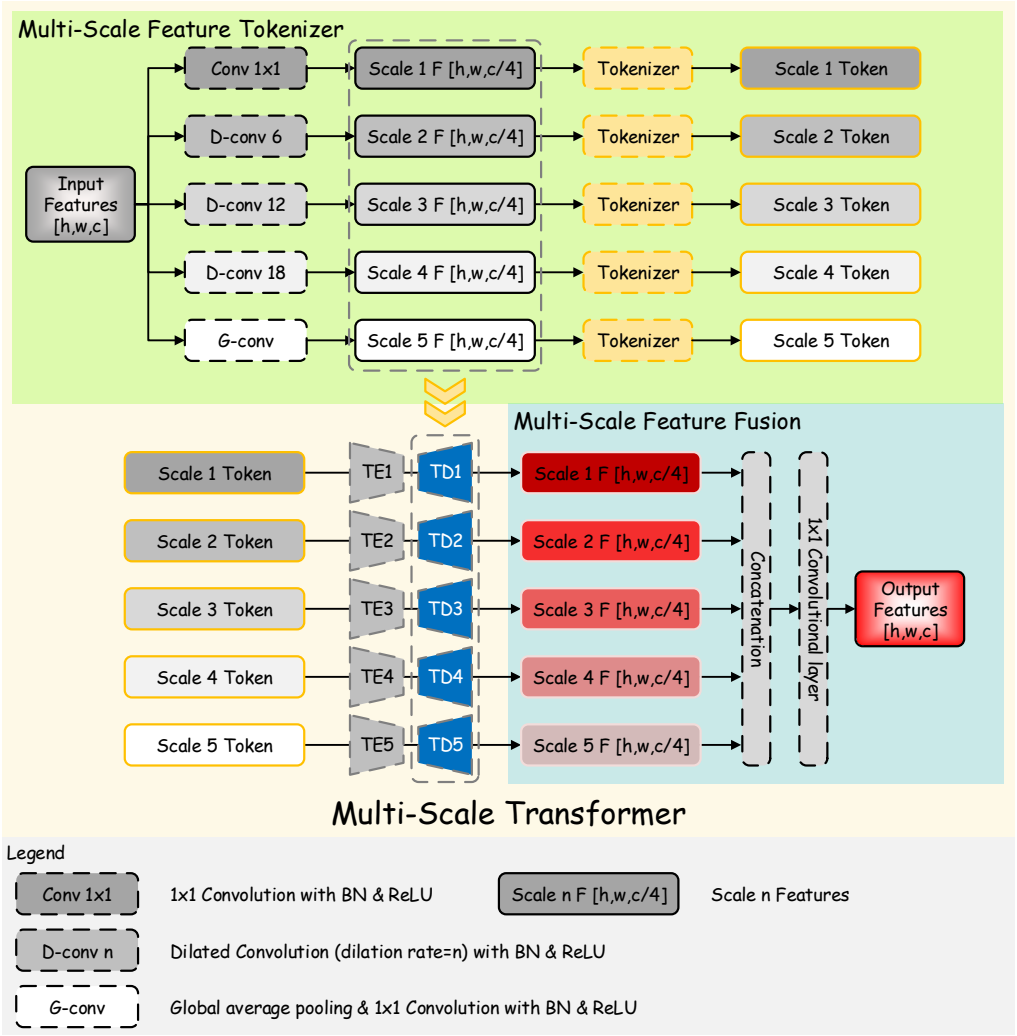

Figure 2. The architecture of the MST module.

2.2.2. Multi-Scale Feature Fusion

After the multi-scale semantic information is enhanced by several parallel Transformers, the multi-scale feature maps are fused and further refined to integrate the multi-scale semantic representation, which can be shown as follows:

$$O = conv\_output(Concat[O_1, O_2, O_3, O_4, O_5]) \tag{7}$$

where $O_n \in \mathbb{R}^{h \times w \times (c/4)}$ $[n = 1, 2, 3, 4, 5]$ and $O \in \mathbb{R}^{h \times w \times c}$ denote the enhanced multi-scale features and the output features, respectively. And $Concat(\cdot)$ and $conv\_output(\cdot)$ represent the concatenation in channel dimension and a convolutional layer with the kernel size of 1×1, respectively. Given the detailed information of the proposed MST module, the complete procedure of the MST module can be clearly demonstrated in Algorithm 1.

Algorithm 1 Procedure of Multi-Scale Transformer

Input:
Input feature maps $I \in \mathbb{R}^{h \times w \times c}$;
Output:
Output feature maps $O \in \mathbb{R}^{h \times w \times c}$;

Acquire multi-scale features I_n $(n = 1, 2, 3, 4, 5)$ with the convolutional layers as follows:
I_n = Conv_Layer_$n(I)$;
Extract multi-scale semantic tokens the tokenizer as follows:
T_n = Tokenizer(I_n);
Acquire rich semantic multi-scale representation as follows:
O_n = Transformer_decoder(Transformer_encoder(T_n),I_n);
Fuse the enhanced multi-scale semantic information as follows:
O = Conv_output([O_1,O_2,O_3,O_4,O_5]);
return Output feature maps O;

Through the proposed MSTrans, we firstly acquire the multi-scale representation of the input features through the multi-scale feature tokenizer. Then, these Transformer-enhanced multi-scale feature maps are fused and input to the following convolutional building annotation layers, which can better perceive the multi-scale semantic building information within.

3. Experiments and Results

In this section, to test the performance of the proposed MSTrans, we perform a series of experiments and analyses on three public datasets and challenge datasets. First, we briefly describe the datasets for the three experimental datasets. Then, we provide the benchmark methods and their implementation details for comparison with the proposed approach. Subsequently, quantitative and visual results of different methods are analyzed for comparison. Finally, we carry out ablation studies and feature visualization analysis of the proposed MST module based on the Massachusetts and EastAsia datasets, respectively.

3.1. Dataset Descriptions

In the experiments, we selected three extensive used and challenging BE datasets, namely the Massachusetts, EastAsia, and Inria datasets, as shown in Figure 3. The detailed information of these datasets is listed in Table 1. All preprocessing strategies such as cropping and partitioning of datasets are based on [59]. The relevant descriptions are as follows.

The Massachusetts BE dataset is a set of VHR aerial images collected in the Boston area. As presented in Figure 3a, different buildings exhibit different shapes, structures, textures, and spatial distributions. The main challenge of the Massachusetts BE dataset is that it contains a large number of densely packed tiny buildings as well as many large buildings. This requires the BE network to be able to adapt to the capabilities of the multi-scale target,

especially when tiny buildings and irregular large targets coexist in the scene. Hence, in our experiments, the Massachusetts BE data are selected to test the performance of the proposed method on buildings of different scales (tiny and large).

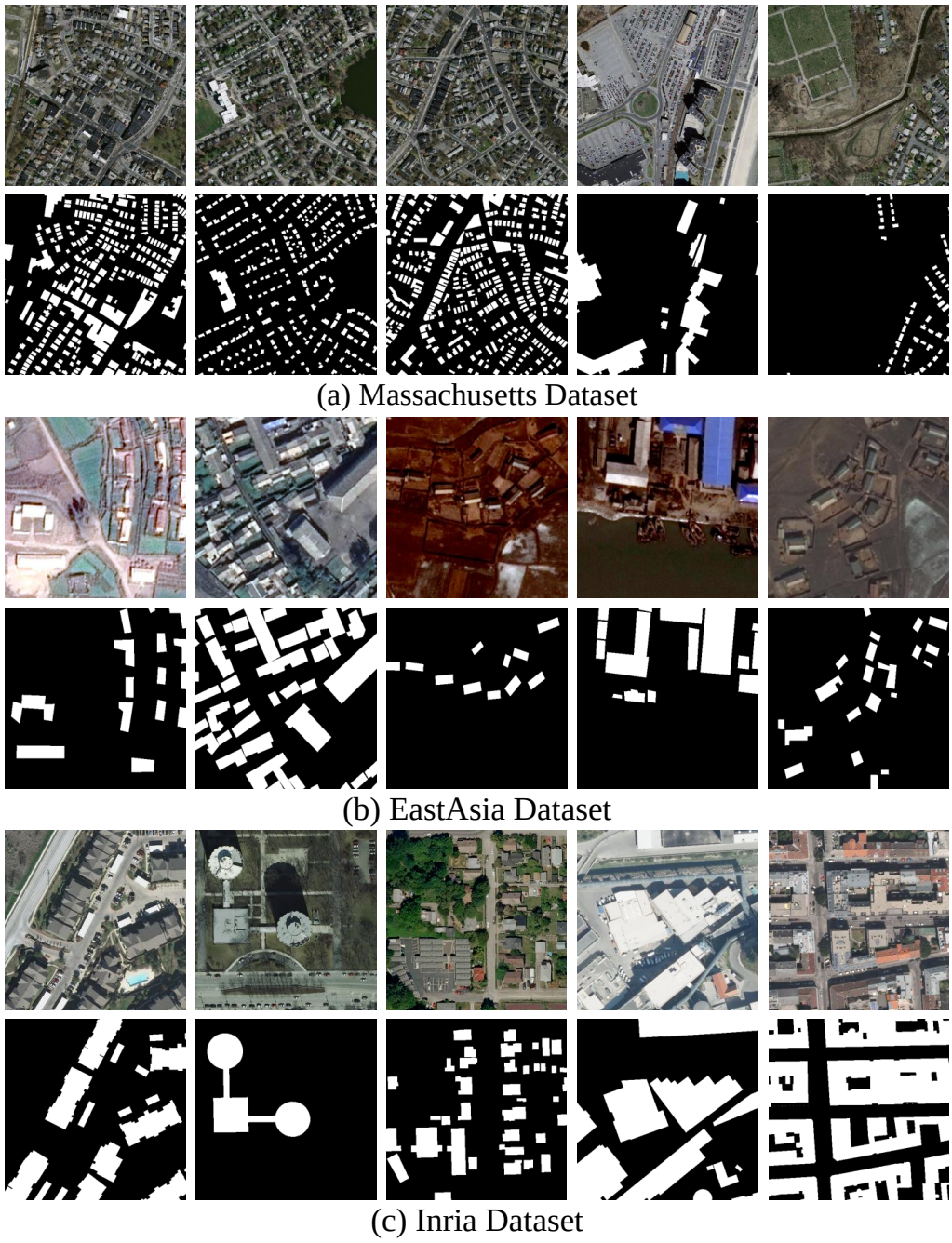

Figure 3. Dataset presentations: (**a**) WHU-CD dataset; (**b**) LEVIR-CD dataset; (**c**) GZ-CD dataset.

Table 1. Experimental datasets description in detail.

Datasets	References	Sensors	Spatial Resolution	Size	Training/Test Samples
Massachusetts	Ji et al. [60]	Aerial	1 m/pixel	512 × 512	548/40
EastAsia	Chen et al. [23]	Satellite	2.7 m/pixel	512 × 512	3153/903
Inria	Peng et al. [61]	Aerial	0.3 m/pixel	512 × 512	2025/891

The EastAsia BE dataset is a set of HR satellite images collected in East Asia. As presented in Figure 3b, buildings in different locations present completely different spectral information. Moreover, buildings and backgrounds are extremely similar, which may cause the accuracy of BE to be limited by the background and other objects. Therefore, this BE dataset has great challenges, which requires the BE network to have the ability to extract robust building features and be able to perceive long-range dependencies such as topology between different buildings.

The Inria BE dataset is a set of VHR aerial images captured in five different urban settles. The five sub-datasets named Austin, Chicago, Kitsap, Tirol, and Vienna make up the Inria BE dataset. Referring to [59], the five sub-datasets are divided into 2025 training samples and 891 testing samples, respectively. As shown in Figure 3c, different sub-datasets may have significant differences in the materials, scales, and shapes of buildings due to the different cities where they were shot. As a result, we chose the Inria BE dataset to verify the robustness and generalization of the proposed approach.

To sum up, the reasons why we selected three sets of BE datasets to verify the performance of different methods mainly include the following aspects. On the one hand, these three sets of data were acquired from different sensors and imaging conditions, including two datasets based on aerial images and one satellite image. On the other hand, the buildings of different scales covered by the different datasets can effectively verify the ability of different methods to extract multi-scale buildings.

3.2. Comparative Approaches and Evaluation Indicators

In this subsection, the comparative approaches and evaluation indicators in the experiments are given. The details are as follows.

3.2.1. Comparative Approaches

- U-Net [20]: this network is a classic semantic segmentation network and also the benchmark network for many BE networks. Here, we choose this network as a base network for comparative analysis.
- SiU-Net [23]: the method is the benchmark model for the EastAsia dataset. It introduces a Siamese branch based on U-Net, which enhances the model's ability to adapt and extract multi-scale building features through a downsampling generation image.
- Res2Unet [24]: this method is a novel deep building detection network that solves the problems of missed detection of small buildings and mixing of building boundaries and background through hierarchical fine-grained multi-scale learning. Therefore, it is necessary to choose this method to compare with the proposed method.
- CFENet [31]: the model is a contextual feature representation focused on buildings by designing a spatial fusion and focus enhancement module. Its purpose is to improve the difficulty of low-level and high-level feature representation of complex and diverse buildings. The challenges of our experimental dataset are the same as the purpose of this method, so this method is selected as a comparison method.
- CBRNet [32]: the network is a boundary refinement network that introduces a boundary refinement module and a coarse-to-fine learning strategy to alleviate false and missed detections caused by small buildings, tree occlusions, and shadow interference.
- BBRNet [28]: the approach focuses on solving the incomplete detection of large-scale buildings and inaccurate detection of complex-shaped buildings by designing residual refinement and prediction modules.
- AGPNet [29]: the network combines attention and pyramid modules to alleviate the limitation that a single receptive field cannot extract multi-scale building features well. Its motivation is similar to that of our proposed model, so it is valuable to choose it as a benchmark comparison method.

3.2.2. Evaluation Indicators

In all experiments, we selected four evaluation indicators, Precision (Pre), Recall (Rec), F1-Score (F1), and mean intersection over union (mIoU), to compare and analyze the performance of the proposed MSTrans and other benchmark models. These indicators can be calculated by a binary confusion matrix consisting of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). They have been widely used in binary classification tasks such as change detection. Please refer to [31,32,59] as this article will not repeat them here.

3.3. Implementation Details

In the experiment, to compare the performance of the models fairly and objectively, and we did not employ any preprocessing or pretraining parameters for the models. For all comparison network deployments and results please refer to [59] to ensure the reliability of the results. In addition, implementation details of the proposed MSTrans are as follows. We employed the PyTorch framework with CUDA 11.3 to deploy the proposed MSTrans, and the MSTrans model was optimized using the Adam optimizer with an initial learning rate of 0.0001. In addition, the network was trained using a single NVIDIA 3090 graphics card in the experiment. Notably, to adapt to the video memory, we set the batch size for training and inference to 4 in our experiments.

3.4. Comparison with Other Approaches

3.4.1. Results on Massachusetts Dataset

As shown in Table 2 and Figure 4, we obtained the quantitative accuracy and visual results of the proposed MSTrans and the compared benchmark models on the Massachusetts BE dataset. The accuracy comparison listed in Table 2 shows that the proposed MSTrans achieves the best accuracy in two comprehensive evaluation indicators, F1 and mIoU, compared with the other seven baseline methods. Although our method does not achieve the best in the Pre and Rec indicators, it is more balanced compared with other methods. Visual comparison results also show the same conclusion, as presented in Figure 4. In our MSTrans detection results (as shown in Figure 4j), the TP pixels in white are obviously the most abundant, while the FN and FP pixels in green and red are relatively few. Overall, the proposed MSTrans network demonstrates good detection capability for tiny buildings.

Table 2. Comparison of the quantitative accuracy (in %) of different models on the Massachusetts BE dataset. **Bold** is the best accuracy.

Methods	Pre (%)	Rec (%)	F1 (%)	mIoU (%)
U-Net [20]	**88.66**	72.19	79.58	79.52
SiU-Net [23]	84.82	75.80	80.06	79.76
Res2Unet [24]	81.04	65.65	72.54	73.79
CFENet [31]	73.48	63.67	68.23	70.34
CBRNet [32]	64.86	67.55	66.18	68.22
BRRNet [28]	79.48	**81.46**	80.46	79.83
AGPNet [29]	84.72	74.86	79.48	79.28
Proposed MSTrans	84.27	80.14	**82.16**	**81.50**

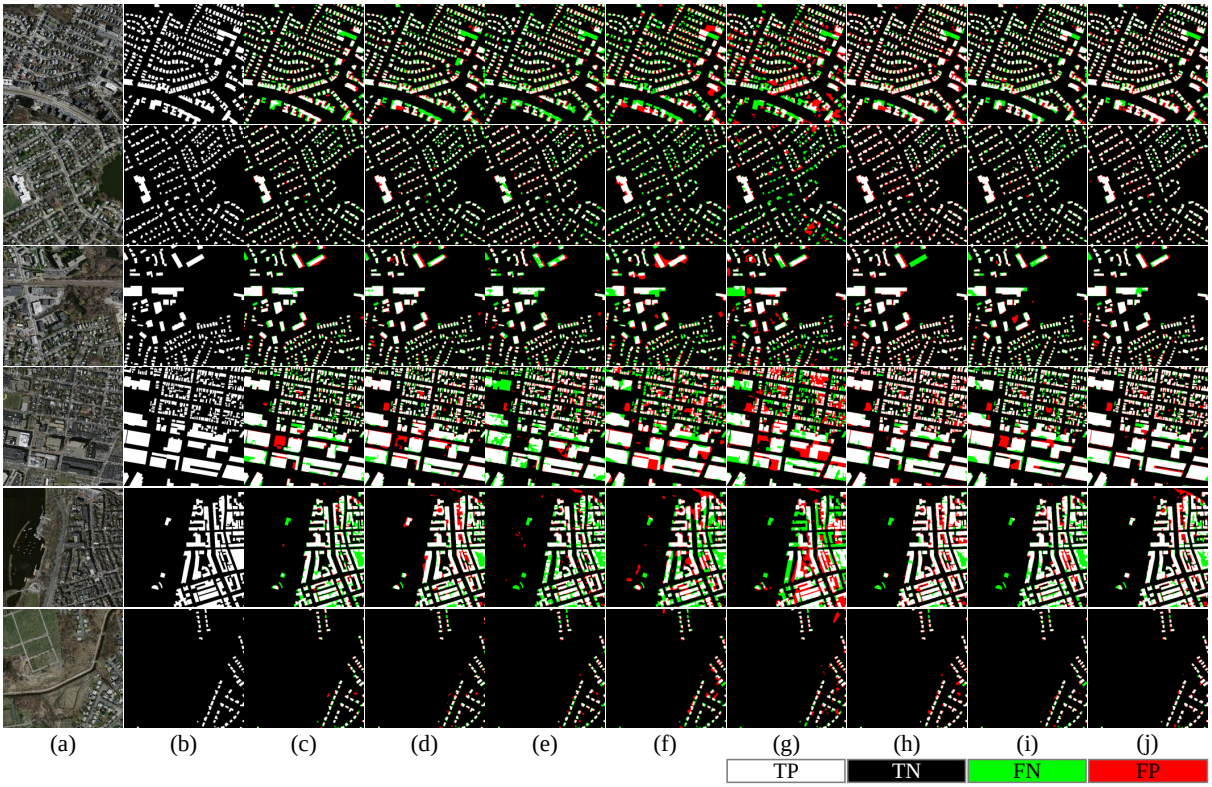

Figure 4. The BE maps of different methods on the Massachusetts dataset: (**a**) image, (**b**) label, (**c**) U-Net [20], (**d**) SiU-Net [23], (**e**) Res2Unet [24], (**f**) CFENet [31], (**g**) CBRNet [32], (**h**) BBRNet [28], (**i**) AGPNet [29], and (**j**) proposed MSTrans.

3.4.2. Results on EastAsia Dataset

In the experiments, the accuracy and visual results of different networks on the EastAsia BE dataset are provided in Table 3 and Figure 5. Compared with other methods, the proposed MSTrans achieves better accuracy in terms of the three evaluation indicators of Rec, F1, and mIoU. Specifically, the F1 and mIoU indicators of the proposed MSTrans are improved by 1.21% and 1.23%, respectively, over the second-best CBRNet [32]. In particular, compared with the second-best CBRNet [32], the Rec of the proposed model is significantly improved by 4.06%. Although U-Net acquires the best Pre, the Rec indicator accuracy is 10.91% lower than that of the proposed method. Visual comparison also supports comparison of quantitative results well. From Figure 5, we can see that the proposed method obtains a more complete building detection result, that is, very few FN and FP pixels in green and red. Moreover, the proposed method outperforms other baseline methods in BE of various scales in complex backgrounds. The effectiveness and superiority of the proposed MSTrans are verified by the quantitative accuracy and visual results on the East Asian BE dataset.

Table 3. Comparison of the quantitative accuracy (in %) of different models on the EastAsia BE dataset. **Bold** is the best accuracy.

Methods	Pre (%)	Rec (%)	F1 (%)	mIoU (%)
U-Net [20]	**88.41**	71.22	78.89	81.57
SiU-Net [23]	88.29	70.85	78.62	81.38
Res2Unet [24]	84.07	69.14	75.88	79.42
CFENet [31]	85.26	73.13	78.73	81.43
CBRNet [32]	84.84	78.07	81.32	83.32

Table 3. *Cont.*

Methods	Pre (%)	Rec (%)	F1 (%)	mIoU (%)
BRRNet [28]	84.06	78.02	80.93	83.02
AGPNet [29]	86.37	76.59	81.19	83.24
Proposed MSTrans	82.93	**82.13**	**82.53**	**84.55**

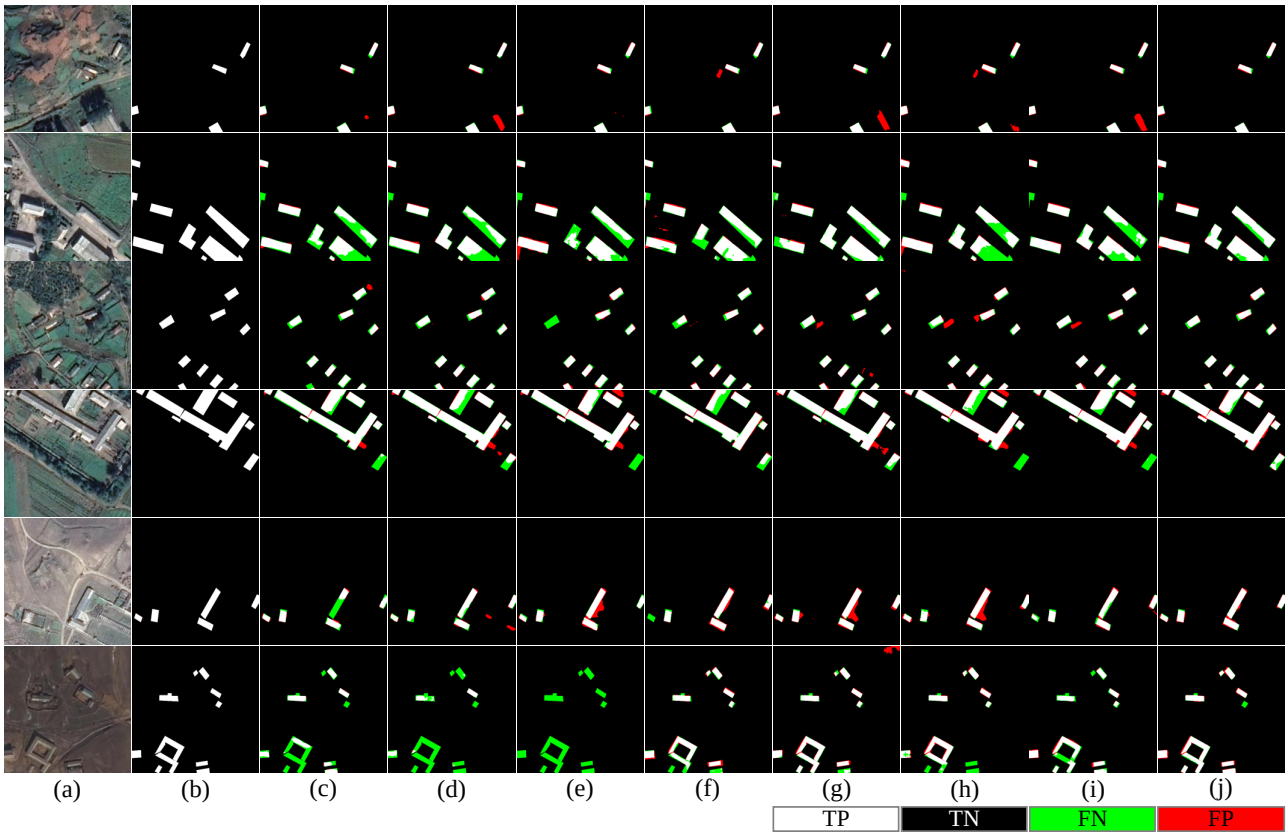

Figure 5. The BE maps of different methods on the EastAsia dataset: (**a**) image, (**b**) label, (**c**) U-Net [20], (**d**) SiU-Net [23], (**e**) Res2Unet [24], (**f**) CFENet [31], (**g**) CBRNet [32], (**h**) BBRNet [28], (**i**) AGPNet [29], and (**j**) proposed MSTrans.

3.4.3. Results on Inria Dataset

The Inria dataset consists of sub-datasets of five cities with different building styles, materials, shapes, and distribution, which can further demonstrate the robustness and superiority of the proposed MSTrans. As listed in Table 4 and Figure 6, the accuracy and visual results of different networks on the Inria BE dataset are provided. Observing the accuracy comparison results of the five sub-datasets, the proposed MSTrans achieves better accuracy performance compared with other models. For example, compared with BRRNet [28] and AGPNet [29], the proposed MSTrans reached improvements ranging from 0.08% to 3.84% in terms of Rec, F1, and mIoU on the four sub-datasets of Austin, Chicago, Kitsap, and Tyrol. On the whole, although AGPNet's Pre accuracy achieved the best performance, the comparison of the average accuracy of the five sub-datasets shows that the proposed method has obvious advantages in terms of Rec, F1, and mIoU in different scenarios. The visual comparison in Figure 6 demonstrates that the proposed MSTrans can obtain results that are closer to real buildings, as displayed in Figure 6j.

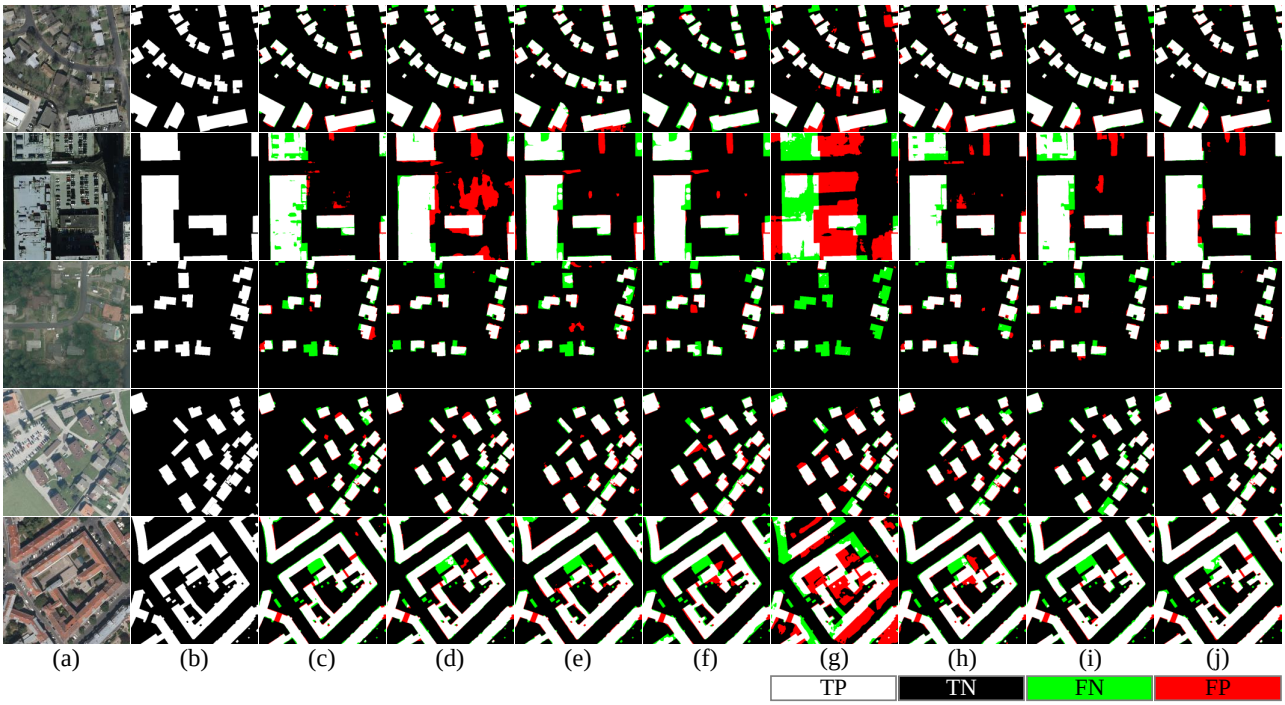

Figure 6. The BE maps of different methods on the Inria dataset: (**a**) image, (**b**) label, (**c**) U-Net [20], (**d**) SiU-Net [23], (**e**) Res2Unet [24], (**f**) CFENet [31], (**g**) CBRNet [32], (**h**) BBRNet [28], (**i**) AGPNet [29], and (**j**) proposed MSTrans.

Table 4. Comparison of the quantitative accuracy (in %) of different models on the Inria BE dataset. **Bold** is the best accuracy.

Methods		U-Net [20]	SiU-Net [23]	Res2Unet [24]	CFENet [31]	CBRNet [32]	BRRNet [28]	AGPNet [29]	Proposed MSTrans
Austin	Pre	89.92	90.48	86.86	89.20	68.14	89.14	**91.72**	90.16
	Rec	87.03	86.95	84.70	77.40	85.67	89.26	86.81	**90.02**
	F1	88.45	88.68	85.77	82.88	75.92	89.20	89.20	**90.09**
	mIoU	88.07	88.29	85.60	83.20	76.83	88.75	88.79	**89.61**
Chicago	Pre	**87.61**	81.40	79.20	81.31	77.76	87.20	86.37	86.18
	Rec	73.50	78.27	78.06	75.96	75.80	75.78	78.69	**79.03**
	F1	79.94	79.81	78.63	78.54	76.77	81.09	82.35	**82.45**
	mIoU	77.26	76.62	75.34	75.49	73.59	78.28	79.39	**79.47**
Kitsap	Pre	84.03	84.42	77.74	82.21	60.62	80.90	**85.91**	84.66
	Rec	73.16	73.55	72.40	72.27	62.72	77.57	76.24	**77.84**
	F1	78.22	78.61	74.97	76.92	61.65	79.20	80.79	**81.11**
	mIoU	80.80	81.08	78.42	79.85	69.77	81.46	82.71	**82.93**
Tyrol	Pre	87.62	88.15	85.61	86.14	84.83	85.39	**90.30**	88.14
	Rec	83.37	82.00	83.09	84.78	79.82	84.43	82.71	**86.50**
	F1	85.44	84.97	84.33	85.45	82.25	84.91	86.34	**87.31**
	mIoU	86.41	86.03	85.49	86.40	83.86	85.95	87.17	**87.96**
Vienna	Pre	89.65	89.49	86.06	88.17	58.33	88.46	**91.45**	88.16
	Rec	85.33	84.76	84.90	81.63	86.70	85.78	85.11	**88.34**
	F1	87.43	87.06	85.48	84.78	69.74	87.10	88.17	**88.25**
	mIoU	85.21	84.82	83.06	82.50	65.60	84.81	**86.04**	85.98

Table 4. *Cont.*

Methods		U-Net [20]	SiU-Net [23]	Res2Unet [24]	CFENet [31]	CBRNet [32]	BRRNet [28]	AGPNet [29]	Prposed MSTrans
Average	Pre	87.77	86.79	83.09	85.41	69.94	86.22	**89.15**	87.46
	Rec	80.48	81.11	80.63	78.41	78.14	82.57	81.91	**84.35**
	F1	83.90	83.82	81.83	81.72	73.27	84.30	85.37	**85.84**
	mIoU	83.55	83.37	81.58	81.49	73.93	83.85	84.82	**85.19**

3.4.4. Comparison and Analysis of Model Efficiency

In this subsection, we construct experiments to further compare and analyze the relationship between the efficiency and accuracy of the proposed approach. As illustrated in Table 5, several U-Net-based methods, such as SiU-Net and Res2Unet, achieve acceptable performance with relatively low graphics memory consumption. The building detection performance of U-Net even surpasses that of some state-of-the-art building extraction methods. The proposed method, which occupies a similar amount of graphics memory as most other SOTA building detection methods, achieves the best change detection performance across multiple datasets.

Table 5. Graphics memory size (M) and performance analysis (in %) of different models on three BE datasets.

Datasets	Graphics Memory (M)	Massachusetts		EastAsia		Inria	
		F1	mIoU	F1	mIoU	F1	mIoU
U-Net [20]	6656	79.58	79.52	78.89	81.57	83.90	83.55
SiU-Net [23]	6664	80.06	79.76	78.62	81.38	83.82	83.37
Res2Unet [24]	6643	72.54	73.79	75.88	79.42	81.83	81.58
CFENet [31]	7670	68.23	70.34	78.73	81.43	81.72	81.49
CBRNet [32]	7626	66.18	68.22	81.32	83.32	73.27	73.93
BBRNet [28]	7662	80.46	79.83	80.93	83.02	84.30	83.85
AGPNet [29]	7667	79.48	79.28	81.19	83.24	85.37	84.82
Proposed MSTrans	7623	82.16	81.50	82.53	84.55	85.84	85.19

4. Discussion

In this section, to further test the effectiveness of our MSTrans, we conduct ablation experiments and feature visualization analysis. The details are as follows.

4.1. Ablation Study on Massachusetts Dataset

In addition to the comparative experiments, we also constructed ablation studies on the Massachusetts BE dataset to verify the effectiveness of the proposed MST module. To achieve this, we roughly summarized the proposed MSTrans method into three submodules, including Backbone, atrous spatial pyramid pooling (ASPP) [22], and MST. Specifically, Backbone is an end-to-end segmentation network consisting of only convolutional feature extraction layers and convolutional building annotation layers. The proposed MST module has a similar structure to ASPP, and ASPP is used here for comparison. Therefore, we set the scale of ASPP to be the same as that of MST to further verify the effect of our MST module.

Based on the aforementioned settings, we obtained results for different submodule combinations on the Massachusetts dataset, as shown in Table 6 and Figure 7. Table 6 demonstrates that the performance using only the Backbone network is low. With the introduction of the Backbone network into ASPP, Rec, F1, and mIoU are improved by 1.29%, 0.43%, and 0.34%, respectively. In addition, when the proposed MST module is combined with Backbone, the performance is the best. The improvements of 0.41–2.68% can be obtained by introducing our MST module into the Backbone network in terms of the

three indicators of Rec, F1, and mIoU. Similarly, the visual results of different submodule combinations also support the conclusion of the accuracy comparison.

Table 6. Quantitative accuracy (in %) of different modules on Massachusetts dataset for ablation studies. **Bold** is the best accuracy.

Methods	Pre (%)	Rec (%)	F1 (%)	mIoU (%)
Backbone	**85.37**	77.46	81.22	80.75
w/ASPP	84.79	78.75	81.65	81.09
w/MST	84.27	**80.14**	**82.16**	**81.50**

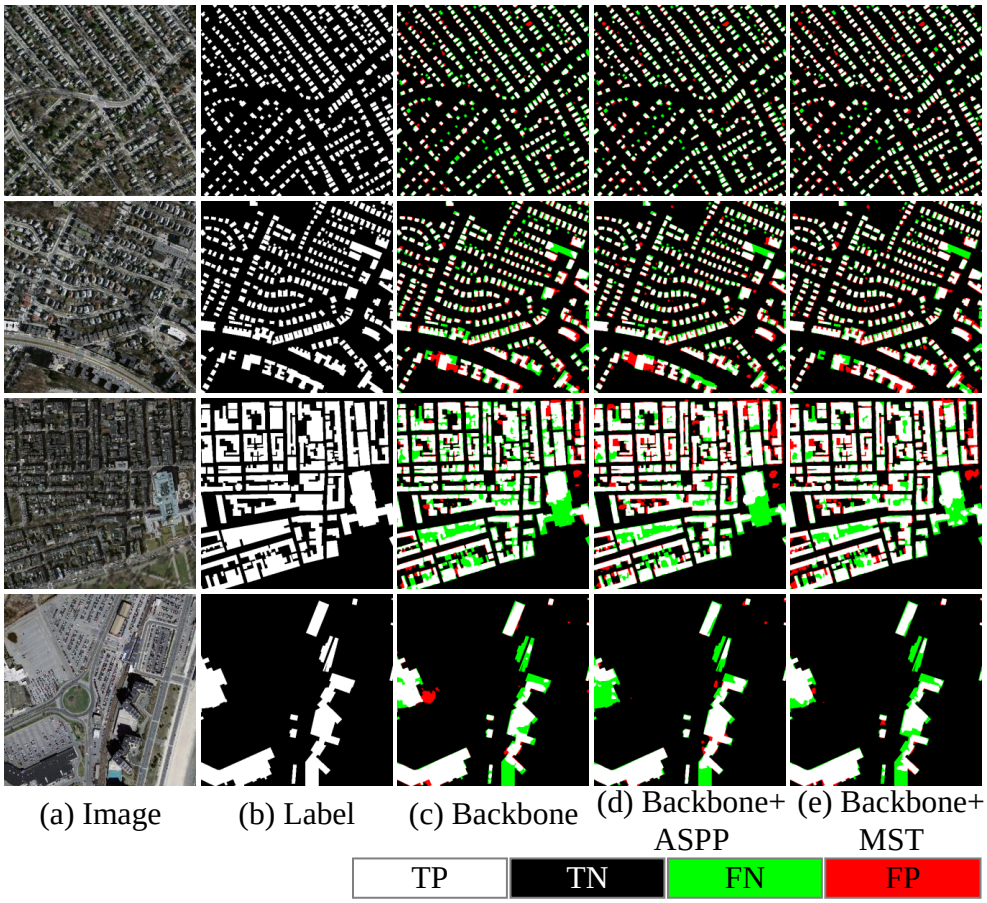

Figure 7. The BE maps of different submodule combinations on the Massachusetts dataset for ablation studies.

4.2. Feature Visualization Analysis on the EastAsia Dataset

To further analyze the impact of our MST module, we acquired the feature heatmaps of different submodule combinations on the EastAsia dataset, as shown in Figure 8. Comparing Figure 8c,d, it can be found that, after introducing ASPP, a more complete BE result can be obtained, and the false detection caused by background interference is significantly reduced. The main reason is that the introduction of ASPP can increase the multi-scale perception ability of the network through atrous convolutions of different scales, which effectively increases the ability to represent building features of various scales. When the MST module is added to the Backbone network, the feature heatmaps of the building are more accurate in BE. Unlike ASPP, the proposed MST module captures global features from different scale perspectives through the multi-scale feature tokenizer and Transformer, which helps to obtain long-range relationships between different buildings. Moreover, local features and global features are subtly aggregated in the proposed MST module

to alleviate the omission of local detail information of buildings that is easily caused by the pure Transformer. Therefore, the proposed MST model can better enhance the ability to extract building features in complex scenes and various scales, thereby improving BE accuracy. In summary, the visualization feature heatmap analysis once again verifies the effectiveness of the proposed MST module.

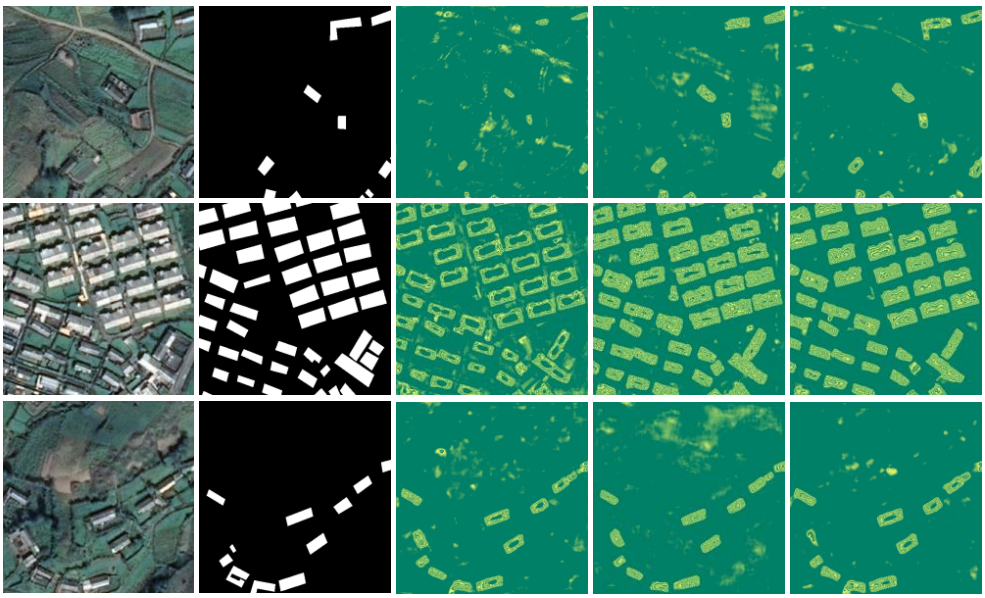

Figure 8. The visual feature heatmaps of the proposed MSTrans on different submodules.

5. Conclusions

In this paper, a novel end-to-end multi-scale Transformer (MSTrans) was proposed for BE from HR remote sensing images. In the proposed MSTrans, a plug-and-play MST module was introduced to fuse the local and global features of buildings, which can enhance multi-scale feature representation. The proposed MSTrans was applied to three real and challenging BE datasets. Experimental results demonstrate the effectiveness of the proposed method, and comparison with the other seven methods also shows the superiority of the proposed method. In addition, ablation studies and feature visualization analysis also verify the effectiveness of the proposed MST module and MSTrans. However, there are two aspects that need to be further studied in future works:

(1) The proposed method still relies on a large number of labeled samples for training to achieve high-performance BE. Future work will focus on reducing the proposed network's dependence on labeled samples while maintaining the network's performance as much as possible.

(2) The effectiveness of the proposed method is only validated under homologous data, and the effectiveness of heterogeneous data still needs further testing. In particular, due to the differences in sensors and imaging principles, heterogeneous remote sensing data can supplement more information and improve BE performance by combining building information from different data. However, how to extract and fuse heterogeneous data to improve the robustness and generalization of BE remains a challenge. At the same time, the BE detection capability of the proposed method in cross-domain scenarios will be studied in the future, which is the focus of promoting the practicality of the algorithm.

Author Contributions: Conceptualization, F.Y. and F.J.; methodology, F.Y., F.J. and J.L.; validation, F.Y., J.L. and L.L.; investigation, J.L. and L.L.; writing—original draft preparation, F.Y. and F.J.; writing—review and editing, F.Y., F.J., J.L. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the National Natural Science Foundation of China (Grant No. 72161035 and No. 72461033), Natural Science and Technology Project Plan in Yulin of China (Grant No. CXY-2020-015), Scientific Research Program Funded by Yulin National High Tech Industrial Development Zone (Program No. CXY-2021-25), the Natural Science Basic Research Program of Shaanxi (Grant No. 2024JC-YBQN-0633), and the Fundamental Research Funds for the Central Universities, China (Grant No. XJSJ24016).

Data Availability Statement: Data are contained within the article.

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

References

1. Weidner, U.; Förstner, W. Towards automatic building extraction from high-resolution digital elevation models. *ISPRS J. Photogramm. Remote Sens.* **1995**, *50*, 38–49. [CrossRef]
2. Luo, L.; Li, P.; Yan, X. Deep learning-based building extraction from remote sensing images: A comprehensive review. *Energies* **2021**, *14*, 7982. [CrossRef]
3. Liu, T.; Gong, M.; Lu, D.; Zhang, Q.; Zheng, H.; Jiang, F.; Zhang, M. Building change detection for VHR remote sensing images via local–global pyramid network and cross-task transfer learning strategy. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 4704817. [CrossRef]
4. Yuan, J. Learning building extraction in aerial scenes with convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 2793–2798. [CrossRef]
5. Yang, H.L.; Yuan, J.; Lunga, D.; Laverdiere, M.; Rose, A.; Bhaduri, B. Building extraction at scale using convolutional neural network: Mapping of the united states. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 2600–2614. [CrossRef]
6. Wang, W.; Shi, Y.; Zhang, J.; Hu, L.; Li, S.; He, D.; Liu, F. Traditional village building extraction based on improved Mask R-CNN: a case study of Beijing, China. *Remote Sens.* **2023**, *15*, 2616. [CrossRef]
7. Jin, X.; Davis, C.H. Automated building extraction from high-resolution satellite imagery in urban areas using structural, contextual, and spectral information. *EURASIP J. Adv. Signal Process.* **2005**, *2005*, 745309. [CrossRef]
8. Fan, Z.; Wang, S.; Pu, X.; Wei, H.; Liu, Y.; Sui, X.; Chen, Q. Fusion-Former: Fusion Features across Transformer and Convolution for Building Change Detection. *Electronics* **2023**, *12*, 4823. [CrossRef]
9. Baltsavias, E.; Mason, S.; Stallmann, D. Use of DTMs/DSMs and orthoimages to support building extraction. In *Automatic Extraction of Man-Made Objects from Aerial and Space Images*; Springer: Berlin/Heidelberg, Germany, 1995; pp. 199–210.
10. Vu, T.T.; Yamazaki, F.; Matsuoka, M. Multi-scale solution for building extraction from LiDAR and image data. *Int. J. Appl. Earth Obs. Geoinf.* **2009**, *11*, 281–289. [CrossRef]
11. Ortner, M.; Descombes, X.; Zerubia, J. Building extraction from digital elevation models. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, 6–10 April 2003; IEEE: New York, NY, USA, 2003; Volume 3, p. III–337.
12. Tournaire, O.; Brédif, M.; Boldo, D.; Durupt, M. An efficient stochastic approach for building footprint extraction from digital elevation models. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 317–327. [CrossRef]
13. Lee, D.S.; Shan, J.; Bethel, J.S. Class-guided building extraction from Ikonos imagery. *Photogramm. Eng. Remote Sens.* **2003**, *69*, 143–150. [CrossRef]
14. Zhang, L.; Huang, X.; Huang, B.; Li, P. A pixel shape index coupled with spectral information for classification of high spatial resolution remotely sensed imagery. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 2950–2961. [CrossRef]
15. Bi, Q.; Qin, K.; Zhang, H.; Zhang, Y.; Li, Z.; Xu, K. A multi-scale filtering building index for building extraction in very high-resolution satellite imagery. *Remote Sens.* **2019**, *11*, 482. [CrossRef]
16. Huang, X.; Zhang, L. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2011**, *5*, 161–172. [CrossRef]
17. Liasis, G.; Stavrou, S. Building extraction in satellite images using active contours and colour features. *Int. J. Remote Sens.* **2016**, *37*, 1127–1153. [CrossRef]
18. Li, Q.; Mou, L.; Sun, Y.; Hua, Y.; Shi, Y.; Zhu, X.X. A Review of Building Extraction from Remote Sensing Imagery: Geometrical Structures and Semantic Attributes. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 4702315. [CrossRef]
19. Dong, X.; Cao, J.; Zhao, W. A review of research on remote sensing images shadow detection and application to building extraction. *Eur. J. Remote Sens.* **2024**, *57*, 2293163. [CrossRef]
20. 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 2015: 18th International Conference, Munich, Germany, 5–9 October 2015 ; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.

21. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
22. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
23. 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.* **2018**, *57*, 574–586. [CrossRef]
24. Chen, F.; Wang, N.; Yu, B.; Wang, L. Res2-Unet, a new deep architecture for building detection from high spatial resolution images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1494–1501. [CrossRef]
25. Ji, S.; Wei, S.; Lu, M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *Int. J. Remote Sens.* **2019**, *40*, 3308–3322. [CrossRef]
26. Ma, J.; Wu, L.; Tang, X.; Liu, F.; Zhang, X.; Jiao, L. Building extraction of aerial images by a global and multi-scale encoder-decoder network. *Remote Sens.* **2020**, *12*, 2350. [CrossRef]
27. Guo, H.; Shi, Q.; Du, B.; Zhang, L.; Wang, D.; Ding, H. Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 4287–4306. [CrossRef]
28. 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]
29. Deng, W.; Shi, Q.; Li, J. Attention-gate-based encoder–decoder network for automatical building extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2611–2620. [CrossRef]
30. Tian, Q.; Zhao, Y.; Li, Y.; Chen, J.; Chen, X.; Qin, K. Multiscale building extraction with refined attention pyramid networks. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8011305. [CrossRef]
31. Chen, J.; Zhang, D.; Wu, Y.; Chen, Y.; Yan, X. A context feature enhancement network for building extraction from high-resolution remote sensing imagery. *Remote Sens.* **2022**, *14*, 2276. [CrossRef]
32. 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]
33. Bischke, B.; Helber, P.; Folz, J.; Borth, D.; Dengel, A. Multi-task learning for segmentation of building footprints with deep neural networks. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE: New York, NY, USA, 2019; pp. 1480–1484.
34. 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.* **2021**, *15*, 1526–1542. [CrossRef]
35. Zhu, Y.; Liang, Z.; Yan, J.; Chen, G.; Wang, X. ED-Net: Automatic building extraction from high-resolution aerial images with boundary information. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4595–4606. [CrossRef]
36. Jung, H.; Choi, H.S.; Kang, M. Boundary enhancement semantic segmentation for building extraction from remote sensed image. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5215512. [CrossRef]
37. Zhou, Y.; Chen, Z.; Wang, B.; Li, S.; Liu, H.; Xu, D.; Ma, C. BOMSC-Net: Boundary optimization and multi-scale context awareness based building extraction from high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5618617. [CrossRef]
38. Yu, B.; Chen, F.; Wang, N.; Yang, L.; Yang, H.; Wang, L. MSFTrans: A multi-task frequency-spatial learning transformer for building extraction from high spatial resolution remote sensing images. *GIScience Remote Sens.* **2022**, *59*, 1978–1996. [CrossRef]
39. Chen, X.; Xiao, P.; Zhang, X.; Muhtar, D.; Wang, L. A Cascaded Network with Coupled High-Low Frequency Features for Building Extraction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2024**, *17*, 10390–10406. [CrossRef]
40. Lu, L.; Liu, T.; Jiang, F.; Han, B.; Zhao, P.; Wang, G. DFANet: Denoising Frequency Attention Network for Building Footprint Extraction in Very-High-Resolution Remote Sensing Images. *Electronics* **2023**, *12*, 4592. [CrossRef]
41. Wagner, F.H.; Dalagnol, R.; Tarabalka, Y.; Segantine, T.Y.; Thomé, R.; Hirye, M.C. U-net-id, an instance segmentation model for building extraction from satellite images—Case study in the joanópolis city, brazil. *Remote Sens.* **2020**, *12*, 1544. [CrossRef]
42. Chen, S.; Ogawa, Y.; Zhao, C.; Sekimoto, Y. Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 129–152. [CrossRef]
43. Wen, Q.; Jiang, K.; Wang, W.; Liu, Q.; Guo, Q.; Li, L.; Wang, P. Automatic building extraction from Google Earth images under complex backgrounds based on deep instance segmentation network. *Sensors* **2019**, *19*, 333. [CrossRef]
44. Wu, T.; Hu, Y.; Peng, L.; Chen, R. Improved anchor-free instance segmentation for building extraction from high-resolution remote sensing images. *Remote Sens.* **2020**, *12*, 2910. [CrossRef]
45. Wang, L.; Fang, S.; Meng, X.; Li, R. Building extraction with vision transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5625711. [CrossRef]
46. Zhang, R.; Wan, Z.; Zhang, Q.; Zhang, G. DSAT-net: Dual spatial attention transformer for building extraction from aerial images. *IEEE Geosci. Remote. Sens. Lett.* **2023**, *20*, 6008405. [CrossRef]
47. Yiming, T.; Tang, X.; Shang, H. A shape-aware enhancement Vision Transformer for building extraction from remote sensing imagery. *Int. J. Remote Sens.* **2024**, *45*, 1250–1276. [CrossRef]

48. Gibril, M.B.A.; Al-Ruzouq, R.; Shanableh, A.; Jena, R.; Bolcek, J.; Shafri, H.Z.M.; Ghorbanzadeh, O. Transformer-based semantic segmentation for large-scale building footprint extraction from very-high resolution satellite images. *Adv. Space Res.* **2024**, *73*, 4937–4954. [CrossRef]
49. Ding, L.; Tang, H.; Liu, Y.; Shi, Y.; Zhu, X.X.; Bruzzone, L. Adversarial shape learning for building extraction in VHR remote sensing images. *IEEE Trans. Image Process.* **2021**, *31*, 678–690. [CrossRef]
50. Hu, A.; Wu, L.; Xu, Y.; Xie, Z. SANET: A Shape-aware Building Footprints Extraction Method in Remote Sensing Images by Integrating Fourier Shape Descriptors. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5632215. [CrossRef]
51. Li, X.; Zhang, G.; Cui, H.; Hou, S.; Chen, Y.; Li, Z.; Li, H.; Wang, H. Progressive fusion learning: A multimodal joint segmentation framework for building extraction from optical and SAR images. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 178–191. [CrossRef]
52. Shi, X.; Gao, J.; Yuan, Y. Enhancing Uni-Modal Features Matters: A Multi-Modal Framework for Building Extraction. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5622013.
53. Chen, X.; Qiu, C.; Guo, W.; Yu, A.; Tong, X.; Schmitt, M. Multiscale feature learning by transformer for building extraction from satellite images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 2503605. [CrossRef]
54. Chang, J.; He, X.; Li, P.; Tian, T.; Cheng, X.; Qiao, M.; Zhou, T.; Zhang, B.; Chang, Z.; Fan, T. Multi-Scale Attention Network for Building Extraction from High-Resolution Remote Sensing Images. *Sensors* **2024**, *24*, 1010. [CrossRef]
55. Xia, L.; Mi, S.; Zhang, J.; Luo, J.; Shen, Z.; Cheng, Y. Dual-stream feature extraction network based on CNN and transformer for building extraction. *Remote Sens.* **2023**, *15*, 2689. [CrossRef]
56. Chang, J.; Cen, Y.; Cen, G. Asymmetric Network Combining CNN and Transformer for Building Extraction from Remote Sensing Images. *Sensors* **2024**, *24*, 6198. [CrossRef] [PubMed]
57. Yuan, Q.; Xia, B. Cross-level and multiscale CNN-Transformer network for automatic building extraction from remote sensing imagery. *Int. J. Remote Sens.* **2024**, *45*, 2893–2914. [CrossRef]
58. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607514. [CrossRef]
59. Gong, M.; Liu, T.; Zhang, M.; Zhang, Q.; Lu, D.; Zheng, H.; Jiang, F. Context–content collaborative network for building extraction from high-resolution imagery. *Knowl.-Based Syst.* **2023**, *263*, 110283. [CrossRef]
60. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto (Canada): Toronto, ON, Canada, 2013.
61. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; IEEE: New Tork, NY, USA, 2017; pp. 3226–3229.