

Article

ME-FCN: A Multi-Scale Feature-Enhanced Fully Convolutional Network for Building Footprint Extraction

Hui Sheng ¹ , Yaoteng Zhang ¹, Wei Zhang ^{1,2,*}, Shiqing Wei ¹ , Mingming Xu ¹  and Yasir Muhammad ¹

¹ College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266404, China; sheng@upc.edu.cn (H.S.); z22160012@s.upc.edu.cn (Y.Z.); wei_sq@upc.edu.cn (S.W.); xumingming@upc.edu.cn (M.X.)

² Land Surveying and Mapping Institute of Shandong Province, Jinan 250102, China

* Correspondence: zhangweightchy@shandong.cn

Abstract: The precise extraction of building footprints using remote sensing technology is increasingly critical for urban planning and development amid growing urbanization. However, considering the complexity of building backgrounds, diverse scales, and varied appearances, accurately and efficiently extracting building footprints from various remote sensing images remains a significant challenge. In this paper, we propose a novel network architecture called ME-FCN, specifically designed to perceive and optimize multi-scale features to effectively address the challenge of extracting building footprints from complex remote sensing images. We introduce a Squeeze-and-Excitation U-Block (SEUB), which cascades multi-scale semantic information exploration in shallow feature maps and incorporates channel attention to optimize features. In the network's deeper layers, we implement an Adaptive Multi-scale feature Enhancement Block (AMEB), which captures large receptive field information through concatenated atrous convolutions. Additionally, we develop a novel Dual Multi-scale Attention (DMSA) mechanism to further enhance the accuracy of cascaded features. DMSA captures multi-scale semantic features across both channel and spatial dimensions, suppresses redundant information, and realizes multi-scale feature interaction and fusion, thereby improving the overall accuracy and efficiency. Comprehensive experiments on three datasets demonstrate that ME-FCN outperforms mainstream segmentation methods.

Keywords: building extraction; deep learning; multi-scale feature; semantic segmentation



Citation: Sheng, H.; Zhang, Y.; Zhang, W.; Wei, S.; Xu, M.; Muhammad, Y. ME-FCN: A Multi-Scale Feature-Enhanced Fully Convolutional Network for Building Footprint Extraction. *Remote Sens.* **2024**, *16*, 4305. <https://doi.org/10.3390/rs16224305>

Academic Editor: Wen Liu

Received: 2 October 2024

Revised: 30 October 2024

Accepted: 8 November 2024

Published: 19 November 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

In urban areas, buildings cover more than 80% of the Earth's surface and constitute one of the most critical land cover types. Building information is essential in various domains, including digital urban planning, smart city development, the detection of illegal construction, and the monitoring of dynamic urban change [1–4]. Consequently, the identification and extraction of buildings has remained a central topic in scientific research.

The emergence of large-scale and diverse remote sensing data makes traditional building extraction methods unable to meet the needs. Methods such as morphology, shape, color, and texture analysis show limited generalization and accuracy [5–11]. In recent years, Convolutional Neural Networks (CNNs) have shown good results in the semantic segmentation of remote sensing images due to their strong generalization and self-learning capabilities [12]. Therefore, some scholars have applied semantic segmentation to building extraction tasks and have achieved good results [13–17]. Thanks to contrastive learning [18,19], emerging basic models such as SAM [20] and DINO v2 [21] can be trained on large-scale datasets. Therefore, they show stronger feature extraction and generalization capabilities. However, training them usually requires a large amount of high-quality data, and the parameters of these models are much larger than those of CNN-based methods.

Although significant progress has been made in building extraction using deep learning methods, accurately extracting buildings of complex styles and structures from high-resolution

remote sensing images remains challenging [22]. Features with large receptive fields lack detailed information, while local features cannot fully extract buildings. Therefore, most studies focus on the integration of multi-scale features to improve the generalizability and accuracy of the network [23–26]. Multi-scale features incorporate both global context and local details, aiding the network in precise localization and segmentation.

The transformer effectively extracts global context information through the attention mechanism, enabling the modeling of long-range dependencies [27]. Swin-Unet [28] integrates the Swin Transformer into a U-shaped architecture, achieving self-attention from local to global contexts in the encoder, which enhances the model's extraction accuracy. However, the computational complexity of transformer-based methods grows quadratically with increasing sequence length. Consequently, some studies [29,30] have proposed combining CNNs with transformers to improve network performance. TransUNet and UNetformer combine CNN encoders with transformer decoders to extract multi-scale features, thus increasing model accuracy [30,31]. However, the accuracy of these models shows limited generalization ability, and the performance needs to be further optimized.

Pyramid pooling modules are commonly used for multi-scale feature extraction, providing a larger receptive field [32]. However, excessive downsampling can result in the loss of local information. Compared with pyramid pooling modules, some studies introduce atrous convolution to achieve feature extraction on different scales [33,34]. However, the combination of hybrid convolutions causes uneven pixel value utilization and reduces the accuracy of features. There are also studies [35–37] that achieve multi-scale feature extraction through cross-layer fusion and greatly improve the accuracy and flexibility of the network. But does single-scale feature extraction based on the input features resolution fully mine semantic information? And how to effectively optimize semantic information?

We designed a powerful Squeeze-and-Excitation U-shaped block (SEUB-N, where N represents the depth of the module) to extract shallow multi-scale information. SEUB can extract multi-scale features in stages without reducing the resolution of the feature map. Additionally, the embedded squeeze-and-excitation (SE) attention modules [38] filter out redundant information and noise, further purifying the multi-scale features. SEUB maximizes the mining of semantic information from the input feature map, providing rich multi-scale information for building extraction. To address the uneven pixel value utilization caused by the combination of hybrid convolutions, we design an Adaptive Multi-scale feature Enhancement Block (AMEB). AMEB introduces a novel branch allowing the network to autonomously assign weights to various receptive field features. This approach reduces randomness in information utilization and enhances adaptive receptive field selection.

The complex background of remote sensing images presents another significant challenge for the accurate extraction of building characteristics [39]. Environmental factors such as shadows and noise can distort the extracted features. This distortion leads to inaccuracies in the description of buildings and introduces erroneous information during the subsequent fusion of features [40,41]. Attention mechanisms can capture global information and integrate data from various hierarchical levels, making it easier to extract internal features. Li et al. [42] designed a dual-branch attention mechanism network to capture both spectral and spatial features, achieving refinement and optimization of feature maps. Xu et al. [43] utilized attention-guided extraction of high-frequency features to enrich detail. Yu et al. [44] integrated attention gates and feature pyramid attention into the U-Net structure to suppress irrelevant information, and expand the receptive field of high-dimensional features. However, the fusion of features at different scales is challenged by discrepancies in semantic information. Current attention-based methods struggle to achieve adaptive interaction and fusion of multi-scale features, and research on multi-scale attention remains limited. To address this issue, this paper proposes a novel multi-scale attention module, Dual Multi-scale Attention (DMSA). DMSA first aligns encoder features with upsampled decoding features using convolutions at three different scales, effectively preventing semantic confusion caused by scale differences. Furthermore, DMSA employs

channel and spatial attention mechanisms to capture critical features, further optimizing the feature and avoiding semantic overlap.

In summary, our work makes the following main contributions:

(1) We introduce an innovative ME-FCN for semantic segmentation of buildings. ME-FCN adopts efficient multi-scale feature extraction and adaptive feature optimization, focusing on the scale correlation of buildings, and achieves accurate building footprint extraction.

(2) We design two powerful multi-scale feature extraction modules, SEUB and AMEB. SEUB and AMEB can adaptively extract and optimize multi-scale features, thus enabling more accurate extraction of building footprints.

(3) We propose a novel attention module named DMSA, which can capture scale sensitivity in both spatial and channel dimensions and realize the interaction between multi-scale features. By adaptively optimizing features, the DMSA module effectively fuses multi-scale features and enhances building representation.

2. Methodology

This study proposes a Multi-scale feature Enhancement Fully Convolutional Network (ME-FCN) for extracting building footprints from remote sensing images, as illustrated in Figure 1.

For each image, we first use the encoder to extract multi-scale features at different resolutions, employing four different-sized Squeeze-and-Excitation U-blocks (SEUBs) and two Adaptive Multi-scale feature Enhancement Blocks (AMEBs). The network also integrates Dual Multi-scale Attention (DMSA) between the encoder and decoder, linking low-level features with high-level semantic information, thereby enhancing the inference of global contextual details and enabling a more accurate representation of multi-scale building features. The decoder, consisting of four SEUBs of different sizes and two AMEBs, generates six building footprint outputs that are then aggregated and processed through convolution and activation to produce the final result.

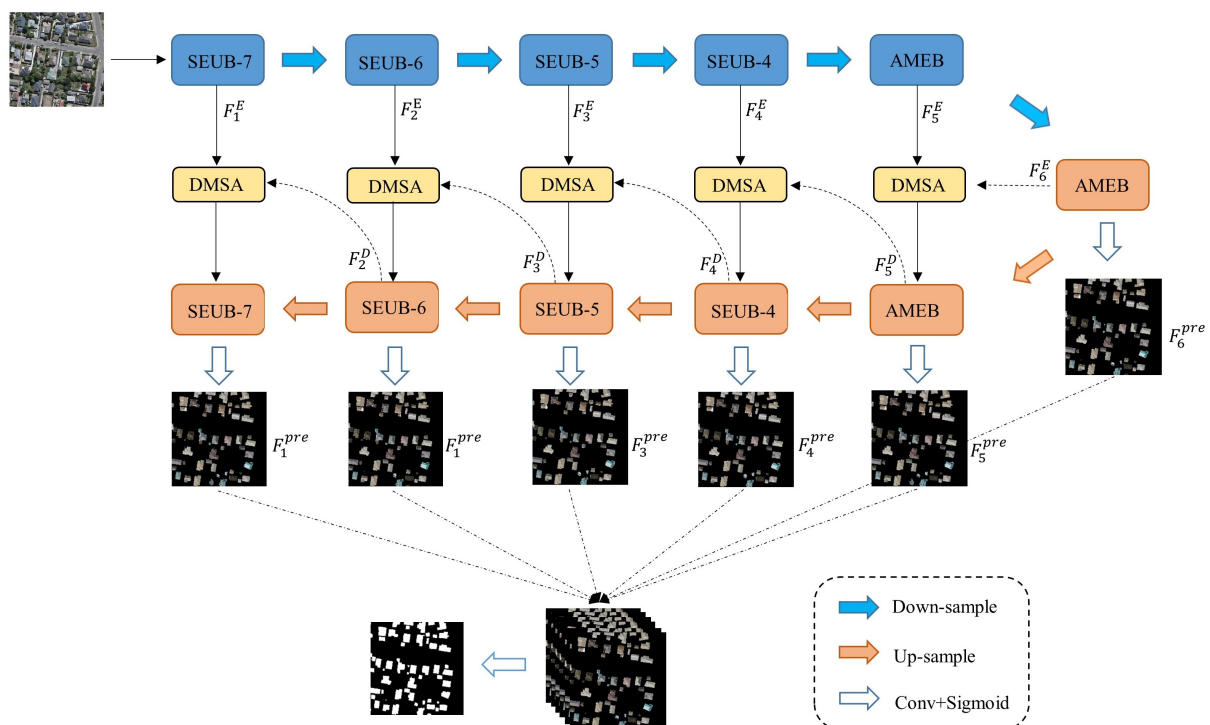


Figure 1. Architecture of the proposed ME-FCN method. (Blue represents the encoder and orange represents the decoder.)

2.1. The Proposed ME-FCN Framework

The proposed ME-FCN is a novel and efficient framework for building footprint extraction, demonstrating robust multi-scale feature extraction capabilities that fully exploit semantic information on different scales. As illustrated in Figure 1, the ME-FCN architecture consists of an encoder network, a DMSA feature enhancement module, and decoder blocks. Unlike previous encoders, our encoder network can extract multi-scale features at arbitrary resolutions and is equipped with different encoding structures (SEUB-N, AMEB) tailored for input feature maps of varying resolutions, where 'N' represents the depth of the module. The original image is first processed by the encoder, passing through SEUB-7 to generate the feature map F_1^E . Subsequently, max-pooling operations reduce the resolution of the feature map by half, and the output is fed into the next encoding structure, producing multiple multi-scale feature maps F_i^E , where $i \in \{1, 2, 3, 4, 5\}$.

The extracted multi-scale features are then optimized using the DMSA module, where they undergo adaptive refinement at both channel and pixel levels, enhancing the interaction of multi-scale features. This process suppresses redundant information and noise, resulting in further purification of the features. The optimized features are subsequently concatenated with the decoded output F_i^D , which has been upsampled and is fed into the $i - 1$ decoder block via skip connections. Finally, the network generates six predicted outputs F_i^{pre} , which are concatenated, convolved, and activated to produce the footprints of the building.

2.2. Squeeze-and-Excitation U-Shaped Block

In previous studies on multi-scale feature fusion, there has been a tendency to extract detailed information from shallow-level feature maps and scale-related information from deep-level feature maps. Typically, features with a larger receptive field are concatenated with shallow features through upsampling to enhance the receptive field. However, we question whether extracting single-scale information at the input feature map resolution fully leverages the semantic information contained within that feature map. Although U2-Net [45] enhances the model's multi-scale feature extraction capabilities, the mined multi-scale information may contain redundant information and noise, further reducing the network extraction accuracy.

We designed an SEUB module that can simultaneously realize multi-scale feature extraction and optimization. As depicted in Figure 2, SEUB adopts a simple U-shaped design, similar to U-Net. Unlike methods that only extract single-scale information at the input feature map resolution, the SEUB structure captures large receptive field features in shallow layers, thereby fully leveraging semantic information. The input features first undergo convolution, batch normalization (BN), and ReLU operations to obtain local information x^1, x^2 . Subsequently, through operations like downsampling convolution, different-scale features x^i , where $i \in (3, 4, 5)$, are obtained.

$$x^i = \text{ReLU}(\theta(\text{Down}(x^{(i-1)}))) \quad (1)$$

where Down represents downsampling, θ represents convolution and batch normalization (BN).

In particular, when obtaining x^6 , we use atrous convolutions with a stride of 2 to capture large-scale building features. To fuse multi-scale features, we employ a layer-by-layer aggregation strategy and improve network stability and noise resistance by incorporating SE modules into skip connections. The SE module optimizes the noise and redundant information in the shallow features, making the multi-scale features generated by fusion purer. The SE module operates in two steps: first, it applies average pooling to capture the global receptive field of the input feature maps, as illustrated in Equation (2):

$$z = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X_{in}(i, j) \quad (2)$$

where X_{in} is the input feature map, H is the height of the feature maps, and W is the width of the feature map. In this way, we can get the 1-D channel feature z .

In the second step, the SE module employs two fully connected bottleneck structures to derive channel correlations, which are then mapped to scaling factors using a sigmoid function.

$$s = Sigmoid(L_2 ReLU(L_1 z)) \quad (3)$$

where L_1 is the first fully connected layer, which reduces the input dimension, and L_2 is the second fully connected layer, which completes the operation of dimension recovery, obtaining the weight s of each feature channel through the sigmoid function. Finally, the weight s is multiplied by the input feature X_{in} , as follows:

$$X_{out} = X_{in} \times s \quad (4)$$

After passing through the SE module, the feature map x^5 is aggregated with the feature map x^6 of the deep layer. This aggregated feature is then processed through convolution, batch normalization (BN), and ReLU operations to produce y^4 . Subsequently, the aggregated feature y^i is combined with the output feature x^i of the shallow layer to generate multi-scale features.

$$y^{(i-1)} = \text{ReLU}(\theta(\text{Concat}(\text{Up}(y^i), \text{SE}(x^i)))) \quad (5)$$

where Up represents the process of upsampling, SE refers to the SE module, and $Concat$ denotes channel-wise concatenation.

The SEUB architecture incorporates a residual concept by replacing conventional single-stream convolutions with a U-shaped structure. It employs a feature representation of a weight layer instead of the original features:

$$F_{SEUB} = F_1(x) + SEUB(F_1(x)) \quad (6)$$

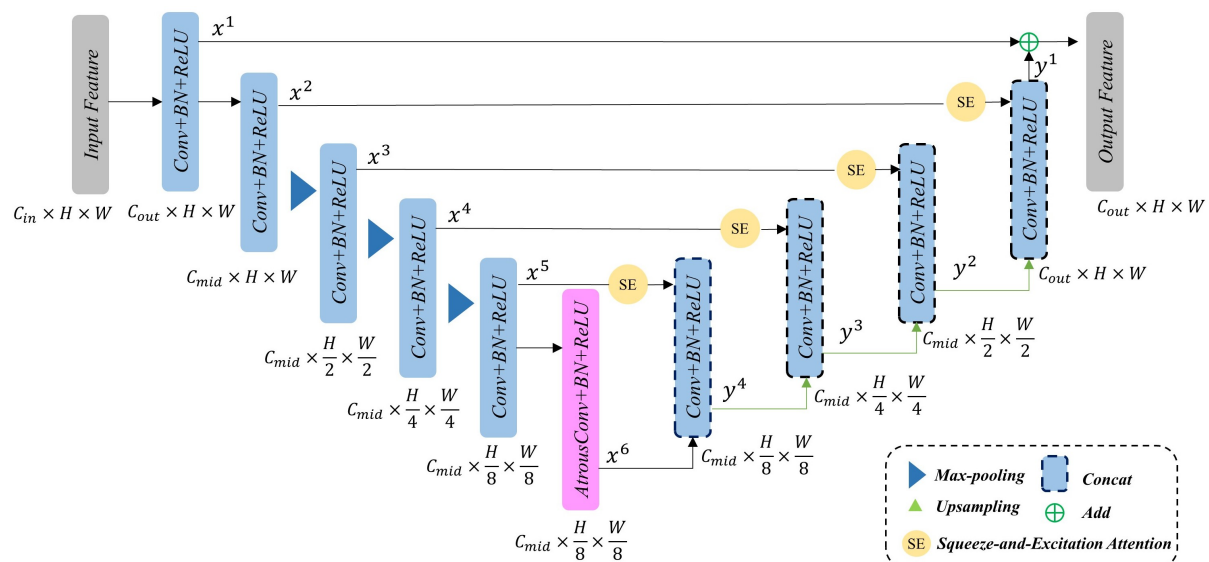


Figure 2. Overall architecture of the SEUB. SEUB adopts a simple U-shaped structure, with the bottom layer utilizing atrous convolution to achieve a larger receptive field, while the skip connections incorporate squeeze-and-excitation attention to enhance network stability.

As illustrated in Figure 3, the SEUB architecture incorporates a residual concept by replacing conventional single-stream convolutions with a U-shaped structure. It employs a feature representation of a weight layer instead of the original features:

$$F_{SEUB} = F_1(x) + seu(F_1(x)) \quad (7)$$

This design allows the network to extract multi-scale features from each residual block. Unlike multiple expansion convolutions applied to input feature maps at their original resolution, which demand significant computational and memory resources, our structure incurs minimal computational costs as most operations are performed on downsampled feature maps.

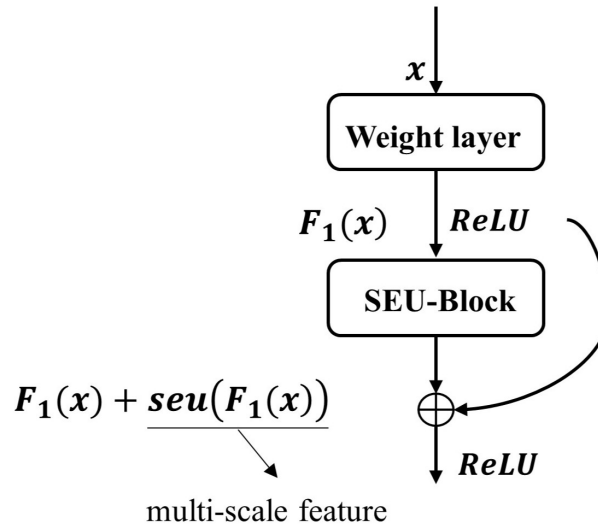


Figure 3. Residual structure schematic of SEUB.

2.3. Adaptive Multi-Scale Feature Enhancement Module

In the last two layers of our network, given that the input feature maps to AMEB have already been downsampled by factors of 16 or 32, we use multi-scale cascaded atrous convolutions instead of further downsampling. Atrous convolutions are beneficial for achieving a larger receptive field while maintaining resolution, which is crucial for effective multi-scale information extraction. However, this approach introduces a challenge: zero-padding can lead to a checkerboard pattern in the receptive field coverage, especially with increasing dilation rates, resulting in a loss of semantic information.

To mitigate this challenge, we first use convolution to obtain initial feature maps, ensuring the preservation of semantic information. Next, we apply a series of atrous convolutions with varying dilation rates $d \in \{2, 4, 8\}$ to generate feature maps with multiple receptive fields. Finally, we integrate these feature maps with the initial features.

However, the combination of hybrid convolutions leads to uneven utilization of feature pixel values, and preset dilation rates result in probabilistic levels of information utilization. Additionally, the preset dilation rate sometimes leads to excessive interference with information, thereby reducing the effective utilization of information.

To address this issue, we introduce a large-scale feature extraction module called the Adaptive Multi-scale feature Enhancement Block (AMEB). AMEB, as shown in Figure 4, employs a cascading strategy of multi-scale atrous convolutions. The AMEB introduces a novel branch that enables the network to autonomously assign weights to different receptive field features, thereby reducing randomness in information utilization and enhancing adaptive field selection. After Global Average Pooling (GAP), the feature map F is processed through two fully connected layers and ReLU for adaptation. The Softmax layer then calculates two scaling factors, λ_1 and λ_2 , which are applied to the multi-scale feature

maps. Finally, the refined high-level feature map is concatenated with the initial feature map to produce the output

$$F_{AMEB} = \lambda_1 \cdot F_1 + \lambda_2 \cdot F_2 \quad (8)$$

With AMEB, information utilization is autonomously optimized based on the relationship between high-level feature maps and initial features. The expansion of receptive fields avoids introducing confounding information and supports the accurate representation of large building structures. Additionally, the process of aggregating multi-scale features ensures that obtaining features with a large receptive field does not compromise detailed information.

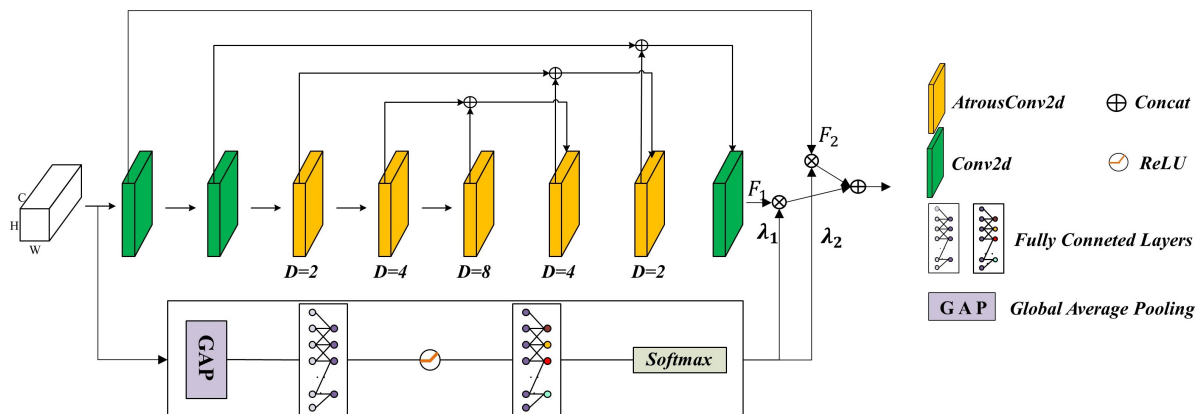


Figure 4. Schematic diagram of the Adaptive Multi-scale feature Enhancement Block. The AMEB consists of atrous convolutions with multiple dilation rates and introduces an auxiliary branch, allowing high-level feature maps to autonomously select information utilization, reducing the randomness of feature utilization.

2.4. The Dual Multi-Scale Attention Module

The U-shaped structure decodes features directly fused with encoder output features through upsampling to create multi-scale features. However, while the resolutions of decoding and encoding features match, there are inherent scale differences in feature maps from different scales, which can lead to semantic confusion during direct fusion. To address this issue, we developed the Dual Multi-scale Attention (DMSA) module, as illustrated in Figure 5. DMSA first extracts features at three different scales and aligns the two multi-scale features through spatial and channel dimensions. It then optimizes these features based on attention mechanisms, effectively mitigating scale differences and enhancing the features of the region of interest. Specifically, the DMSA module employs a parallel structure to separately extract multi-scale features at both channel and spatial levels.

In the channel attention component, feature maps are first processed through convolutional layers at three different scales to produce multi-scale feature maps along the channel dimension. Each branch operates at channel dimensions of $(2C, C, C)$, corresponding to scale sizes of $(1, 3, 5)$. This approach allows each branch to independently learn multi-scale spatial information and generate corresponding feature maps:

$$F_i = \text{Conv2d}(k, d)(X) \quad (9)$$

where k denotes the kernel size of the convolution, and d represents the dilation rate. Multi-scale features are generated by concatenating feature maps with different kernel sizes and dilation rates:

$$F_i = \text{Concat}(F_0, F_1, F_2) \quad (10)$$

Next, Efficient Channel Attention [46] is applied to generate attention maps for these feature maps. Channel attention weights are extracted from feature maps, yielding weight vectors for different scales. These attention weights can be represented as

$$att_i^{ch} = ECA(F_i), \quad i = 0, 1, 2 \quad (11)$$

We concatenate and fuse the attention weights att_i^{ch} to obtain multi-scale channel attention weights Att :

$$Att = Concat(att_0^{ch}, att_1^{ch}, att_2^{ch}) \quad (12)$$

Finally, we multiply the multi-scale channel attention weights with the multi-scale feature maps to obtain the feature maps F^{ch} containing channel attention:

$$F^{ch} = Att \odot F \quad (13)$$

In the spatial attention component, the feature maps F_x and F_g are used as inputs, where F_x represents the output of the encoder and F_g is the upsampled output of the decoder. These feature maps undergo convolutional operations at three different scales, enabling multi-scale spatial position interaction and providing finer pixel-level attention for multi-level features. The corresponding feature maps F_x^{ms} and F_g^{ms} are obtained by adding the two feature maps to enhance the regions of interest. Subsequently, the ReLU function is introduced to increase network sparsity. Following this, convolutional layers compress the regions of interest into a single channel, and the sigmoid function quantifies these regions, producing pixel-level attention coefficients:

$$f_x^{add} = F_x^{ms} + F_g^{ms} \quad (14)$$

$$att_x^{sp} = Sigmoid(Conv(ReLU(f_x^{add}))) \quad (15)$$

Finally, the feature map F^{ch} , which includes channel attention, is element-wise multiplied with the spatial attention map to produce a new feature map that incorporates pixel-level attention weights:

$$F^{dmsa} = att_x^{sp} \odot F^{ch} \quad (16)$$

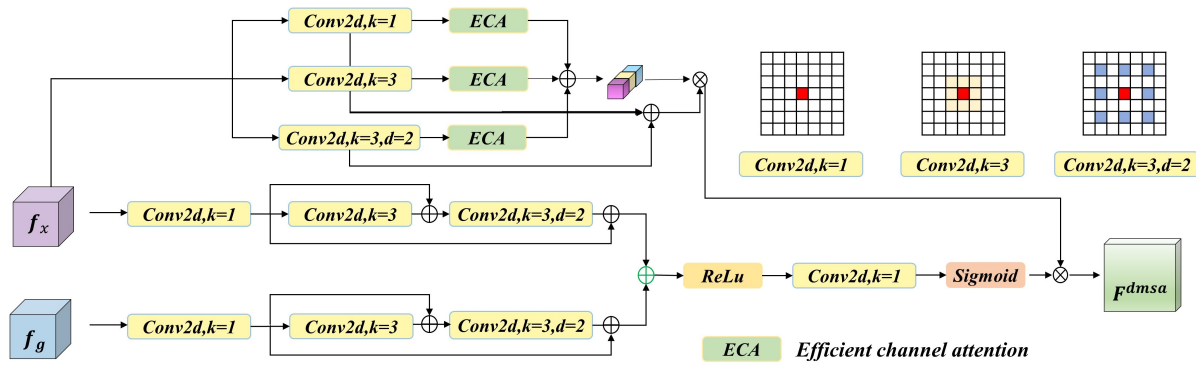


Figure 5. Dual Multi-scale Attention structure. DMSA performs feature alignment at three scales to alleviate semantic differences and achieve feature interaction and fusion.

3. Experiment

3.1. Datasets

To assess the building extraction performance of ME-FCN, three different datasets were used in this study: the WHU aerial building dataset [47], the Massachusetts building dataset [48], and the GF-2 building dataset.

The WHU aerial building dataset exhibits a complex distribution, including buildings with different materials, sizes, and shapes. The WHU aerial building dataset covers an

area of approximately 450 km² in Christchurch, New Zealand, and includes approximately 220,000 individual buildings. The ground resolution of the aerial images is 0.3 m. The images in the dataset were cropped into 512 × 512 pixels image patches. All cropped images were divided into 4736 training set images, 2416 test set images, and 1088 validation set images.

The Massachusetts building dataset was introduced by Mnih in 2013. It contains a total of 151 RGB remote sensing images of the Boston area, with each remote sensing image having a size of 1500 × 1500 pixels and a spatial resolution of 1 m. The labels consist of two categories: buildings and nonbuildings. The images were randomly cropped into 512 × 512 for training, testing, and prediction, resulting in 2430 training set images, 72 test set images, and 180 prediction set images.

The GF-2 building dataset was created by selecting high-resolution satellite images from the GF-2 satellite for the years 2017, 2019, and 2021, covering the Qingdao city area. The ground resolution of this self-made dataset is 0.8 m, and the original images were cropped into 512 × 512 pixels data for training, testing, and prediction. The dataset consists of 891 training set images, 297 test set images, and 371 prediction set images.

3.2. Implementation Details

ME-FCN was conducted in PyTorch using NVIDIA GeForce RTX 3060Ti. During the training phase, ME-FCN used the Adam optimizer, the initial learning rate was set to 0.001, and the epoch was set to 100. To avoid model overfitting, ME-FCN used the early stopping strategy. Specifically, if the accuracy of the validation set did not improve significantly within five epochs, we stopped training. Considering the 8GB GPU memory limitation, the batch size of ME-FCN was set to 4.

3.3. Evaluation Metrics

To evaluate the performance of ME-FCN, five metrics were used to evaluate the accuracy of building extraction: overall accuracy (OA), recall, Precision, F1 score (F1), and IoU. These metrics are defined as follows:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (17)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (18)$$

$$Precision = \frac{TP}{TP + FP} \quad (19)$$

$$Recall = \frac{TP}{TP + FN} \quad (20)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (21)$$

OA is the percentage of correctly predicted samples from the total samples. Precision is the probability of true positive samples among all samples predicted as positive. Recall represents the probability that a true positive sample is predicted as positive. The F1 score comprehensively considers the relationship between Precision and recall to simultaneously maximize and balance them. IoU measures the extent to which target features are captured, maximizing the overlap between predicted labels and ground truth annotations, and assessing the model's accuracy.

3.4. Experiment with the WHU Aerial Building Dataset

To demonstrate the effectiveness of SEUB, AMEB, and DMSA and explore the superiority of ME-FCN, we conducted extensive experiments. Our network was compared with a va-

riety of methods, including the CNN-based approaches FCN [49], U-Net [36], PSPNet [32], HRNet [50], MANet [51], UNetformer [31], and CMTFNet [29], which are based on the fusion of transformer and CNN, and the transformer-based method DC-Swin [52]. FCN uses the VGG-16 backbone network. HRNet uses HRNet-w18 backbone network. U-Net, PSPNet, MANet, UNetFormer, CMTFNet, and ME-FCN use the Resnet50 backbone network. DC-Swin uses the Swin-S backbone network.

Table 1 illustrates the comparative analysis of various semantic segmentation methods, with our proposed ME-FCN showing a noteworthy lead over its counterparts. U-Net procures feature maps at multiple scales through a downsampling process and amalgamates these into multi-scale feature maps via cascade fusion. This strategy is effective in preserving both local characteristics and global context, facilitating the accurate delineation of building structures, and securing a commendable second-place ranking. MANet distinguishes itself by leveraging multiple efficient attention modules to extract contextual dependencies, thereby bolstering the semantic comprehension of building footprints, and achieving an IoU score of 88.80%. The accuracy of CMTFNet is close to that of MANet. CMTFNet uses an integrated approach of CNNs and transformers for multi-scale feature extraction, culminating in an IoU score of 88.24%. In contrast, ME-FCN adopts a more powerful multi-scale strategy, which can fully extract the semantic information of the input feature map. In addition, in the feature fusion part, the scale difference of semantic information at different scales is taken into account, and DMSA is designed to align multi-scale features, while retaining key local information, so that our model has excellent anti-interference ability and extraction effect. On the WHU aerial image dataset, our method outperforms U-Net, PSPNet, HRNet, MANet, CMTFNet, and DC-Swin by 1.44%, 9%, 2.85%, 1.81%, 2.37%, and 4.29%, respectively. We also tested the foundation-model method SAM-RSIS [53] on the WHU aerial building dataset; however, it only achieved an IoU of 77.40%. This is because SAM-RSIS is designed for instance segmentation tasks and has disadvantages in pixel-level evaluation indicators. Given that the parameter number of the SAM backbone network is significantly higher than that of other comparative methods, we believe this comparison is not entirely fair. Consequently, we only evaluated SAM-RSIS on the WHU aerial building dataset. We used FPS to measure models' inference speed, indicating how many images can be processed per second. Generally, a higher FPS means that the model can make inferences faster. Compared with CMTFNet based on the CNN-Transformer method, ME-FCN has significantly better inference speed than CMTFNet. Although UNetFormer maintains a fast inference speed, the extraction accuracy is very low. DC-Swin uses a lightweight backbone network, and the model inference speed is better than ME-FCN, but the extraction accuracy is much lower than ME-FCN.

Table 1. Quantitative comparison of various methods on the WHU aerial building dataset.

Method	Precision (%)	Recall (%)	OA (%)	F1 (%)	IoU (%)	FPS
FCN	69.50	87.23	95.12	77.36	63.08	12.11
U-Net	93.32	92.34	98.21	92.83	89.17	14.21
PSPNet	90.53	89.23	98.08	89.88	81.61	18.71
HRNet	93.72	93.24	98.76	93.48	87.76	18.86
MANet	93.97	94.17	98.87	94.07	88.80	41.66
UNetFormer	89.75	87.42	97.84	88.57	79.48	89.13
CMTFNet	95.07	92.48	98.82	93.75	88.24	8.34
DC-Swin	93.62	91.71	98.61	92.66	86.32	17.54
SAM-RSIS	96.27	79.79	97.45	87.26	77.40	-
ME-FCN	95.55	94.61	99.06	95.07	90.61	14.27

Figure 6 presents a visual comparison of our proposed ME-FCN against other state-of-the-art methods, qualitatively underscoring ME-FCN's superior performance. Rows 3 and 4 distinctly showcase ME-FCN's prowess in capturing large buildings, where it markedly surpasses its counterparts. The depiction of large buildings is more holistic, and the clarity of the overall output is enhanced. This achievement is credited to our multi-scale extraction module, which extends the model's capacity to encompass larger receptive fields. Despite MANet and CMTFNet achieving multi-scale information extraction based on attention mechanism, the absence of robust supervision in their fused multi-scale information introduces uncertainty, leading to incomplete building extraction. Row 2 exemplifies ME-FCN's proficiency in extracting small buildings with precision, accurately delineating building edges. In contrast to U-Net's straightforward multi-scale fusion, ME-FCN's DMSA module substantially mitigates interference in shallow features. This refinement of shallow features yields more precise boundary and positional information, aligning building contours more closely with the actual ground truth. Rows 1 and 3 highlight ME-FCN's capacity to discern non-building objects within complex backgrounds. While methods such as FCN, MANet, PSPNet, HRNet, UNetformer, DC-Swin, and CMTFNet exhibit varying degrees of false positives, ME-FCN's comprehensive feature capture and subsequent optimization significantly bolster the network's recognition capabilities, effectively circumventing false detections.

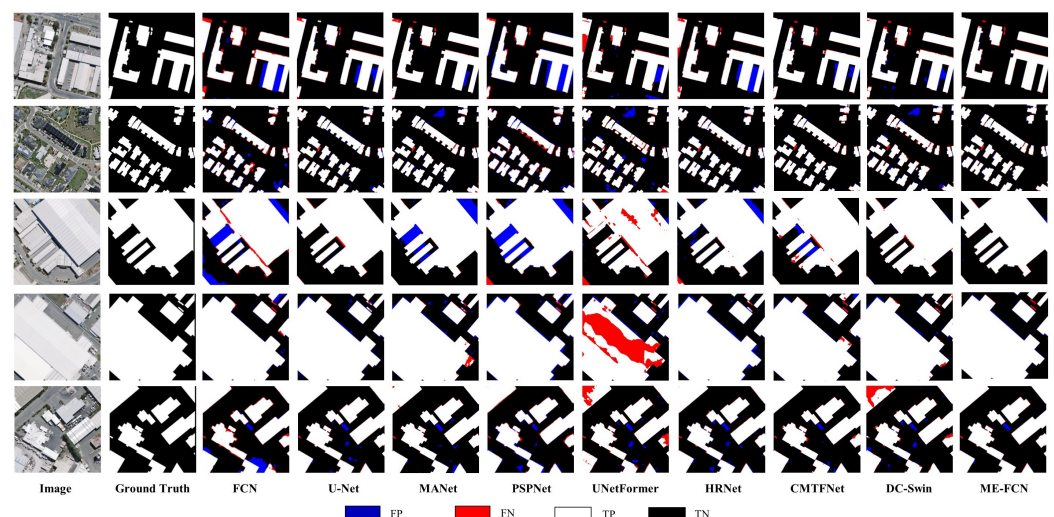


Figure 6. The set of predicted results obtained from different algorithms on the WHU aerial dataset (blue indicates the number of false negatives, red represents the number of false positives, white indicates the correct identification of positive samples, and black shows the correct identification of negative samples).

3.5. Experiment with the Massachusetts Building Dataset

The Massachusetts building dataset, characterized by lower resolution and predominantly small, densely distributed buildings, presents challenges in segmenting the shapes of these small structures. The dataset also features similar foreground and background textures and colors, along with shadow occlusions that hinder the completeness of large building extraction. Compared to other methods, ME-FCN excels in optimizing building extraction while preserving the original shapes to the greatest extent, demonstrating strong applicability to both large and small buildings.

Table 2 quantitatively demonstrates the outstanding performance of ME-FCN across various metrics. The significant improvement in IoU by 3.67% compared to the second-best method, even in complex environments with small buildings, highlights the method's excellent generalization capability.

Table 2. Quantitative comparison of various methods on the Massachusetts building dataset.

Method	Precision (%)	Recall (%)	OA (%)	F1 (%)	IoU (%)
FCN	67.30	51.75	86.19	58.51	41.35
U-Net	85.21	78.18	93.34	81.54	68.84
PSPNet	67.26	70.31	87.97	68.75	52.38
HRNet	79.70	76.91	91.97	78.28	64.31
MANet	84.74	77.73	93.17	81.08	68.18
UNetFormer	82.41	59.73	90.02	69.26	52.97
CMTFNet	83.50	78.17	92.98	80.75	67.71
DC-Swin	79.97	69.37	90.97	74.29	59.10
ME-FCN	87.09	81.25	94.20	84.07	72.51

Figure 7 showcases the remarkable extraction capability of ME-FCN in challenging environments affected by factors such as shadows. Firstly, when extracting small buildings from low-resolution images, as demonstrated in rows 1 and 3, ME-FCN effectively recognizes and accurately reconstructs the shapes of these structures. Secondly, in scenarios where large buildings suffer from incomplete extraction due to shadow occlusions, as illustrated in rows 2, 4, and 5, ME-FCN successfully restores the original shapes to a high degree. Compared to other methods, the buildings extracted by ME-FCN are more complete and exhibit clearer boundaries. While SEUB and AMEB utilize large kernel convolutions to capture features with extensive receptive fields, the incorporation of the DMSA module in ME-FCN better preserves locally relevant information and filters out redundant features, enabling the accurate restoration of buildings of varying sizes during the decoding process.

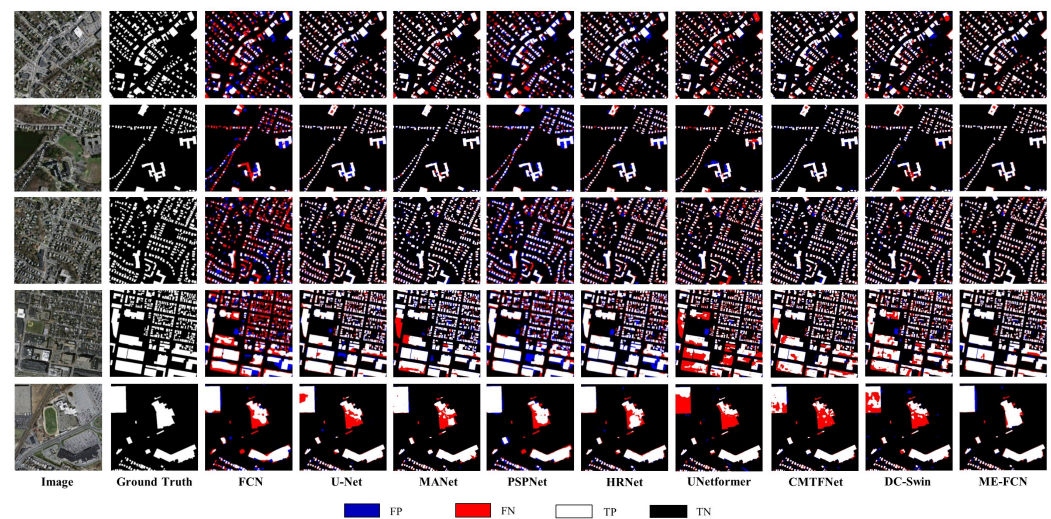


Figure 7. The set of predicted results obtained from different algorithms on the Massachusetts dataset (blue indicates the number of false negatives, red represents the number of false positives, white indicates the correct identification of positive samples, and black shows the correct identification of negative samples).

3.6. Experiment with the GF-2 Building Dataset

To validate the versatility of our method, we used GF-2 satellite imagery for building extraction. There are several differences compared to the two building datasets mentioned earlier. These differences include a greater variety of building roof colors and increased image noise.

As shown in Table 3, on the GF-2 building dataset, ME-FCN outperforms all other methods, further validating its ability to ensure the completeness of building extraction even in the presence of noise and various complex roof materials. CMTFNet achieves feature optimization by integrating CNN and attention, which increases the stability and noise resistance of the network. HRNet and U-Net struggle to adequately manage noise in shallow features, resulting in error-prone fused features that produce high recall but low IoU. In contrast, ME-FCN accounts for the impact of noise during feature fusion, with the SE attention mechanism in the SEUB module effectively suppressing noisy elements. Furthermore, our proposed DMSA module enhances the effectiveness of features at different scales by removing redundant information through both channel and spatial dimensions, thereby improving the network's resistance to noise. As a result, ME-FCN achieves an IoU of 88.78% on the GF-2 dataset, which is 0.68% higher than the second-best method. DC-Swin achieves an IoU of 85.67 with the transformer architecture, which is lower than the ME-FCN.

Table 3. Quantitative comparison of various methods on the GF-2 building dataset.

Method	Precision (%)	Recall (%)	OA (%)	F1 (%)	IoU (%)
FCN	69.06	61.94	88.31	65.31	48.49
U-Net	91.96	91.20	97.02	91.58	84.46
PSPNet	86.97	87.04	95.39	87.00	77.00
HRNet	89.40	89.32	96.23	89.36	80.77
MANet	92.36	89.40	96.81	90.86	83.25
UNetFormer	91.79	72.14	94.47	80.79	67.77
CMTFNet	94.00	93.36	97.76	93.68	88.10
DC-Swin	94.84	89.86	97.33	92.28	85.67
ME-FCN	94.65	93.47	97.90	94.06	88.78

To test the stability of the model, we conducted three experiments on the GF-2 building dataset. As shown in Table 4, ME-FCN maintains stable extraction accuracy in the three experiments, with an IoU of 88.78%, 89.01%, and 88.69% respectively, and a standard deviation of 0.134%.

Table 4. The quantitative results obtained from repeated experiments on the GF-2 building dataset.

	Precision (%)	Recall (%)	OA (%)	F1 (%)	IoU (%)
1	94.65	93.47	97.90	94.06	88.78
2	94.71	93.55	97.99	94.12	89.01
3	94.65	93.21	97.67	93.92	88.69

Figure 8 displays the extraction results of ME-FCN on the GF-2 building dataset. In scenarios where the edges of large buildings are affected by shadow occlusions, as shown in rows 1 and 4, ME-FCN effectively distinguishes shadows from building edges, preserving the regularity of building boundaries. The extracted building boundaries are nearly comparable to manual extractions. In situations where noise might cause false positives, such as in rows 1, 2, and 3, ME-FCN demonstrates minimal false positives overall. The SEUB and DMSA modules exhibit strong noise resistance, enabling our network to effectively disregard the impact of noise. Regarding the challenge of edge loss for small buildings, as illustrated in rows 1 and 2, the AMEB module facilitates the extraction of large receptive field features while preserving local features, ensuring that the edges of small buildings are well maintained.

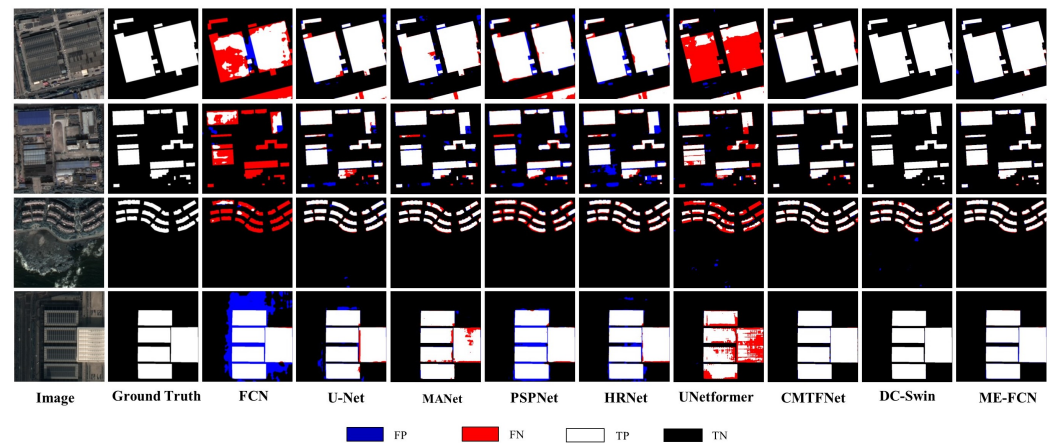


Figure 8. The set of predicted results obtained from different algorithms on the GF-2 building dataset (blue indicates the number of false positives, red represents the number of false negatives, white indicates the correct identification of positive samples, and black shows the correct identification of negative samples).

3.7. Ablation Experiment

We conducted ablation experiments on the WHU aerial building dataset to evaluate the contributions of different modules within the network.

We evaluated the effectiveness of the proposed module, with the quantitative results presented in Table 5. Using U-Net as the baseline, the incorporation of the SEUB module in the network model results in improvements in all metrics, notably enhancing IoU by 0.76%. This indicates that extracting single-scale semantic information from the input feature map resolution does not fully exploit the relevant semantic information. The SEUB module enables multi-scale feature extraction on the input resolution feature map. The embedded SE module within SEUB strengthens pertinent semantic information while suppressing redundant features and noise, ensuring the purity of the extracted semantic features. SEUB facilitates the extraction of multi-scale semantic information, providing richer multi-scale data for subsequent feature fusion, thereby reducing the information disparities between different decoding features. In the first row of Figure 9d, the SEUB module enables the complete extraction of the building. The SEUB module significantly enhances the model's multi-scale feature extraction capabilities, providing richer scale semantic information. In the second row, compared to the false detection observed in the baseline network, the SEUB module effectively considers multiple scale features, greatly improving the model's accuracy in recognition and extraction.

Furthermore, we evaluated the DMSA module, which led to further improvements in model accuracy, with IoU reaching 90.36%. DMSA is integrated into the skip connection part of the network, connecting the encoder and decoder. Given the potential semantic differences between the decoder features obtained by upsampling and the encoding features, DMSA first extracts features at three different scales in both spatial and channel dimensions, facilitating the interaction of features across scales. It then assigns weight parameters to the encoding features to further optimize them. DMSA mitigates semantic differences between encoding and decoding features at various scales. Compared to direct concatenation, DMSA more finely and accurately detects in the region of interest, reducing the risk of false or missed detections caused by semantic discrepancies across scales. As shown in Figure 9e, after adding the DMSA module, the false detection phenomenon of the model is significantly improved. DMSA enables the model to focus on the region of interest and ignore redundant semantic features.

Lastly, we evaluated the effectiveness of the AMEB module. Given the importance of receptive fields in understanding building size, we designed cascaded dilated convolutions to extract semantic information from large receptive fields. As shown in Table 5, the addition of the AMEB module further improves the accuracy of the model, with the IoU reaching 90.

66%. This shows that AMEB improves the ability of the model to understand and extract features from large buildings. However, overemphasizing large receptive fields can result in the loss of detailed information. To address this, we incorporated an auxiliary branch that improves the network to learn and determine when large receptive field information is necessary. AMEB achieves a balance between obtaining large receptive fields and preserving local information, which is crucial for the model's generalization capabilities.

Table 5. Ablation experiments of different modules on the WHU aerial building dataset.

SEUB	DMSA	AMEB	Precision (%)	Recall (%)	F1 (%)	IoU (%)
			93.32	92.34	92.83	89.17
✓			93.57	94.72	94.14	89.93
✓	✓		95.31	94.56	94.94	90.36
✓	✓	✓	95.54	94.60	95.07	90.66

We conducted a comparative analysis of mainstream attention modules by replacing the DMSA in the skip connection with the Efficient Channel Attention (ECA) and Attention Gates (AGs) modules. Table 6 presents the quantitative results of this comparison. The accuracies of ECA and AGs are similar, but DMSA significantly outperforms both, with an IoU increase of 0.44% and 0.45%, respectively. Although ECA achieves adaptive feature optimization from the channel dimension, it does not account for the relationships between features at different scales. While AGs address semantic differences across scales, their feature optimization through two branches does not fully enable semantic interaction between features, limiting model accuracy. DMSA, on the other hand, aligns encoding and decoding features through three-scale feature extraction, ensuring accurate feature interaction. It further optimizes features across two dimensions, leading to more precise building information extraction.

Table 6. Quantitative comparison experiment of different attention modules.

Attention Module	Precision (%)	Recall (%)	F1 (%)	IoU (%)
ECA	95.37	94.36	94.86	90.22
AGs	94.96	94.75	94.85	90.21
DMSA	95.54	94.60	95.07	90.66

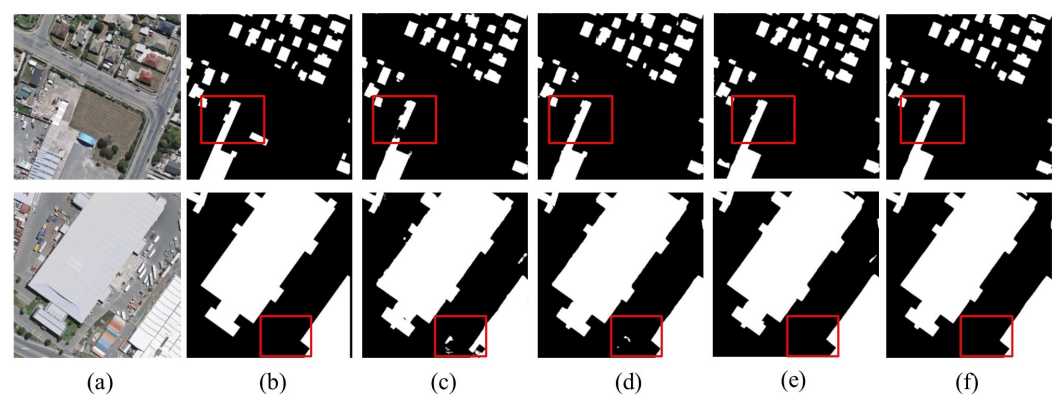


Figure 9. Ablation experiment results with different module combinations. (a) Image, (b) ground truth, (c) baseline, (d) Baseline + SEUB (e) baseline + SEUB + DMSA, and (f) baseline + SEUB + DMSA + AMEB.

4. Conclusions

This paper proposes a powerful building extraction network, ME-FCN. ME-FCN retains the overall architecture of the encoder–decoder. In the feature extraction part, two novel multi-scale feature extraction modules were designed: Squeeze-and-Excitation U-

Block (SEUB) and Adaptive Multi-scale feature Enhancement Block (AMEB). SEUB uses a simple U-shaped structure to achieve multi-scale feature extraction, which greatly improves the upper limit of the network's information acquisition. It has an embedded SE attention module to screen and filter noise and redundant information. AMEB focuses on deep multi-scale feature extraction. Considering the information loss caused by downsampling, AMEB uses cascaded atrous convolutions to achieve larger receptive field feature acquisition to ensure complete extraction of building contours. Additionally, AMEB introduces a novel branch that enables the network to autonomously assign weights to different receptive field features, thereby reducing randomness in information utilization and enhancing adaptive field selection. ME-FCN introduces a Dual Multi-scale Attention (DMSA) module in skip connections. DMSA realizes multi-scale feature interaction and fusion, improves the accuracy of multi-scale building feature representation, and greatly reduces the misclassification of similar features. In three building datasets, the extraction accuracy of ME-FCN is better than other mainstream semantic segmentation methods, which shows the excellent generalizability of ME-FCN. In addition, we verified the effectiveness of the SEUB and AMEB components proposed in this paper in multi-scale feature extraction through ablation experiments, and DMSA was also better than the mainstream attention module.

Author Contributions: Conceptualization, H.S. and W.Z.; methodology, Y.Z. and S.W.; software, Y.Z.; validation, Y.Z.; formal analysis, Y.Z. and H.S.; investigation, H.S. and Y.M.; data curation, M.X. and Y.M.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z. and S.W.; visualization, M.X. and W.Z.; supervision, S.W. and H.S.; project administration, Y.M. and W.Z.; funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Postdoctoral Fellowship Program CPSF under Grant Number (No. GZB20240849), the Natural Science Foundation of Shandong Provincial under Grant Number (No. ZR2024QD133), the Fundamental Research Funds for the Central Universities under Grant Number (No. 24CX06022A) and the National Natural Science Foundation of China under Grant Number (No. 41776182).

Data Availability Statement: Data are contained within the article.

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

References

1. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [\[CrossRef\]](#)
2. Liao, W.; Van Coillie, F.; Gao, L.; Li, L.; Zhang, B.; Chanussot, J. Deep learning for fusion of APEX hyperspectral and full-waveform LiDAR remote sensing data for tree species mapping. *IEEE Access* **2018**, *6*, 68716–68729. [\[CrossRef\]](#)
3. Lechner, A.M.; Foody, G.M.; Boyd, D.S. Applications in remote sensing to forest ecology and management. *One Earth* **2020**, *2*, 405–412. [\[CrossRef\]](#)
4. Alganci, U.; Soydas, M.; Sertel, E. Comparative research on deep learning approaches for airplane detection from very high-resolution satellite images. *Remote Sens.* **2020**, *12*, 458. [\[CrossRef\]](#)
5. 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. [\[CrossRef\]](#)
6. 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. [\[CrossRef\]](#)
7. Huang, H.; Chen, Y.; Wang, R. A lightweight network for building extraction from remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5614812. [\[CrossRef\]](#)
8. 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. [\[CrossRef\]](#)
9. Jabari, S.; Zhang, Y.; Suliman, A. Stereo-based building detection in very high resolution satellite imagery using IHS color system. In Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 2301–2304.
10. 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. [\[CrossRef\]](#)

11. Pesaresi, M.; Gerhardinger, A.; Kayitakire, F. A robust built-up area presence index by anisotropic rotation-invariant textural measure. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2008**, *1*, 180–192. [\[CrossRef\]](#)
12. Yasir, M.; Liu, S.; Pirasteh, S.; Xu, M.; Sheng, H.; Wan, J.; de Figueiredo, F.A.; Aguilar, F.J.; Li, J. YOLOShipTracker: Tracking ships in SAR images using lightweight YOLOv8. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *134*, 104137. [\[CrossRef\]](#)
13. Abdollahi, A.; Pradhan, B. Integrating semantic edges and segmentation information for building extraction from aerial images using UNet. *Mach. Learn. Appl.* **2021**, *6*, 100194. [\[CrossRef\]](#)
14. Shrestha, S.; Vanneschi, L. Improved fully convolutional network with conditional random fields for building extraction. *Remote Sens.* **2018**, *10*, 1135. [\[CrossRef\]](#)
15. Zhao, K.; Kang, J.; Jung, J.; Sohn, G. Building extraction from satellite images using mask R-CNN with building boundary regularization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 247–251.
16. Zhang, Z.; Wang, Y. JointNet: A common neural network for road and building extraction. *Remote Sens.* **2019**, *11*, 696. [\[CrossRef\]](#)
17. 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. [\[CrossRef\]](#)
18. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, Online, 13–18 July 2020; pp. 1597–1607.
19. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
20. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 4015–4026.
21. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. Dinov2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193.
22. Liu, Y.; Zhao, Z.; Zhang, S.; Huang, L. Multiregion Scale-Aware Network for Building Extraction from High-Resolution Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5626310. [\[CrossRef\]](#)
23. Kang, W.; Xiang, Y.; Wang, F.; You, H. EU-Net: An efficient fully convolutional network for building extraction from optical remote sensing images. *Remote Sens.* **2019**, *11*, 2813. [\[CrossRef\]](#)
24. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [\[CrossRef\]](#)
25. 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. [\[CrossRef\]](#)
26. Wang, T.; Borji, A.; Zhang, L.; Zhang, P.; Lu, H. A stagewise refinement model for detecting salient objects in images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4019–4028.
27. Vaswani, A. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
28. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 205–218.
29. Wu, H.; Huang, P.; Zhang, M.; Tang, W.; Yu, X. CMTFNet: CNN and multiscale transformer fusion network for remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 2004612. [\[CrossRef\]](#)
30. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
31. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [\[CrossRef\]](#)
32. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
33. Chen, L.C. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
34. 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. [\[CrossRef\]](#)
35. 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. [\[CrossRef\]](#)
36. 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; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
37. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
38. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

39. Xu, S.; Deng, B.; Meng, Y.; Liu, G.; Han, J. ReA-Net: A Multiscale Region Attention Network with Neighborhood Consistency Supervision for Building Extraction from Remote Sensing Image. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 9033–9047. [\[CrossRef\]](#)
40. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
41. Park, J. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
42. Li, R.; Zheng, S.; Duan, C.; Yang, Y.; Wang, X. Classification of hyperspectral image based on double-branch dual-attention mechanism network. *Remote Sens.* **2020**, *12*, 582. [\[CrossRef\]](#)
43. Xu, X.; Zhou, Y.; Lu, X.; Chen, Z. FERA-Net: A Building Change Detection Method for High-Resolution Remote Sensing Imagery Based on Residual Attention and High-Frequency Features. *Remote Sens.* **2023**, *15*, 395. [\[CrossRef\]](#)
44. Yu, M.; Chen, X.; Zhang, W.; Liu, Y. AGs-Unet: Building Extraction Model for High Resolution Remote Sensing Images Based on Attention Gates U Network. *Sensors* **2022**, *22*, 2932. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [\[CrossRef\]](#)
46. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
47. 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. [\[CrossRef\]](#)
48. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto (Canada): Toronto, ON, Canada, 2013.
49. 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.
50. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [\[CrossRef\]](#)
51. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5607713. [\[CrossRef\]](#)
52. Wang, L.; Li, R.; Duan, C.; Zhang, C.; Meng, X.; Fang, S. A novel transformer based semantic segmentation scheme for fine-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6506105. [\[CrossRef\]](#)
53. Luo, M.; Zhang, T.; Wei, S.; Ji, S. SAM-RSIS: Progressively adapting SAM with box prompting to remote sensing image instance segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4413814. [\[CrossRef\]](#)

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.