electronics

MDPI

Article

Multi-Scale Feature Fusion Attention Network for Building Extraction in Remote Sensing Images

Jia Liu, Hang Gu *, Zuhe Li, Hongyang Chen and Hao Chen

School of Computer Science and Technology, Zhengzhou University of Light Industry, Zhengzhou 450002, China; ieliujia@zzuli.edu.cn (J.L.); zuheli@zzuli.edu.cn (Z.L.); 332207050701@email.zzuli.edu.cn (H.C.); 332305040480@email.zzuli.edu.cn (H.C.)

* Correspondence: 332207050689@email.zzuli.edu.cn

Abstract: The efficient semantic segmentation of buildings in high spatial resolution remote sensing images is a technical prerequisite for land resource management, high-precision mapping, construction planning and other applications. Current building extraction methods based on deep learning can obtain high-level abstract features of images. However, the extraction of some occluded buildings is inaccurate, and as the network deepens, small-volume buildings are lost and edges are blurred. Therefore, we introduce a multi-resolution attention combination network, which employs a multiscale channel and spatial attention module (MCAM) to adaptively capture key features and eliminate irrelevant information, which improves the accuracy of building extraction. In addition, we present a layered residual connectivity module (LRCM) to enhance the expression of information at different scales through multi-level feature fusion, significantly improving the understanding of context and the capturing of fine edge details. Extensive experiments were conducted on the WHU aerial image dataset and the Massachusetts building dataset. Compared with state-of-the-art semantic segmentation methods, this network achieves better building extraction results in remote sensing images, proving the effectiveness of the method.

Keywords: remote sensing; building extraction; multi-scale fusion; attention mechanism

check for updates

Citation: Liu, J.; Gu, H.; Li, Z.; Chen, H.; Chen, H. Multi-Scale Feature Fusion Attention Network for Building Extraction in Remote Sensing Images. *Electronics* **2024**, *13*, 923. https://doi.org/10.3390/electronics13050923

Academic Editor: Silvia Liberata Ullo

Received: 30 January 2024
Revised: 21 February 2024
Accepted: 26 February 2024
Published: 28 February 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

With the continuous advancement of aerospace remote sensing technology, we can easily obtain large-scale high spatial resolution (HSR) image data containing rich architectural details. The semantic segmentation of high-resolution remote sensing buildings is a major component of our remote sensing Earth observation technology. By analyzing the acquired remote sensing images, we extract relevant feature information on buildings and classify the various elemental targets in the images [1]. These endeavors provide foundational support for numerous tasks, including land resource management, detailed cartography, monitoring land use changes, and assessing suitability for human habitation and the environment. Although remote sensing images offer a wide range of potential applications, characteristics such as large computational demands, complex processes, and redundant information present certain challenges. The structure of buildings is complex, with subtle distribution differences, and the surrounding environment is dynamic and variable. These factors severely interfere with building segmentation, affecting aspects like illumination, seasonal changes, shooting angles, unclear boundaries, and complex background information [2]. These interference factors can result in issues like blurred building boundaries and the loss of small targets.

With the continuous development of computer vision technology, extensive research on the semantic segmentation of high-resolution remote sensing buildings has been increasingly undertaken, leading to the emergence of numerous semantic segmentation methods for remote sensing images [3,4]. Building segmentation methods can generally

be categorized into two major types: traditional feature-based methods and deep learning feature-based methods. In traditional feature-based methods, various algorithms for building extraction have been proposed by researchers in this field. Although some success has been achieved, significant shortcomings persist, such as the limited generalization capability of manually designed features leading to insufficient segmentation accuracy. Deep learning convolutional neural networks (CNNs) have demonstrated remarkable performance in tasks like car navigation and medical image segmentation, gradually becoming the dominant methodology in the field of remote sensing segmentation [5,6]. The fully convolutional network (FCN) is a commonly used model in deep learning for semantic segmentation and is capable of achieving end-to-end target extraction [7]. However, the effective utilization of features of different scales and the integration of high-level semantic information with low-level location spatial information remain pressing research challenges. Researchers have proposed methods of feature fusion for image segmentation [8,9]. The attention mechanism, which enhances strong correlation features and suppresses weak correlation features, is introduced to address differences in various areas of the feature map [10,11]. However, accurately capturing the regional properties of buildings on a large scale with a pixel-level attention mechanism remains a challenging task.

Despite advancements and satisfactory achievements in semantic segmentation tasks for buildings, significant challenges still need to be addressed. The downsampling process reduces the resolution of the feature map, leading to the loss of much small detailed feature information. For small buildings and those with complex edge features, this reduction in resolution may lead to their being overlooked during the extraction process, resulting in issues like the loss of small building predictions and blurring of building edges [12].

To address the aforementioned challenges, we introduce a new deep learning network specifically designed to enhance the building extraction task. The key contributions of our research are summarized as follows:

1. Multi-resolution attention network: We introduce a network structure named the multi-resolution attention network. This network fully utilizes the attention mechanism, concentrating on the most critical features and effectively addressing the issue of blurred boundaries in the segmentation process through the integration of multi-scale information.
2. Multiscale channel and spatial attention module: In response to the unique characteristics of remote sensing building images, we have designed the MCAM. This module operates adaptively, more effectively capturing crucial features and filtering out irrelevant information. This capability enables the model to selectively concentrate on the critical parts of the image, thus enhancing the accuracy of building extraction.
3. Layered residual connectivity module: We designed the LRCM. This module aims to augment the expression of information across various scales by merging features from multiple levels. This enhancement not only improves context understanding but also yields significant results in capturing fine edge details and fusing high-level abstract features. This approach offers an effective means of enhancing the performance of building extraction models.

The remainder of our paper is structured as follows. Section 2 provides an overview of current building segmentation methods, discussing the existing research background and approaches. Section 3 offers a detailed introduction to the architecture of the building extraction network, including the MCAM and the crucial LRCM component. Section 4 reports and analyzes experimental results and relevant demonstrations, evaluating the performance and effectiveness of the proposed network. Finally, Section 5 concludes the paper.

2. Related Works

In this section, our study delves into related research in the field of building segmentation in remote sensing images, covering both traditional feature-based and deep feature-based building segmentation methods.

2.1. Traditional Feature-Based Segmentation Methods

A number of researchers have proposed various algorithms for building extraction; however, a significant portion of these algorithms predominantly rely on manually designed features. These features encompass geometric, textural, shadow, and edge characteristics [13–16]. Furthermore, techniques such as support vector machines [13], AdaBoost [14], conditional random fields (CRF) [17], and random forests [18] are frequently used for pixel labeling. Building extraction is achieved by leveraging the mathematical geometric relationships among building line features. Nitin L. proposed a morphology-based method for automatic building extraction from high resolution satellite (HRS) images [19]. This method incorporates morphological top-hat filtering, facilitating the detection of buildings of varying sizes and shapes. Xu et al. successfully extracted buildings by combining their symmetry and color features [20]. They utilized the K-means clustering method for region segmentation to accomplish building extraction. However, it is important to note that this method falls short in effectively segmenting building edges. Cheng et al. [21] introduced an innovative active contour algorithm that utilizes the image's HSV representation, effectively extracting building outlines by leveraging the distinct color characteristics of buildings to minimize the impact of vegetation and shadows. Dai et al. employed a multi-stage processing approach, utilizing various techniques including the support vector machine (SVM) and W-k-means clustering algorithm [22]. This approach is designed to standardize the process from vegetation point extraction to building ground contouring, leading to a more comprehensive building extraction.

Traditional methods exhibit certain limitations in building segmentation, particularly in handling complex buildings and diverse background scenarios [23]. These methods often necessitate additional manual intervention and post-processing steps, and are typically suited only for certain types of buildings or specific remote sensing images. The generalization capabilities of these methods are limited, making them challenging to adapt to different scenarios.

2.2. Segmentation Based on Deep Learning Methods

Following the advent of convolutional neural networks, target extraction algorithms have progressively shifted from manually designed feature-based methods to deep feature-based techniques. With the advancement of CNNs, FCNs have for the first time enabled end-to-end and pixel-level image segmentation. In remote sensing building extraction, numerous methods based on deep CNNs have been developed. The fundamental concept involves labeling each pixel of the image as either a building or non-building. Most building segmentation methods typically employ FCNs or their variants as the foundational structure, enhancing the model's multi-scale learning capabilities. High-level features provide rich global semantic information, whereas low-level features offer detailed local insights [24]. A lack of global semantic information can result in omissions in large-scale building predictions, and a paucity of local details might cause small buildings to be overlooked. Semantic segmentation algorithms using the U-net [25] encoder-decoder structure, which capitalizes on residual and dense connectivity, have yielded significant advances in high-resolution remote sensing building segmentation [26,27]. The U-net structure encodes by downsampling the image and employs compressed feature representation for decoder upsampling, aiming to restore image resolution and facilitate predictions. In the upsampling process, the deep feature map is combined with the shallow feature map of identical resolution, reducing the loss of positional information incurred during encoding. However, the low resolution of this method presents challenges in precisely locating target boundaries. Conversely, shallow feature maps, despite their higher resolution, can result in prediction errors owing to a lack of sufficient semantic information.

In convolutional neural network-based methods, downsampling operations often result in the loss of local details in the input image. For example, DeepLab utilizes dilated convolutions to mitigate this issue [28]. Guo et al. developed a module that incrementally refines building predictions from coarse to fine, leveraging the strengths of both high-level

and low-level features [29]. Fang et al. utilized the Res2Net concept to create Res2-UNet [30]. They integrated modified residual modules in the downsampling phase to address the confusion of boundary pixels with background objects and to prevent the overlooking of small-sized buildings. Recently, researchers have introduced attention mechanisms into diverse fields including machine translation, pose estimation, and image processing. Attention mechanisms are capable of capturing contextual information in spatial or channel dimensions. Zhang et al. employed the Swin Transformer for feature extraction, aiming to enhance modeling of distant spatial dependencies [31]. The atrous spatial pyramid pooling block with depthwise separable convolution (SASPP) is utilized to acquire multi-scale context, combined with channel attention blocks and auxiliary boundary detection to constrain edge segmentation. Zhou et al. developed the AGs to link low-level features with high-level semantic information [32]. Liu et al. utilized the convolutional block attention module (CBAM) and deformable convolutional network (DCN) to improve feature extraction accuracy and generalization capacity [33]. Wang et al. incorporated residual learning and spatial attention units into the encoder [34]. Ku, Yang, and Zhang proposed a novel convolutional neural network (CNN) architecture aimed at improving the accuracy and efficiency of semantic segmentation [35]. Zhang et al. developed a novel dual skip connection structure in the model [36].

Owing to the fixed size of convolutional kernels and insufficient contextual information in images, current methods require further enhancement in representing object features. To address this limitation, multi-scale information generated at various stages of the network can be utilized and integrated for computation. These operations aid in reducing potential information loss during the downsampling and upsampling processes. In building segmentation tasks, this approach enables a finer delineation of edges, thus preventing the loss of smaller, less prominent buildings.

Traditional methods, with their reliance on manually designed features and multi-stage processing, exhibit limited generalization capabilities in complex scenes. Conversely, deep learning-based methods, utilizing fully convolutional networks, achieve end-to-end and pixel-level image segmentation. Recent advances in building segmentation have significantly improved accuracy and generalization capabilities through the introduction of techniques like attention mechanisms, residual learning, and multi-scale information fusion. However, when dealing with high-resolution buildings, issues such as blurred boundaries and inaccurate localization remain pressing challenges to be addressed.

3. Method

This section primarily focuses on the multi-resolution attention fusion network presented in our paper. It includes an overview of the network's architecture and its two key modules: MCAM and LRCM.

3.1. Multi-Resolution Attention Network Framework

In contemporary research, the majority of segmentation models utilize an encoder-decoder structure. Nonetheless, maintaining higher resolution in deeper-level features during an encoder's downsampling process is challenging. Repeated convolution operations can lead to the loss of boundary details and unclear information.

We employed the MCAM and the LRCM modules for feature fusion. This fusion strategy incrementally enhances features layer by layer and leverages information from various levels to improve model performance. The LRCM design facilitates more efficient gradient propagation within the network, effectively mitigating the common issue of gradient vanishing in deep networks. Such optimization of information flow ensures the preservation of key features, enabling the development of deeper and more complex model structures without compromising performance. Simultaneously, LRCM's ability to integrate features across different scales effectively addresses the challenge of merging high-level abstract information with low-level detail in images. This dual emphasis on optimizing information flow and enhancing multi-scale feature fusion plays a crucial role.

To address this issue, we propose a multi-resolution attention fusion network. In Algorithm 1 below, we introduce the network algorithm structure proposed in our study. This table encapsulates the model's entire processing flow from the input image to the output segmentation map, including the initialization of the VGG16 backbone, feature extraction, enhancement through attention mechanisms, feature fusion, and upsampling steps. We adopt VGG16 as the backbone network due to its deep feature extraction capabilities and excellent generalizability, which is widely applied in various visual tasks, providing a solid feature foundation for subsequent model learning and recognition tasks. The network employs an attention mechanism to adaptively enhance the importance of features at each position. Furthermore, the network fuses the features of different resolutions by utilizing feature maps of varying sizes created during the downsampling process. This network consists of four key components:

1. Multiple pooling layers in the multi-layer feature extraction section capture features at various scales, ensuring continuous feature extraction for subsequent modules.
2. The adaptive feature enhancement attention module implements an attention mechanism on the input of the feature extraction section, achieving adaptive feature enhancement by assigning different weights to different features.
3. The task of the feature aggregation fusion module is to organically fuse high-resolution features with low-resolution ones, thereby enhancing detail extraction.
4. In the progressive upsampling section, during the decoding process, further fusion of the already integrated features with bottom-layer features is performed to ultimately obtain the prediction results.

Algorithm 1 Enhanced Segmentation Model with VGG Backbone and Attention Mechanisms

Input: Input image
Output: Segmentation map
1: Initialize a VGG16 backbone with optional pretrained weights.
2: Use the first four layers of VGG16 to extract initial feature maps from the input image.
3: Employ subsequent layers of VGG16 to extract deeper, more complex features.
4: Apply MRAM modules on the extracted features for channel and spatial attention enhancement.
5: Utilize LRCM modules to integrate and fuse features from various hierarchical levels.
6: Perform progressive upsampling and feature concatenation:
    a. For each upsampling stage, concatenate the upsampled feature with the corresponding LRCM-enhanced feature.
    b. Apply convolution operations followed by ReLU activations.
7: Generate the final segmentation map using a convolutional layer.
8: Output the segmentation map.

As shown in Figure 1, the multi-resolution attention fusion network initially conducts multi-level feature extraction on the input remote sensing building images. Each hierarchical structure includes the multi-resolution attention fusion structure (MCAM + LRCM). The multi-resolution attention fusion structure, while maintaining the number and size of each channel unchanged, can fuse multi-level attention features to enhance contextual feature information. The channels in the network are divided into four groups, each passing through a top-down multi-resolution module. Each module receives attention features from the subsequent layer, with reception quantities of 4, 3, 2, and 1, respectively. This design achieves the fusion of multi-resolution features. Subsequently, the decoder is responsible for receiving bottom-layer features and the features from the multi-resolution attention fusion structure, performing progressive upsampling operations, and ultimately restoring the image to its original size. This process aims to effectively integrate features from the bottom layer and multi-resolution attention to enhance the representation and segmentation performance of building images.

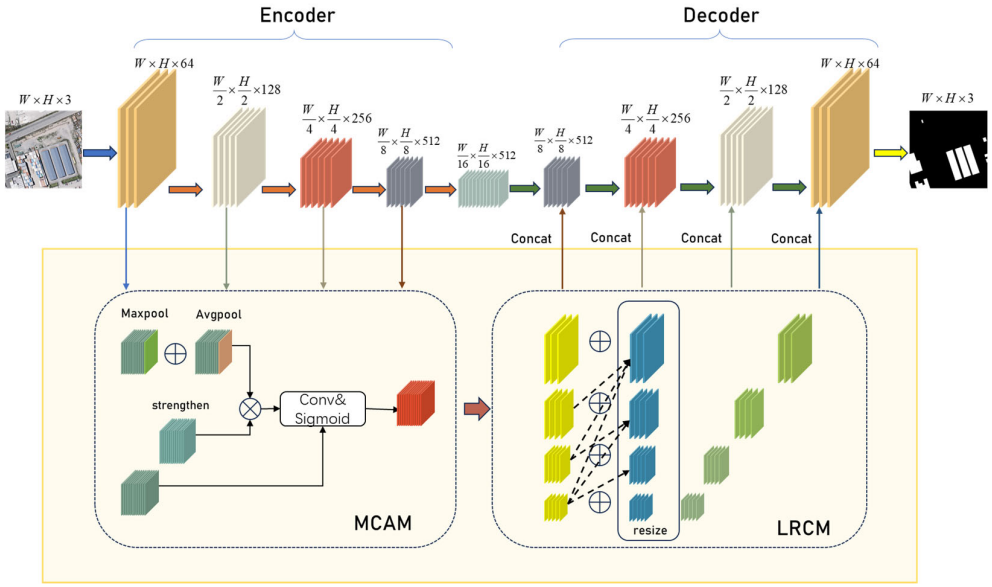


Figure 1. Schematic diagram of the network framework. The encoder performs downsampling to extract features, which are then enhanced by the MCAM and LRCM modules before being transmitted to the decoder for progressive upsampling to restore the original size.

3.2. Multiscale Channel and Spatial Attention Module

High-resolution remote sensing images typically have a large resolution, relatively small targets for segmentation, and a significant proportion of irrelevant background, often leading to the neglect of small buildings by extraction networks. To address this issue, introducing attention mechanisms can effectively utilize global statistical information, enhancing the salient features of segmentation targets in remote sensing images. Simultaneously, without reducing the spatial resolution of the image, attention mechanisms help suppress non-essential features such as background. However, conventional attention mechanisms usually focus on specific dimensions, such as spatial or channel, and primarily emphasize local features, potentially overlooking global contextual information.

To tackle the mentioned challenges, our paper introduces the multiscale channel and spatial attention module (MCAM) and integrates it into the skip-connection section. In Algorithm 2 presented in the text, the table delineates the entire process from the input feature maps to the output refined feature maps, encompassing the application of both channel and spatial attention modules. This includes performing average and max pooling on the input feature maps to capture global information, processing the pooled features through a multi-layer perceptron (MLP) to generate the channel attention map, and applying the spatial attention module to further refine the spatial distribution of the feature maps.

The design of MCAM aims to comprehensively capture and leverage global contextual information by focusing on features from multiple channels, enhancing the perceptual capability for various types of buildings in remote sensing images. This innovative mechanism helps maintain high resolution while more effectively highlighting targets like buildings, thereby enhancing segmentation performance. By integrating spatial and channel attention, it adjusts weights for each pixel in the image at the pixel level, capturing subtle local patterns. Additionally, it simultaneously considers relationships between different channels, ensuring a balance with global contextual information. Figure 2 illustrates the algorithmic process and structural schematic of the multiscale channel and spatial attention module.

Algorithm 2 Multiscale Channel and Spatial Attention Module

Input: Feature maps $X \in \mathbb{R}^{C \times H \times W}$
Output: The refined feature maps Y
1: for each feature map in the batch do
2: Channel Attention Module:
3: Perform average pooling and max pooling along spatial dimensions (H and W).
4: Apply shared Multi-Layer Perceptron (MLP) on both pooled features.
5: Sum the outputs of the MLP.
6: Apply sigmoid function to obtain channel attention map.
7: Spatial Attention Module:
8: Perform average pooling and max pooling along the channel axis.
9: Concatenate the pooled features along the channel axis.
10: Convolve concatenated features and apply sigmoid function to obtain spatial attention map.
11: Multiply the channel-refined feature maps with the spatial attention map.
12: end for
13: Output the refined feature maps Y after sequentially applying channel and spatial attention mechanisms.

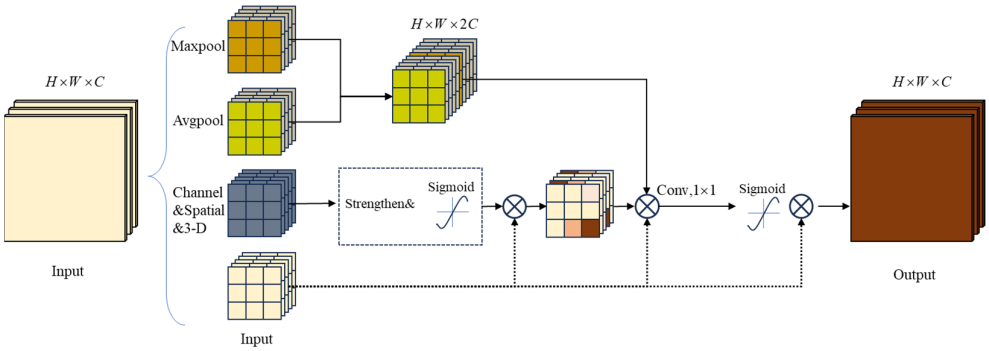


Figure 2. Multiscale channel and spatial attention module.

The multiscale channel and spatial attention module (MCAM) integrates channel attention, similarity activation, and spatial attention. Channel attention obtains channel statistics through global average pooling and global max pooling. Channel correlations are introduced through a fully connected layer and Sigmoid activation. For the input features, $F \in R^{C*H*W}$ (where C is the number of channels, and H and W are the height and width of the input feature map, respectively). Channel, similarity, and spatial dimension statistical information are extracted through global average pooling (1) and global max pooling (2) operations:

$$\text{mean} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{c,i,j} \tag{1}$$

$$\underline{Max}out = \max_{i,j}(F_{c,i,j}) \tag{2}$$

Through the channel attention mechanism, dimensionality reduction is achieved using convolution operations and introducing the Sigmoid activation function to obtain the weights of channel attention. Subsequently, the similarity activation mechanism (4) adjusts features by calculating the mean and variance and introducing the similarity activation formula:

$$\text{var} = \frac{1}{H \times W - 1} \sum_{i=1}^{H} \sum_{j=1}^{W} (F_{c,i,j} - \text{mean})^2 \tag{3}$$

$$\text{Similarity_activation} = \text{Sigmoid}\left(\frac{(F_{c,i,j} - \text{mean})^2}{4(\text{var} + \lambda)} + 0.5 \right) \tag{4}$$

where λ is a small constant. The spatial attention mechanism obtains the weights of spatial attention through channel concatenation, convolution operations, and Sigmoid activation:

$$\text{Spatial_attention_weights} = \text{Sigmoid}(\text{Conv}([\text{avg_out}; \text{max_out}])) \quad (5)$$

Finally, the weights of channel attention, similarity activation, and spatial attention are fused with the input feature tensor through element-wise multiplication. This design, employing different attention mechanisms, enables the MCAM module to better allocate attention weights for different pixel-level information when dealing with multiscale feature relationships.

3.3. Layered Residual Connectivity Module

During the encoding process, as the downsampling deepens, the network might lose some detailed features, leading to unclear boundary delineation in the image. Relying solely on deep-level feature information may not accurately reconstruct the predicted image, so a combination of multiscale feature maps is needed to compensate for the lack of local feature information, enabling accurate pixel classification. To tackle this problem, our study has designed the layered residual connectivity module (LRCM). The structural schematic of the module is shown in Figure 3. This module enhances the network's ability to delineate edge portions by improving the skip-connection section of each hierarchical network, enabling the network to incorporate information from other hierarchical levels. The introduction of this module helps improve the network's perception of image details and achieve more accurate pixel classification in terms of local contextual information. The interrelationships between different LRCMs are illustrated in Figure 4.

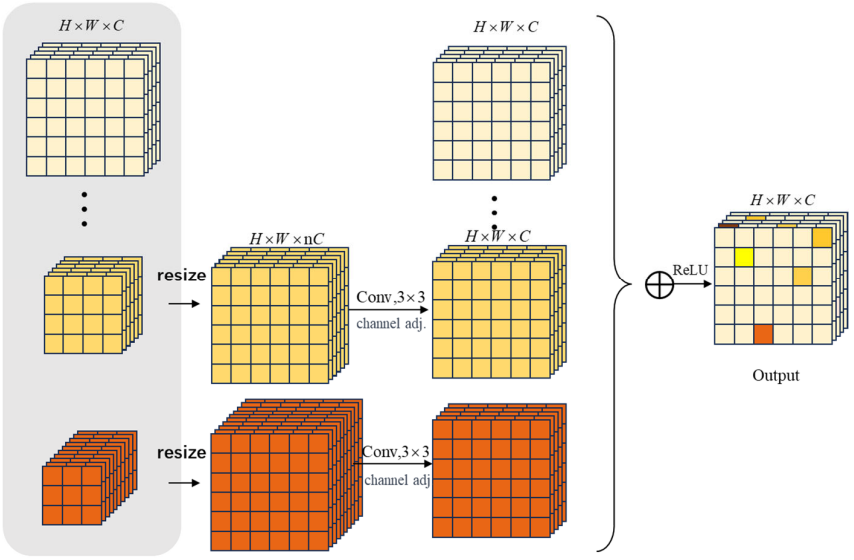


Figure 3. Layered residual connectivity module.

By introducing hierarchical residual connections, the aim is to enhance the model's feature transmission capacity and gradient flow efficiency. Using the bilinear interpolation method, the module resizes feature maps from different hierarchies to adapt to inputs with different resolutions.

$$Q = (1-x)(1-y)A + x(1-y)B + (1-x)yC + xyD \quad (6)$$

A, B, C, D represent the values of the points at the top-left, top-right, bottom-left, and bottom-right corners, respectively.

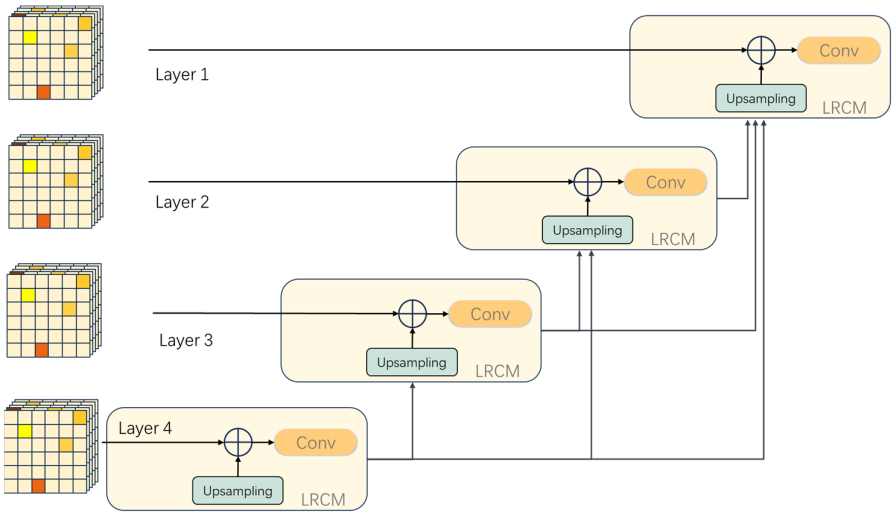


Figure 4. Schematic diagram of feature map transmission in LRCM.

Convolutional operations ($\mathcal{F}$), batch normalization (BN), and rectified linear units ($ReLU$) are employed to process input features. $\mathcal{H}$ represents the process of hierarchical residual feature fusion.

$$\text{Out} = \text{Input} + \mathcal{F}(ReLU(BN(\mathcal{F}(\text{Input})))) + \text{Resize}(\mathcal{H}(\text{Input})) \quad (7)$$

In this process, multiple stages of LRCM operations are introduced, and after each stage, feature size adjustment is performed to ensure the organic flow of information between different levels. This approach has significant advantages in enhancing feature transmission and mitigating gradient-related issues.

4. Experiments

4.1. Data and Hardware Environment

The COCO dataset [28] comprises over 200 K images. In our experiments, we utilized two datasets to validate the proposed method:

1. WHU dataset [37]. This dataset comprises two subsets of remote sensing images (aerial and satellite images). The aerial image subset was selected to validate the proposed method. The original aerial images were obtained from the New Zealand Land Information Service website, located in Christchurch, New Zealand, and contain 187,000 buildings with a total of 8188 images. Among these, 4736 images were used as the training set, 1036 images as the validation set, and 2416 images as the test set. Each image has a size of 512 × 512 pixels, a spatial resolution of 0.3 m, and includes three bands (red, green, and blue).
2. Massachusetts building dataset [38]. It comprises 151 remote sensing images of Boston city and suburbs, each with a size of 1500 × 1500 pixels. The dataset was divided into 136 images for model training, 11 for model testing, and the remaining ones for model validation. Due to memory constraints, the original image size exceeded the available memory, so the images were cropped to 512 × 512 in these experiments.

The experimental graphics processing unit used was the Tesla V100-PCIE, with a memory of 32 GB. The required software packages include Python 3.7, CUDA 11.4, the deep learning framework PyTorch 1.8.0, and related libraries such as cuDNN, OpenCV, Pandas, and NumPy.

4.2. Evaluation Metrics

The proposed building extraction network determined whether each pixel of the input remote sensing image belongs to the building category. *TP* and *TN* represent the cases of

positive and negative classes correctly identified by the model, respectively, while *FP* and *FN* represent false alarms and missed cases, respectively. Popular pixel-level evaluation metrics were used to quantitatively assess the network performance, including precision (*P*) (8), recall (*R*) (9), intersection over union (*IoU*) (10), and *F*1-score (11). The specific formulas are as follows:

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{10}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{11}$$

4.3. Experiment Analysis

1. Quantitative Results

In this section, we test the model's performance on two public datasets. The model's performance is first tested on the WHU dataset, and its segmentation capability is compared with other models.

The comparative results are outlined in Table 1. In the WHU aviation image dataset, the network proposed in our paper exhibited a significant performance improvement and demonstrated outstanding results. Its accuracy reached 95.0%, the recall rate was 95.2%, and the intersection over union (IoU) was 90.66%.

Table 1. Comparison of other commonly used semantic segmentation networks with the proposed network on the WHU aerial image dataset.

Method	IoU	Recall	Precision	F1
U-net [25]	88.26	93.66	94.34	93.77
Segnet [39]	85.20	91.55	92.48	92.00
DeepLabv3+ [40]	87.07	92.65	93.59	93.12
HRNet [41]	86.51	92.67	93.66	93.16
RSR-Net [42]	88.69	92.31	95.61	93.93
LightFGCNet-C [43]	89.87	94.57	-	94.86
DS-Net [44]	90.4	95.06	94.85	94.96
Ours	90.66	95.2	95.0	95.1

The F1 score increased by 1.33%, 2.56%, 1.98%, and 1.94% compared to U-Net, SegNet, DeepLabv3+, and HRNet, respectively. In comparison to these commonly used benchmark models, the proposed model in our paper demonstrated a notable enhancement across various performance metrics. Compared to the second-best performing DS-Net, MRAN's performance metrics were higher by 1.97%, 2.89%, −0.61%, and 1.17%, respectively.

From Table 2, we can observe that in the Massachusetts building dataset, the method proposed in our paper achieves an accuracy of 81.12%, a recall rate of 82.96%, an IoU of 69.53%, and an F1 score of 82.03. Compared to classic networks such as SegNet, Res-Unet, DeepLabv3+, and HRNet, the proposed method exhibits respective improvements in accuracy by 15.06%, 4.15%, −3.61%, and 1.61%. The recall rates show improvements of 0.21%, 0.38%, 3.86%, and 0.68% respectively. The IoU improvements are 11.46%, 3.32%, 0.3%, and 1.64%, and F1 score improvements are 8.56%, 2.36%, 0.21%, and 1.16%, respectively.

Table 2. Comparison between other semantic segmentation networks and the network proposed in our paper using the Massachusetts building dataset.

Method	IoU	Recall	Precision	F1
Segnet [39]	58.07	82.75	66.06	73.47
Res-Unet [45]	66.21	82.58	76.97	79.67
DeepLabv3+ [40]	69.23	79.1	84.73	81.82
HRNet [41]	67.89	82.28	79.51	80.87
Ours	69.53	82.96	81.12	82.03

In summary, the experimental results outlined above show that MRAN performs significantly better than classic networks such as U-Net, SegNet, DeepLabv3+, HRNet, and CFENet in terms of performance. In comparison to models like U-Net and DeepLabv3+ that utilize feature maps with insufficient semantic information during the upsampling process, MRAN cleverly reallocates weights in the skip connections through the effective use of attention mechanisms. This mechanism enhances the model's perceptual ability at various positions in the image, contributing to a more accurate restoration of detailed information. Its meticulously designed network structure takes into full consideration the distribution of semantic information and the multi-scale representation of features. By introducing attention mechanisms, it effectively integrates features of different scales. These advantages make MRAN an outstanding advanced model in building segmentation tasks.

2. Qualitative results

In the presented extraction results, black regions denote background areas, while white regions indicate buildings. Areas incorrectly predicted as buildings when they are background are highlighted in red, representing false positives. Conversely, areas where buildings are incorrectly predicted as background are marked in green, indicative of false negatives.

In the comparison samples, all models used for comparison exhibited some errors and omissions in the detection results. Specifically, for the test image Figure 5a, where buildings have roofs with special textures, U-Net, SegNet, DeepLabv3+, and HRNet showed less smooth performance in determining building boundaries, leading to the misidentification of surrounding ground items as buildings. Similarly, for the test image Figure 5c, where ground items are placed side by side with buildings, these networks failed to clearly define the building boundaries, while MRAN demonstrated smoother handling of boundaries with superior results. In remote sensing images, there is a tendency to misjudge items with shadows and trees, as seen in test image Figure 5b where buildings and trees are adjacent and shadows overlap. U-Net and SegNet incorrectly identified the shadow portion as a building, while MRAN handled this issue well. Additionally, in test image Figure 5d where containers are neatly arranged on the ground, DeepLabv3+ and HRNet incorrectly classified them as small buildings. In contrast, MRAN's building extraction results show a smoother delineation of building edges, effectively addressing cases of both false positives and false negatives.

Due to the poor imaging quality of the Massachusetts building dataset, the influence of lighting and shadows leads to indistinct boundaries of buildings, and there are numerous small-sized buildings in the images, making the visual similarity between target and background buildings challenging for segmentation tasks. The following is a comparison of segmentation results from different networks, as shown in the figures. In Figure 6a,d, the red circular areas represent regions with dark roofs and shadows on buildings. In these cases, other segmentation models like SegNet and Res-Unet exhibit issues of missing detection, defining these areas as multiple small buildings. MRAN, on the other hand, is able to more accurately delineate the boundaries of buildings in these areas, avoiding excessive segmentation of small buildings. In Figure 6b, multiple small buildings are arranged adjacent to each other, and other networks often define them as a single connected structure. Thanks to the LRCM module in MRAN, which provides a clear definition of fine boundaries, this

network finely delineates the edges of buildings, resulting in a more accurate definition of the number of buildings. These results highlight the robustness and efficiency of MRANet in handling the challenging Massachusetts building dataset, especially in scenarios with uneven lighting, severe shadows, and a high density of small buildings.

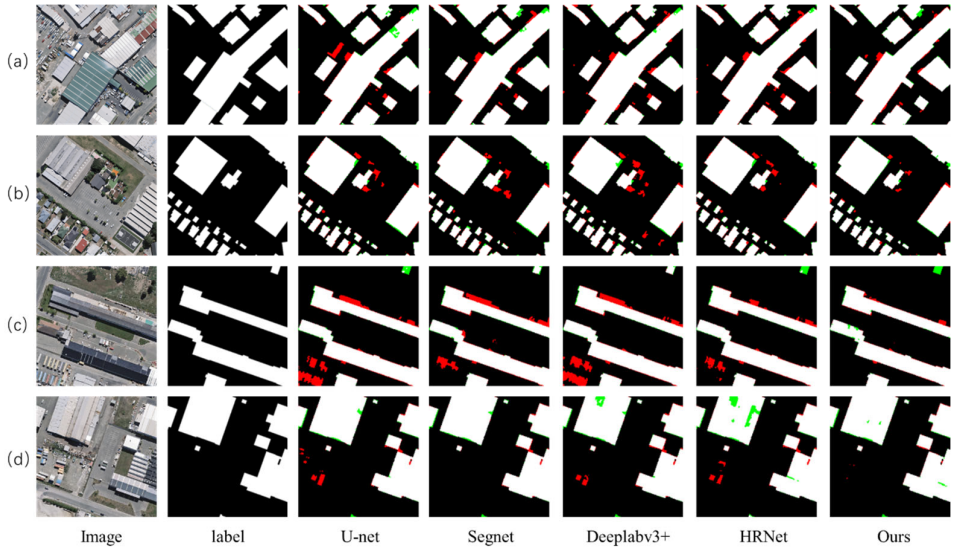


Figure 5. Qualitative comparison results of the WHU aerial imagery dataset. Visualization results of four remote sensing building images (**a**–**d**).

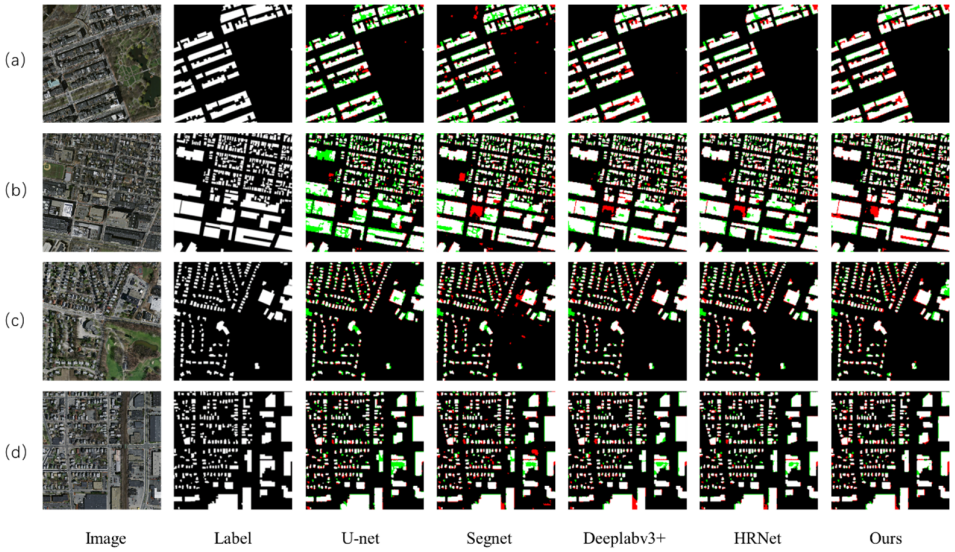


Figure 6. Qualitative comparison results of the Massachusetts building dataset. Visualization results of four remote sensing building images (**a**–**d**).

4.4. Ablation Study

1. Quantitative Results

We conducted ablation experiments on the WHU aerial dataset in this study to assess the effectiveness of designing various modules and search for the contributions of these modules within our network. In comparison to the baseline [25], this section tested the roles of MCAM and LRCM. The experimental results are shown in the Table 3.

Table 3. Ablation experiments on WHU dataset.

Method	IoU	Recall	Precision	F1
Baseline	87.02	92.87	93.25	93.06
Baseline + MCAM	89.15	94.62	93.92	94.27
Baseline + LRCM	88.97	94.75	93.58	94.16
Baseline + MCAM + LRCM	90.66	95.2	95.0	95.1

The results show that when only adding the LRCM structure, the four evaluation metrics are higher than the baseline by 1.95%, 1.88%, 0.28%, and 1.1%, respectively. Adding only the MCAM achieves an IoU of 89.15%, a precision of 93.92%, and an F1-score of 94.27%, which are higher than the baseline by 2.13%, 0.68%, and 1.21%, respectively. When continuing to add both MCAM and LRCM structures to the network, the performance is further improved, with the IoU, precision, and F1-score being higher than the baseline by 3.54%, 2.75%, and 2.04%, respectively. All metrics are higher than when using a single structure. The conflict between these two modules is relatively small, and both can independently improve the model's performance.

2. Qualitative results

For the qualitative analysis of the performance of each module, a qualitative comparison was conducted among four models: Baseline, Baseline + MCAM, Baseline + LRCM, and Baseline + MCAM + LRCM. The visualized extraction results of the network were analyzed to compare the impact of each module, as shown in Figure 7.

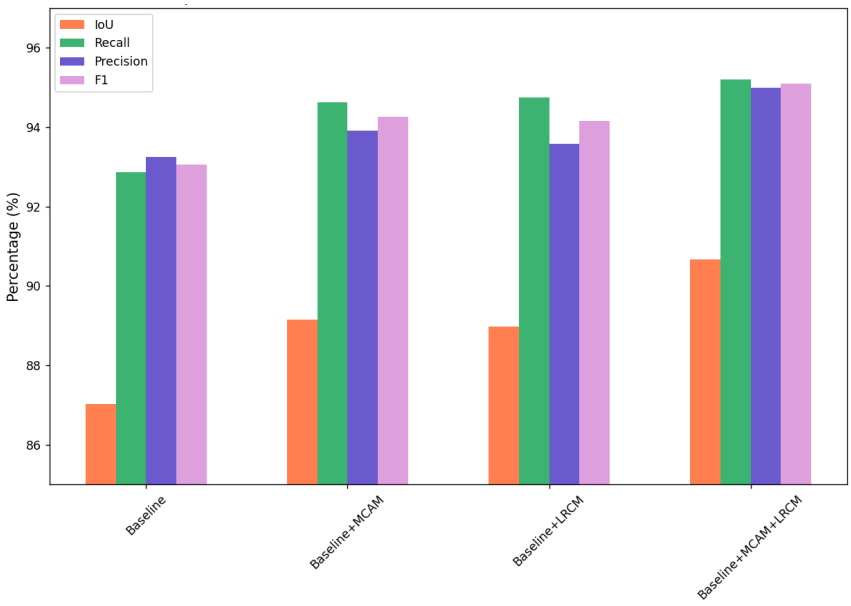


Figure 7. Ablation experiments on WHU dataset.

According to Figure 7, the illustration shows that the MCAM attention module demonstrates a significant effect on building extraction. It effectively prevents the neglect of some small buildings and enhances the weights for occluded buildings. Specifically, as shown in Figure 8a,d, the MCAM module notices small buildings in the images and integrates them into the final prediction by strengthening attention. In Figure 8d, where buildings are obscured by trees, MCAM adjusts attention weights through reallocation, enabling the model to predict partially obscured buildings.

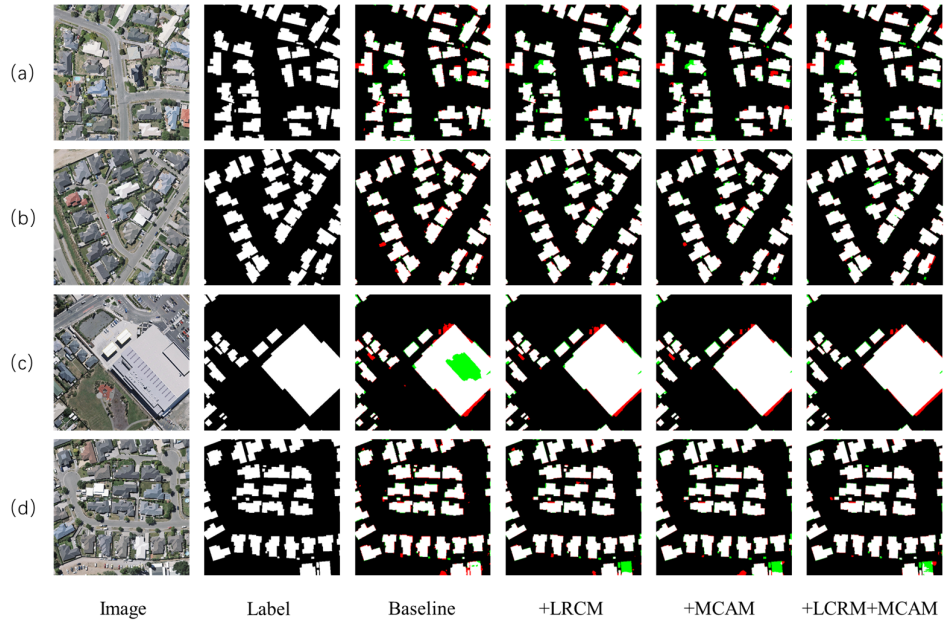


Figure 8. Example of results with the Baseline, Baseline + MCAM, Baseline + LRCM, and Baseline + MCAM + LRCM. Visualization results of four remote sensing building images (**a**–**d**).

Additionally, the LRCM receives semantic information at different scales, and after fusing this information, it helps differentiate building boundaries, effectively dividing building edges and ensuring smoothness. For example, in the red portions in Figure 8b, LRCM can discern protruding parts of buildings to determine whether they are independent structures. This helps prevent connected buildings from being erroneously identified as independent, ensuring accurate predictions of the number of buildings.

By fully leveraging the advantages of both the MCAM and LRCM modules and successfully integrating them, the model avoids missing buildings and accurately delineates boundaries, resulting in outstanding building extraction results.

5. Conclusions

Our study tackles the challenge of low accuracy in building segmentation within remote sensing image extraction, which is often caused by factors such as blurry boundaries, tree cover, and shadow obstructions. A solution named the multi-resolution attention fusion network is proposed. The network extracts semantic information of buildings in remote sensing images during the encoder phase and progressively upsamples through the decoder to restore feature maps. In the skip-connection part, a multi-channel attention module and a mechanism for multi-scale feature fusion are introduced. The uniqueness of this network design lies in its effective fusion of semantic information at different scales, assigning varying attention weights to enhance the extraction capability of building boundaries and contextual understanding. This design effectively improves the smoothness of the network in capturing target boundaries and enhances the accuracy in predicting obscured buildings. To validate the performance of the proposed algorithm, the researchers conducted quantitative experiments, qualitative experiments, and ablation experiments on publicly available aerial image datasets and the Massachusetts dataset. The experimental results demonstrate the effectiveness of the algorithm in improving the accuracy and robustness of building segmentation.

Author Contributions: Conceptualization, H.C. (Hongyang Chen); Data curation, J.L., H.G. and H.C. (Hongyang Chen); Formal analysis, H.G. and H.C. (Hao Chen); Funding acquisition, J.L.; Methodology, J.L., H.G., H.C. (Hongyang Chen) and H.C. (Hao Chen); Project administration, J.L. and Z.L.; Software, Z.L.; Supervision, Z.L.; Visualization, Z.L. and H.C. (Hongyang Chen); Writing—original draft, H.G. and H.C. (Hao Chen); Writing—review & editing, H.C. (Hongyang Chen). All authors have read and agreed to the published version of the manuscript.

Funding: This paper was supported by the Henan Provincial Science and Technology Research Project under Grant 222102210304, 232102211006, 232102210044 and 232102211017, the Science and Technology Innovation Project of Zhengzhou University of Light Industry under Grant 23XNKJTD0205, the Undergraduate Universities Smart Teaching Special Research Project of Henan Province under Grant Jiao Gao [2021] No. 489-29.

Data Availability Statement: The WHU Building Aerial Imagery and Massachusetts Buildings datasets used in the experiment can be downloaded at http://gpcv.whu.edu.cn/data/building_dataset.html (accessed on 25 February 2024) and https://www.cs.toronto.edu/~vmnih/data/ (accessed on 25 February 2024), respectively.

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

References

1. Kotaridis, I.; Lazaridou, M. Remote sensing image segmentation advances: A meta-analysis. *ISPRS J. Photogramm. Remote. Sens.* **2021**, *173*, 309–322. [CrossRef]
2. Ok, A.O. Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. *ISPRS J. Photogramm. Remote. Sens.* **2013**, *86*, 21–40. [CrossRef]
3. Zhao, W.; Persello, C.; Stein, A. Building instance segmentation and boundary regularization from high-resolution remote sensing images. In Proceedings of the IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 3916–3919.
4. Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Huang, H.; Zhu, J.; Liu, Y.; Li, H. DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 1194–1206. [CrossRef]
5. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2017**, *60*, 84–90. [CrossRef]
6. Abdollahi, A.; Pradhan, B. Integrating semantic edges and segmentation information for building extraction from aerial images using UNet. *Mach. Learn. Appl.* **2021**, *6*, 100194. [CrossRef]
7. 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.
8. Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Multi-Resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks. *Remote. Sens.* **2018**, *10*, 1636. [CrossRef]
9. Dong, S.; Chen, Z. A Multi-Level Feature Fusion Network for Remote Sensing Image Segmentation. *Sensors* **2021**, *21*, 1267. [CrossRef] [PubMed]
10. Zheng, Z.; Du, S.; Taubenböck, H.; Zhang, X. Remote sensing techniques in the investigation of aeolian sand dunes: A review of recent advances. *Remote Sens. Environ.* **2022**, *271*, 112913. [CrossRef]
11. Yuan, X.; Shi, J.; Gu, L. A Review of Deep Learning Methods for Semantic Segmentation of Remote Sensing Imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
12. Cheng, Y.; Wang, W.; Zhang, W.; Yang, L.; Wang, J.; Ni, H.; Guan, T.; He, J.; Gu, Y.; Tran, N.N. A Multi-Feature Fusion and Attention Network for Multi-Scale Object Detection in Remote Sensing Images. *Remote. Sens.* **2023**, *15*, 2096. [CrossRef]
13. Inglada, J. Automatic Recognition of Man-Made Objects in High Resolution Optical Remote Sensing Images by SVM Classification of Geometric Image Features. *ISPRS J. Photogramm. Remote. Sens.* **2007**, *62*, 236–248. [CrossRef]
14. Cetin, M.; Halici, U.; Aytekin, O. Building detection in satellite images by textural features and Adaboost. In Proceedings of the 2010 IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS 2010), Istanbul, Turkey, 22–22 August 2010; pp. 1–4.
15. Peng, J.; Liu, Y.C. Model and Context-Driven Building Extraction in Dense Urban Aerial Images. *Int. J. Remote. Sens.* **2007**, *26*, 1289–1307. [CrossRef]
16. Wei, Y.; Zhao, Z.; Song, J. Urban Building Extraction from High-Resolution Satellite Panchromatic Image Using Clustering and Edge Detection. In Proceedings of the IGARSS 2004—2004 IEEE International Geoscience and Remote Sensing Symposium, Anchorage, AK, USA, 20–24 September 2004; Volume 3, pp. 2008–2010.
17. Li, E.; Femiani, J.; Xu, S.; Zhang, X.; Wonka, P. Robust rooftop extraction from visible band images using higher order CRF. *IEEE Trans. Geosci. Remote. Sens.* **2015**, *53*, 4483–4495. [CrossRef]
18. Du, S.; Zhang, F.; Zhang, X. Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach. *ISPRS J. Photogramm. Remote. Sens.* **2015**, *105*, 107–119. [CrossRef]

19. Gavankar, N.L.; Ghosh, S.K. Automatic building footprint extraction from high-resolution satellite image using mathematical morphology. *Eur. J. Remote. Sens.* **2018**, *51*, 182–193. [CrossRef]
20. Xu, L.; Kong, M.; Pan, B. Building Extraction by Stroke Width Transform from Satellite Imagery. In Proceedings of the Second CCF Chinese Conference Computer Vision CCCV 2017, Tianjin, China, 11–14 October 2017; Springer: Singapore, 2017; pp. 340–351.
21. Cheng, B.; Cui, S.; Ma, X.; Liang, C. Research on an Urban Building Area Extraction Method with High-Resolution PolSAR Imaging Based on Adaptive Neighborhood Selection Neighborhoods for Preserving Embedding. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 109. [CrossRef]
22. Dai, Y.; Gong, J.; Li, Y.; Feng, Q. Building Segmentation and Outline Extraction from UAV Image-Derived Point Clouds by a Line Growing Algorithm. *Int. J. Digit. Earth* **2017**, *10*, 1077–1097. [CrossRef]
23. Adegun, A.A.; Viriri, S.; Tapamo, J.-R. Review of Deep Learning Methods for Remote Sensing Satellite Images Classification: Experimental Survey and Comparative Analysis. *J. Big Data* **2023**, *10*, 9. [CrossRef]
24. Liu, S.; Shi, Q.; Zhang, L. Few-Shot Hyperspectral Image Classification with Unknown Classes Using Multitask Deep Learning. *IEEE Trans. Geosci. Remote. Sens.* **2020**, *59*, 5085–5102. [CrossRef]
25. 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, Munich, Germany, 5–9 October 2015.
26. Tong, Z.; Li, Y.; Li, Y.; Fan, K.; Si, Y.; He, L. New Network Based on Unet++ and Densenet for Building Extraction from High Resolution Satellite Imagery. In Proceedings of the 2020 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Waikoloa, HI, USA, 26 September 2020; pp. 2268–2271.
27. Dey, M.S.; Chaudhuri, U.; Banerjee, B.; Bhattacharya, A. Dual-Path Morph-UNet for Road and Building Segmentation From Satellite Images. *IEEE Geosci. Remote. Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
28. 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.* **2019**, *40*, 834–848. [CrossRef] [PubMed]
29. 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]
30. 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]
31. Zhang, C.; Jiang, W.S.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C.J. Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote. Sens.* **2022**, *60*, 1–20. [CrossRef]
32. 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*, 1–17. [CrossRef]
33. Liu, T.; Yao, L.; Qin, J.; Lu, N.; Jiang, H.; Zhang, F.; Zhou, C. Multi-Scale Attention Integrated Hierarchical Networks for High-resolution Building Footprint Extraction. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *109*, 102768. [CrossRef]
34. Wang, Y.; Zeng, X.; Liao, X.; Zhuang, D. B-FGC-Net: A Building Extraction Network from High Resolution Remote Sensing Imagery. *Remote. Sens.* **2022**, *14*, 269. [CrossRef]
35. Ku, T.; Yang, Q.; Zhang, H. Multilevel Feature Fusion Dilated Convolutional Network for Semantic Segmentation. *Int. J. Adv. Robot. Syst.* **2021**, *18*, 20. [CrossRef]
36. Zhang, R.; Zhang, Q.; Zhang, G. SDSC-UNet: Dual Skip Connection ViT-based U-shaped Model for Building Extraction. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
37. 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]
38. Huang, Z.; Cheng, G.; Wang, H.; Li, H.; Shi, L.; Pan, C. Building Extraction from Multi-Source Remote Sensing Images Via Deep Deconvolution Neural Networks. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1835–1838.
39. 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] [PubMed]
40. 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 Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
41. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Wang, J. High-Resolution Representations for Labeling Pixels and Regions. *arXiv* **2019**, arXiv:1904.04514.
42. Huang, H.; Chen, Y.; Wang, R. A Lightweight Network for Building Extraction from Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [CrossRef]
43. Chen, Y.; Jiang, W.; Wang, M.; Kang, M.; Weise, T.; Wang, X.; Zhang, C. LightFGCNet: A Lightweight and Focusing on Global Context Information Semantic Segmentation Network for Remote Sensing Imagery. *Remote. Sens.* **2022**, *14*, 6193. [CrossRef]

44. Guo, H.; Su, X.; Tang, S.; Du, B.; Zhang, L. Scale-Robust Deep-Supervision Network for Mapping Building Footprints From High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2021**, *14*, 10091–10100. [CrossRef]
45. Zhang, Z.; Liu, Q.; Wang, Y. Road Extraction by Deep Residual U-Net. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 749–753. [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.