electronics

Article

Multi-Scale Frequency-Spatial Domain Attention Fusion Network for Building Extraction in Remote Sensing Images

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

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.); 332207050689@zzuli.edu.cn (H.G.)

* Correspondence: 332305040480@zzuli.edu.cn

Abstract: Building extraction from remote sensing images holds significant importance in the fields of land resource management, urban planning, and disaster assessment. Encoder-decoder deep learning models are increasingly favored due to their advanced feature representation capabilities in image analysis. However, because of the diversity of architectural styles and issues such as tree occlusion, traditional methods often result in building omissions and blurred boundaries when extracting building footprints. Given these limitations, this paper proposes a cutting-edge Multi-Scale Frequency-Spatial Domain Attention Fusion Network (MFSANet), which consists of two principal modules, named Frequency-Spatial Domain Attention Fusion Module (FSAFM) and Attention-Guided Multi-scale Fusion Upsampling Module (AGMUM). FSAFM introduces frequency domain attention and spatial attention separately to enhance the feature maps, thereby strengthening the model's boundary-detection capabilities and ultimately improving the accuracy of building extraction. AGMUM first resizes and concatenates attention enhancement maps to enhance contextual understanding and applies attention guidance to further improve prediction accuracy. Our model demonstrates superior performance compared to existing semantic segmentation methods on both the WHU building data set and the Inria aerial image data set.

Keywords: remote sensing; building extraction; dual-domain learning; multi-scale fusion

Citation: Liu, J.; Chen, H.; Li, Z.; Gu, H. Multi-Scale Frequency-Spatial Domain Attention Fusion Network for Building Extraction in Remote Sensing Images. *Electronics* **2024**, *13*, 4642. https://doi.org/10.3390/electronics13234642

Academic Editor: Fernando De la Prieta Pintado

Received: 24 October 2024
Revised: 22 November 2024
Accepted: 23 November 2024
Published: 25 November 2024

1. Introduction

Advances in deep learning are fueling rapid growth in building extraction from remote sensing images. Analyzing remote sensing images has enabled the application of techniques such as pixel-level object classification and building extraction across diverse fields, including land resource management [1], urban planning [2,3], and disaster assessment [4]. In comparison with lower- and medium-resolution counterparts, high-resolution remote sensing images provide richer details of the targets, but they also increase the computational volume and complexity of the image-processing process, posing various challenges to the building-extraction task. Furthermore, the varying shapes and sizes of buildings, along with different lighting conditions, shadows, and occlusions in the surrounding environment, complicate the segmentation process, leading to problems such as blurred boundaries, missed targets, and incomplete extraction areas in the final predictions [5].

In recent years, various semantic segmentation models have emerged. Most deep learning-based building-extraction methods focus on spatial domain information processing. Despite their proficiency in overall segmentation accuracy, they still face challenges in handling regions with gray-level variations, such as edges and shadows [6–8]. The frequency domain features are more sensitive to information in these regions. Therefore, incorporating frequency domain information into building-extraction networks is necessary. The utilization of frequency domain information in digital image processing [9–12] has garnered increasing attention. Ref. [13] introduced a new neural network input method based on frequency domain learning. By selectively retaining essential frequency components, this method reduces the input data size and enhances the accuracy of instance segmentation tasks. The discrete cosine transform (DCT) efficiently transforms image data into the frequency domain, providing superior energy compaction compared to the discrete Fourier transform (DFT). This enables the DCT to gather more crucial image information while discarding relatively unimportant frequency domain regions, as illustrated in Figure 1. Accordingly, the DCT represents an optimal choice for image compression [14,15]. However, solely depending on frequency domain data could lead to a forfeiture of spatial information, since spatial domain characteristics include a variety of semantic details across different styles and categories. Thus, utilizing the inherent frequency-spatial domain characteristics of remote sensing images for accurate large-scale building analysis is still a formidable challenge [16].

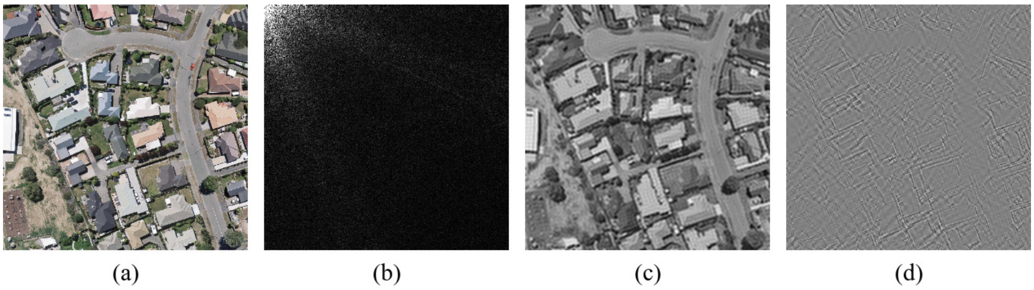

Figure 1. Illustration of frequency domain features: (**a**) original image, (**b**) 2D-DCT transformed spectrograms, (**c**) low-frequency components, and (**d**) high-frequency components.

While existing building-extraction methods have achieved high accuracy rates, they still face significant challenges. Downsampling enables the extraction of features at varying levels of an image, but it simultaneously reduces resolution, which can lead to information loss, particularly for small or irregularly shaped buildings. The loss of relevant feature information leads to problems such as lost predictions and blurred boundaries [17,18].

To address these challenges, we propose a deep learning network that innovatively integrates frequency domain components and utilizes multi-scale feature fusion to achieve high accuracy in building extraction. Key contributions are outlined below:

1. Frequency-Spatial Domain Attention Fusion Module (FSAFM): We integrate a frequency domain component into the attention module, which enhances the detection of features of interest within the feature maps. In this module, by integrating frequency and spatial feature enhancements, we enable the model to concentrate its attention on the most critical features, leading to a more sensitive perception of boundary information.
2. Attention-Guided Multi-scale Fusion Upsampling Module (AGMUM): The module is divided into two parts: (1) Multi-Scale Features Fusion (MSF): This part integrates attention feature maps from various depths, improving the model's comprehension of contextual information and facilitating the capture of boundary details. (2) Attention Upsampling (AU): Through frequency-space fusion attention, this part effectively compensates for the losses incurred during upsampling.
3. Multi-Scale Frequency-Spatial Domain Attention Fusion Network (MFSANet): Based on the above structure, we propose the MFSANet architecture, which integrates rich frequency domain information while preserving spatial domain details. This network is designed in two distinct phases: the initial phase employs dual-domain attention mechanisms to refine feature extraction, directing the network's focus towards salient features; the subsequent phase utilizes multi-scale feature concatenation along with attention-guided upsampling, enabling the network to perceive multi-scale information effectively.

The structure of this paper is organized as follows: Section 2 provides a review of the current landscape in building segmentation of remote sensing imagery, focusing on feature enhancement methods in the frequency domain. Section 3 delivers a comprehensive overview of the entire architecture of the building-extraction network, covering the two key components, FSAFM and AGMUM. Section 4 evaluates the model's efficacy through experimental outcomes using two data sets of remote sensing imagery. Section 5 presents a synthesis of the research outcomes and further explores future research directions.

2. Related Works

This section provides an overview of methods and recent advances in semantic segmentation for remotely sensed images, with a particular focus on techniques related to the frequency domain.

2.1. CNN-Based Semantic Segmentation

As the potential of deep learning is continuously explored, CNN-based algorithms are increasingly replacing traditional building segmentation methods. Conventional building-extraction algorithms, including support vector machines [19], conditional random fields (CRF) [20], and random forests [21], typically require additional manual intervention and exhibit a more limited scope of applicability. However, with the advent of CNNs, the field of building extraction has reached a new height. For the first time, FCNs have achieved pixel-level classification through end-to-end convolution operations, generating segmentation maps with the same resolution as the input images, allowing for automatic building extraction. Subsequent building-extraction methods are largely based on FCNs or their variants. The most classic semantic segmentation network, U-Net [22], is a symmetrical architecture featuring an encoder-decoder structure. By utilizing efficient skip connections, more contextual information is preserved in image-segmentation tasks, significantly improving the segmentation accuracy. However, during the encoding process, while more contextual information is captured, the spatial resolution of feature maps gradually decreases, potentially leading to the loss of certain minute details. During the decoding process, restoring low-resolution feature maps to their original resolution may cause blurriness and an inability to accurately reconstruct the boundaries.

To address the issue of information loss during downsampling, Chen et al. [23] proposed Res2-UNet, which decomposes the feature map into different sub-feature maps by channels, organized using a residual network structure. This method enhances multi-scale learning capabilities. Ali et al. [24] proposed EPD, which uses the class uncertainty within local windows to generate soft labels. This method effectively preserves edge details and significantly improves the network's ability to produce high-quality predictions at low resolutions. With the emergence of attention mechanisms, their application in remote sensing building extraction is continually being explored. Attention mechanisms are resource allocation strategies that allow models to focus on the most important aspects of input data, improving accuracy and efficiency in processing and prediction. Since Hu et al. [25] proposed SENet, which introduced channel attention mechanisms to enhance feature representation in convolutional neural networks, various attention mechanisms have emerged. Zhou and Wei [26] proposed FANet, which combines the Pyramid Vision Transformer to capture global features and further optimizes these features through the Feature Aggregation Module and Difference Elimination Module to form a unified representation.

CNN-based methods typically operate solely within the spatial domain, overlooking the spectral characteristics of the image. This underutilization of information can lead to distortion when extracting contextual information.

2.2. Transformer-Based Semantic Segmentation

Transformer-based network architectures have also been extensively studied and applied in remote sensing image analysis. Through the self-attention mechanism, Transformer networks can capture global contextual information, thereby effectively improving the accuracy of semantic segmentation in complex backgrounds of remote sensing images. Transformer-based methods can be divided into two categories: the first category is the fully Transformer encoder-decoder. Xie et al. [27] proposed SegFormer, a semantic segmentation model that unifies Transformers with MLP decoders for efficient and accurate image analysis, surpassing previous methods in performance and efficiency. Cao et al. [28] introduced Swin-Unet, a pioneering pure Transformer model for medical image segmentation, which incorporates the Swin Transformer in both the encoder and decoder. It has achieved significantly higher accuracy than traditional convolutional networks in tasks such as multi-organ and cardiac segmentation. The second category combines the advantages of CNNs and Transformers. Zhang et al. [29] proposed the Swin Transformer, a model that combines the advantages of Transformers and CNNs. It effectively captures global contextual information of images and utilizes local feature details to enhance segmentation accuracy, particularly when dealing with complex objects and rich details in images. Wang et al. [30] introduced UNetFormer, a semantic segmentation model for remote sensing imagery, which combines a lightweight CNN encoder with a Transformer decoder. It outperforms other models in accuracy and efficiency on urban scene data sets.

Unlike traditional Convolutional Neural Networks (CNNs), Transformer networks enhance feature representation by modeling long-range dependencies, overcoming the limitations of CNNs in handling large-scale contextual information.

2.3. Learning in Frequency Domain

It is crucial to learn frequency domain information from images because frequency domain analysis allows models to understand image content from frequency levels. By transforming to the frequency domain, image details and edges (high-frequency information) as well as smooth areas and overall structure (low-frequency information) can be distinctly identified and processed separately. Dong et al. [31] proposed a headless lightweight semantic segmentation architecture called AFFormer, which introduces a lightweight adaptive frequency filter, significantly reducing parameters and computational complexity while preserving high precision. Huang et al. [32] proposed FSDR, which learns a model with strong generalization ability by randomizing specific frequency components of images in the frequency space. Qin et al. [33] proposed FcaNet, which compresses channels based on DCT and utilizes frequency analysis to address the compression challenges in traditional channel attention caused by large amounts of information. Zhu et al. [34] proposed MDNet, which employs a dual-dimension Discrete Cosine Transform attention module (D3AM) and a multi-scale Discrete Cosine Transform pyramid (MDP) to effectively harness frequency domain information, thereby enhancing the feature representation for change detection tasks. Fan et al. [35] proposed MIFNet, an approach that significantly enhances the model's resilience to diverse interferences by leveraging the complementary advantages of CNNs and Transformers, while also incorporating frequency domain information. Zhang et al. [36] proposed a Dual-Domain Transformer method, achieving comprehensive feature extraction by combining the Fast Fourier Transform and boundary-aware modules.

Frequency domain compressed representations encompass vital patterns essential for tasks related to image comprehension [13]. So far, frequency domain learning methods have not seen widespread application. In the field of building extraction, incorporating spectral analysis in models can help identify subtle differences in the spectral features of various buildings, particularly in data where features are similar but spectral characteristics differ, such as boundary areas obscured by shadows.

In conclusion, CNN-based methods are primarily concerned with the exploitation of spatial domain data, which ultimately results in the incomplete utilization of the inherent information present within the image. Furthermore, this approach may introduce blurring in building-extraction tasks, leading to inaccurate boundary predictions. Therefore, incorporating frequency domain information to achieve frequency-spatial domain fusion can enhance the model's sensitivity to boundary details and improve its learning of spatial domain information, ultimately leading to more accurate results in remote sensing applications.

3. Method

In the following, the general structure of the Multi-Scale Frequency-Space Domain Attention Fusion Network proposed in this paper is first introduced, followed by a detailed discussion of the two key modules of the MFSANet: FSAFM and AGMUM.

3.1. Multi-Scale Frequency-Spatial Domain Attention Fusion Network (MFSANet)

Figure 2 illustrates the comprehensive architecture of MFSANet, which is specifically engineered for high-resolution remote sensing imagery building-extraction tasks. Recent research indicates that most models are based on an architecture that includes both an encoder and a decoder. During the encoding phase, as downsampling takes place, the resolution of feature maps gradually decreases. This results in the loss of certain boundary details and blurriness of other information, which subsequently impacts the precision recovery during upsampling. Additionally, inadequate design of the feature fusion strategy during upsampling may impede the effective integration of contextual information, diminishing the model's predictive power. To address these issues, we use FSAFM and AGMUM to perform layer-by-layer enhancement and scale fusion of different layers of feature information. The FSAFM design enhances the model's capacity to concentrate on critical details across multiple scales, thereby improving its feature representation and learning capacity. The AGMUM design combines high-level abstract information with low-level detailed features, significantly enhancing the network's ability to capture contextual information.

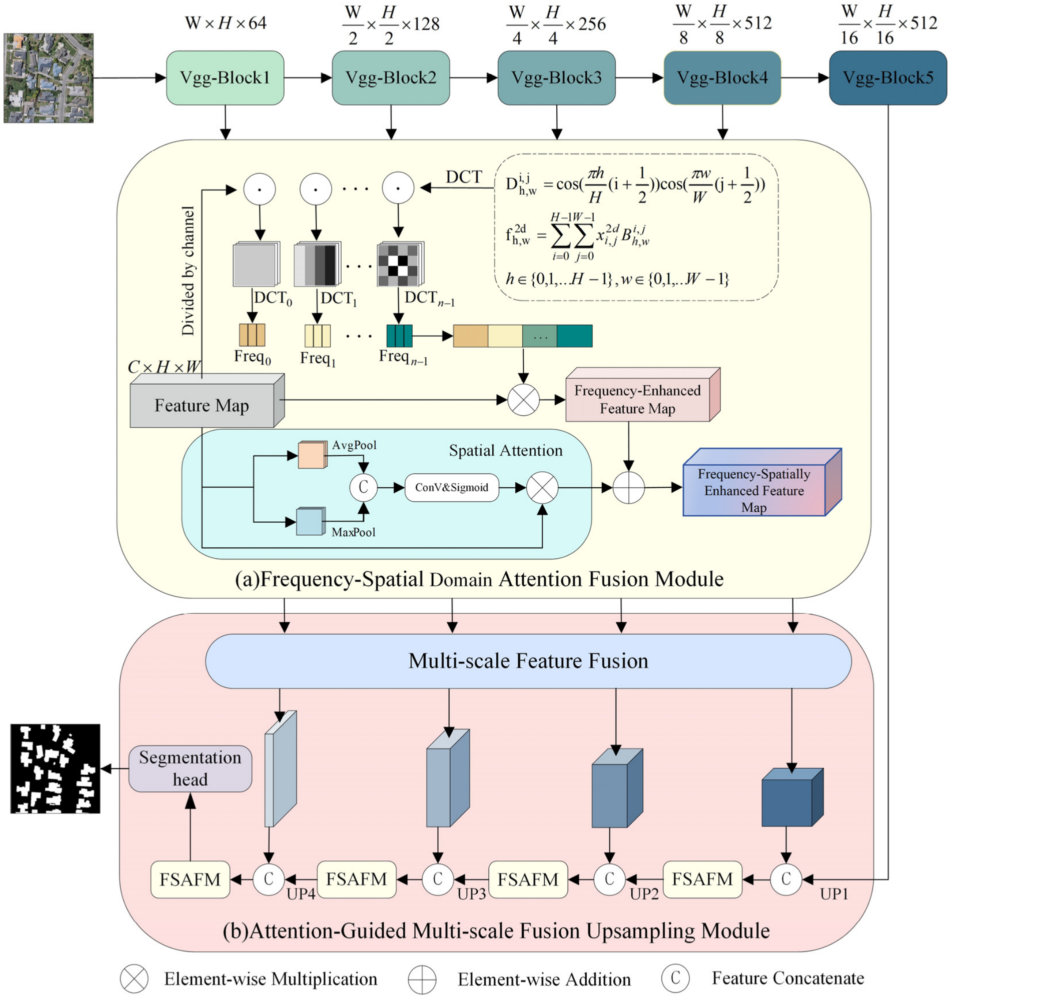
W × H × 64
W/2 × H/2 × 128
W/4 × H/4 × 256
W/8 × H/8 × 512
W/16 × H/16 × 512
Vgg-Block1
Vgg-Block2
Vgg-Block3
Vgg-Block4
Vgg-Block5
DCT
$D_{h,w}^{i,j} = \cos(\frac{\pi h}{H}(i+\frac{1}{2}))\cos(\frac{\pi w}{W}(j+\frac{1}{2}))$
$f_{h,w}^{2d} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d} B_{h,w}^{i,j}$
$h \in \{0,1,..H-1\}, w \in \{0,1,..W-1\}$
Divided by channel
DCT_0
DCT_1
DCT_{n-1}
$Freq_0$
$Freq_1$
$Freq_{n-1}$
C × H × W
Feature Map
Frequency-Enhanced Feature Map
AvgPool
Spatial Attention
C
ConV&Sigmoid
MaxPool
Frequency-Spatially Enhanced Feature Map
(a)Frequency-Spatial Domain Attention Fusion Module
Multi-scale Feature Fusion
Segmentation head
FSAFM
C
UP4
FSAFM
C
UP3
FSAFM
C
UP2
FSAFM
C
UP1
(b)Attention-Guided Multi-scale Fusion Upsampling Module
Element-wise Multiplication
Element-wise Addition
Feature Concatenate

Figure 2. The network architecture of MFSANet consists of four components: a backbone, (**a**) FSAFM, (**b**) AGMUM, and a Segmentation head.

Algorithm 1 offers a comprehensive overview of the algorithmic structure of our proposed model. It includes four key components: feature extraction, frequency-space domain enhancement, multi-scale fusion, and attention-guided upsampling. The deep structure of VGG16 enables the network to learn diverse levels of intricate feature information, which is vital for downstream tasks such as image segmentation. Furthermore, the pre-trained weights of VGG16 have been optimized on a multitude of standard data sets, thereby facilitating transfer learning. MFSANet is comprised of four principal components:

Algorithm 1: Segmentation Model Based on Dual-Domain Enhancement

```
Input: $512 \times 512$ remote sensing images
Output: Building segmentation map
1: Use VGG16 to extract the feature maps of five layers of the input image
2: Apply FSAFM for frequency-space domain attention enhancement on the first four
layers of feature maps
3: Apply MSF to fuse the enhanced multi-scale features
4: Apply attention-guided upsampling
    a. Skip connections
    b. Apply FSAFM enhancement again to the concatenated features
5: Utilize the segmentation head to produce the final segmentation map
6: Output the building segmentation map
```

Figure 2 illustrates that MFSANet first extracts five feature maps at different levels from the input remote sensing image. The first four feature maps are enhanced by FSAFM, followed by multi-scale attention information integration through MSF. The fused feature maps maintain consistency in both channel dimensions and spatial size with the original feature maps at each level, while exhibiting higher attention and clarity for boundary information. The deepest feature map is first upsampled and then concatenated with the processed features from the fourth layer. The concatenated features are subsequently guided by FSAFM to compensate for the upsampling loss and enhance feature representation. This step is repeated for the other three layers. This process effectively enhances and integrates information from different levels to achieve high-performance building segmentation.

3.2. Frequency-Spatial Domain Attention Fusion Module (FSAFM)

Incorporating attention mechanisms into building extraction aims to boost the model's targeted emphasis on critical feature zones, refine the richness of feature representation, mitigate overfitting, and ensure higher accuracy in image-segmentation tasks. In addition, the mechanism improves the adaptability of the model to complex remote sensing environments. Traditional building-extraction methods have primarily focused on the spatial domain, underutilizing the potential of frequency domain information. This indicates that frequency domain enhancement holds significant promise for research and application.

To fully utilize the inherent information of images, this paper introduces a Frequency-Spatial Attention Fusion Module (FSAFM), as shown in Figure 2a. This module is applied to the skip connection part. The FSAFM is designed to maximize the utilization of image data and emphasize the overall features of buildings. It enhances the model's capacity to identify object boundaries and improve the clarity of the regions of interest in the input feature maps, ultimately achieving enhanced segmentation performance. The overall algorithm flow of FSAFM is shown in Algorithm 2.

Algorithm 2: Frequency-Spatial Domain Attention Fusion Module

```
Input: Multi-level feature maps $X \in \mathbb{R}^{C \times H \times W}$
Output: Fusion attention enhancement feature map Z
For the feature maps derived from various tiers of the backbone network:
 Frequency-domain attention component
   Segment into n parts across the channel dimension
   Assign 2D-DCT frequency components to each part
   Output 2D DCT enhanced feature map $Y_1$
 Spatial-domain attention component
   Refine the feature map through the application of average and max pooling
   Concatenate the two feature maps
   Integrate the spatial weight map with the original feature map, resulting in a spatial attention-enhanced feature map $Y_2$
   Add the frequency-domain enhanced feature map and the spatial attention map pixel-wise
end for
Output the feature map refined by frequency-space domain attention Z
```

3.2.1. Spatial Domain Attention Enhancement

Spatial domain attention models enhance the model's sensitivity to spatial information by focusing on key areas within the image. Essentially, these models transform spatial information from the original image into another space through a spatial-transformation module, preserving critical information, generating weighted masks for each location, and outputting the results with weighting, thereby enhancing the target area and weakening irrelevant background areas.

Initially, apply average pooling and max pooling to the input feature map. Next, combine the outcomes of these two processes. Apply the Sigmoid function to generate attention maps. Finally, perform a pixel-wise multiplication between these attention maps and the original input feature map.

$$AvgPool = \frac{1}{h \times w} \sum_{p=i}^{i+h-1} \sum_{q=j}^{j+w-1} x_{p,q} \tag{1}$$

$$MaxPool = \max\left(\left\{x_{p,q} \middle| p \in [i, i+h-1], q \in [j, j+w-1]\right\}\right) \tag{2}$$

where $x_{p,q}$ represents the value at position (p,q) in the input feature map, $h \times w$ is the size of the pooling window, and i and j are the positions in the pooled output feature map.

$$X = Conv\left(Cat\left(AvgPool\left(x_{p,q}\right), MaxPool\left(x_{p,q}\right)\right)\right) \tag{3}$$

$$Y_2 = Sigmoid(X) \otimes X \tag{4}$$

where X denotes the feature map with one channel resulting from $Conv$ applied to the concatenation of $AvgPool$ and $MaxPool$ outputs of the input feature map $x_{(p,q)}$. The kernel shape of the $Conv$ is (in = 2, out = 1, k = 7, p = 3). Y_2 represents the feature map enhanced by spatial domain attention, and $\bigotimes$ represents element-wise multiplication.

3.2.2. Frequency Domain Attention Enhancement

The calculation of the 2D DCT is represented by the following formula:

$$F_{h,w}^{2d} = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j}^{2d} D_{h,w}^{i,j} \tag{5}$$

$$D_{h,w}^{i,j} = \cos\left(\frac{(2i+1)\pi h}{2H}\right) \cos\left(\frac{(2j+1)\pi w}{2W}\right) \tag{6}$$

where $D_{h,w}^{i,j}$ is the basis function of the 2D DCT, $F_{h,w}^{2d} \in \mathbb{R}^{H \times W}$ is the frequency spectrum representation, $x^{2d} \in \mathbb{R}^{H \times W}$ is the input feature map, and H and W correspond to its height and width, respectively.

In the frequency domain enhancement section, the input feature map is initially split into n segments along the channel dimension, denoted as $[X_0, X_1, \cdots X_{n-1}]$, $X_i \in \mathbb{R}^{C' \times H \times W}$, $i \in \{0, 1, \cdots, n-1\}$, $C' = \frac{C}{n}$. A 2D DCT is then applied to each segment to obtain the frequency components for each part as shown in Equation (7). Then, all parts are concatenated using Equation (8) and the compressed overall information is output as frequency domain attention enhancement.

$$Freq_i = DCT(X_i) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X_{h,w}^{i} D_{h,w} \tag{7}$$

$$Freq = cat\left(\left[Freq_0, Freq_1, \cdots Freq_{n-1}\right]\right) \tag{8}$$

where $Freq_i \in \mathbb{R}^{C'}$ represents the frequency components of the i-th segment X_i after splitting along the channel dimension, with a dimension of C'. The calculation for the final output of the frequency domain attention enhancement is as follows:

$$Y_1 = Sigmoid(fc(Freq)) \otimes X \tag{9}$$

where fc denotes a sequence of fully connected layers used for both dimensionality reduction and expansion, while maintaining the same size, followed by activation and output of attention weights, X represents the input feature map, and Y_1 represents the feature map after frequency domain attention enhancement.

Fuse the feature map enhanced by frequency domain attention with the feature map enhanced by spatial domain attention using element-wise addition.

$$Z = Y_1 \oplus Y_2 \tag{10}$$

where Z represents the feature map enhanced by frequency-spatial domain fusion attention, and $\oplus$ represents element-wise addition. This design, by leveraging the characteristics of different domains, can comprehensively perceive the details of the image's features and better allocate attention weights.

3.3. Attention-Guided Multi-Scale Fusion Upsampling Module (AGMUM)

As downsampling continues, it may lead to the degradation of key information in the original data. Furthermore, downsampling is likely to disrupt the continuity of data through time and space, causing discontinuities in time series and geographic distributions, thus affecting the accuracy and reliability of subsequent analysis. Relying solely on feature maps from the downsampling process for reconstructing the predicted image can lead to significant errors. We introduce an Attention-Guided Multi-Scale Fusion Upsampling Module (AGMUM) to tackle this problem. The framework of this module is depicted in Figure 2b, with the Multi-Scale Feature Fusion (MSF) part shown in Figure 3a. In the multi-scale fusion part, improvements are made to the skip connections, effectively integrates contextual information by combining shallow edge features with deep detail features. This enables precise recovery of image details during the upsampling and reconstruction process, thereby improving segmentation accuracy.

$$F_1 = Q_1 \tag{11}$$

$$F_2 = Cat(RS(Q_1), Q_2) \tag{12}$$

$$F_3 = Cat(RS(Q_1, Q_2), Q_3) \tag{13}$$

$$F_4 = Cat(RS(Q_1, Q_2, Q_3), Q_4) \tag{14}$$

where F_i represents the feature maps at each level output after fusion and $Q_i, i \in \{1, 2, 3, 4\}$ represents the feature maps at each level of downsampling. RS denotes the adjustment of the input feature map's scale to match the dimensions of a specific tier. Cat denotes concatenation along the channel dimension.

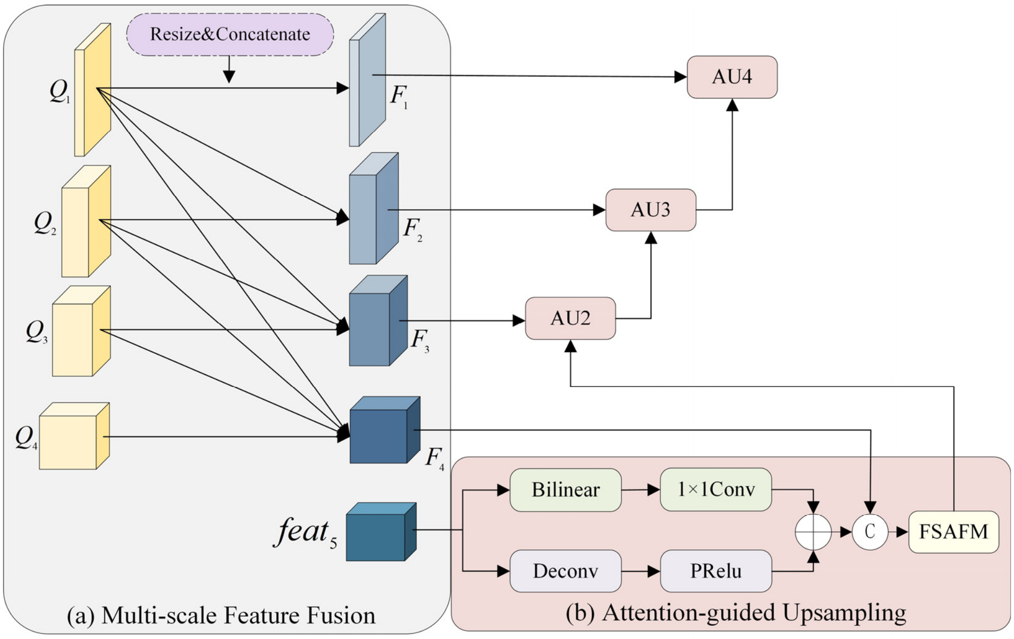

Figure 3. The AGMUM framework diagram consists of two parts: (**a**) Multi-Scale Feature Fusion and (**b**) Attention-guided Upsampling. "Bilinear" refers to bilinear interpolation upsampling, and "Deconv" denotes transposed convolution upsampling.

In the feature-reconstruction phase of upsampling, we leverage the complementary strengths of bilinear interpolation upsampling and transposed convolution upsampling, as depicted in Figure 3b. Additionally, an attention mechanism is introduced during the upsampling process to achieve high-accuracy feature reconstruction. Bilinear interpolation upsampling effectively preserves edge details of the image. Transposed convolution upsampling not only effectively restores image details and textures but also enhances the model's capacity to capture details. It uses multi-scale convolution kernels to facilitate the fusion of features across various levels, thereby improving image quality. After transposed convolution, the PReLU activation layer introduces non-linearity, which improves the training efficiency and generalization ability of the model. This facilitates a quicker training phase and elevates the clarity of the resulting imagery. The fusion of the two upsampling methods enhances the model's adaptability and robustness to different types of images. After performing upsampling of deep features and skip connections with shallow features, attention is introduced, namely the FSAFM module, which enhances the feature maps by frequency and spatial domain attention.

$$up_i = FSA(Conv(UpSample(x, Bilinear)) \oplus PRelu(ConvTranspose2d(x))) \tag{15}$$

where up_i, represents the output feature map after attention-guided upsampling. FSA represents frequency-spatial domain attention fusion. $Bilinear$ represents bilinear interpolation upsampling, and $ConvTranspose2d$ represents transposed convolution upsampling.

The incorporation of multi-scale feature fusion, coupled with attention-guided upsampling, significantly enhances the model's ability to detect a comprehensive range of both high- and low-frequency information. This approach effectively addresses common issues such as detail omission and boundary blurring that occur during the upsampling process, ultimately achieving greater accuracy in reconstructing the predicted feature maps.

4. Experiments

4.1. Data Sets and Hardware Environment

To evaluate the efficacy and usability of our method, tests were carried out on two accessible data sets in the public domain. These data sets contain aerial images of buildings captured in different urban environments. The data sets are described in detail below.

1. WHU building data set [37]. The data set is composed of aerial and satellite imagery. For our experiments, we utilize the aerial image data set to validate the model's effectiveness. It contains 187,000 buildings across more than 450 square kilometers, totaling 8189 images, each sized at 512 × 512 pixels. In the experiment, for the training, there are 4736 images (60% of the total data set); for validation, the count is 1036 images (15%); and for testing, the number stands at 2416 images (25%).
2. Inria aerial image data set [38]. The data set consists of five cities, containing a total of 360 remote sensing images, each measuring 5000 × 5000 pixels. From this data set, we selected 180 images for experimentation, cropping the originals into patches of 512 × 512 pixels. The cropped images were subsequently allocated into training (12,600 images, 70%), validation (2700 images, 15%), and test (2700 images, 15%) sets.

Experiments were run on an NVIDIA GeForce RTX 4090 24 GB GPU, utilizing Python 3.8, Pytorch 1.11.0, CUDA 11.3, along with libraries like NumPy, Pillow, and OpenCV.

4.2. Evaluation Metrics

We employed standard metrics for remote sensing image segmentation, including F1 score (F1), Intersection over Union (IoU), Precision, and Recall. TP represents the pixels accurately identified as positive, while FP denotes those incorrectly identified as positive, and FN signifies the pixels mistakenly identified as negative. The evaluation metrics are defined as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{16}$$

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{19}$$

4.3. Experiment Analysis

4.3.1. Quantitative Comparison Results

We performed performance validation of MFSANet on two public data sets. Table 1 presents the overall quantitative evaluation results of different methods on the WHU building data set.

Table 1. Numerical comparison of different methods on the WHU data set.

Method	IoU	Recall	Precision	F1
UNet [22]	88.28	93.81	94.74	93.77
SegNet [39]	85.35	91.31	92.65	91.97
HRNet [40]	86.51	92.67	93.66	93.16
UNetFormer [30]	87.33	92.84	93.64	93.24
LCS [41]	90.71	94.86	95.38	95.12
BuildFormer [42]	90.73	95.14	95.15	95.14
MRANet [43]	90.59	**95.22**	94.90	95.06
Ours	**91.01**	95.12	**95.47**	**95.29**

Bold font indicates the best performance for each attribute.

Our proposed model shows significant improvement over other methods, with an Intersection over Union (IoU) of 91.01%, Recall of 95.12%, and Accuracy of 95.47%. These improvements are all attributed to FSAFM and AGMUM. Compared to the classic benchmark networks UNet, SegNet, and HRNet, our proposed network improves F1 scores by 1.52%, 3.29%, and 2.13%, respectively. When compared to the high-performance BuildFormer, the metrics show improvements of 0.28%, −0.02%, 0.32%, and 0.15%, respectively. These data indicate that our proposed MFSANet has outstanding feature-extraction and prediction capabilities.

For additional evaluation of the efficacy and usability of our model, we conducted experiments on the Inria aerial image data set, and the findings are detailed in Table 2, where our proposed MFSANet demonstrates superior performance compared to other models across all metrics.

Table 2. Numerical Comparison of Methods on the Inria data set.

Method	IoU	Recall	Precision	F1
UNet	74.40	84.28	86.39	85.32
SegNet	72.00	82.33	84.69	83.49
HRNet	75.03	84.92	86.56	85.73
LCS	78.82	86.77	89.58	88.15
BuildFormer	81.24	88.78	90.65	89.71
MRANet	81.79	**90.72**	89.26	89.99
Ours	**82.45**	90.09	**90.68**	**90.38**

Bold font indicates the best performance for each attribute.

Experiments on both data sets demonstrate that our proposed MFSANet not only substantially outperforms traditional segmentation networks but also shows enhancements compared to the high-performing method BuildFormer. This success can be attributed to the two key modules we introduced, FSAFM and AGMUM. By effectively leveraging the inherent frequency and spatial domain information of images and integrating multi-level features, MFSANet achieves its outstanding performance.

4.3.2. Qualitative Results

We performed visual experiments across two data sets to demonstrate the efficacy of our approach, with the associated visualizations depicted in Figures 4 and 5. In these figures, black pixel regions indicate the background, while white pixel regions denote buildings. Areas where the background is mistakenly identified as buildings are highlighted in red, and buildings incorrectly categorized as background appear in green.

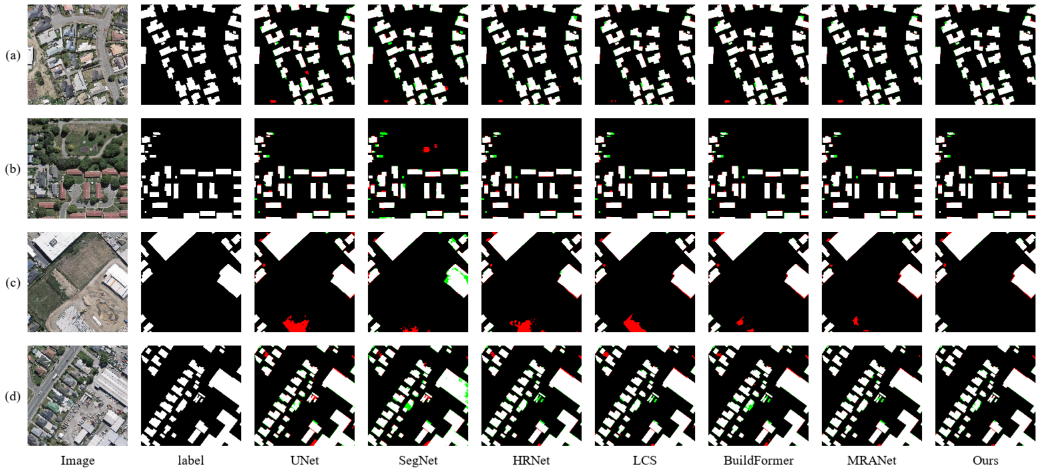

Figure 4. The comparative qualitative findings from the WHU building data set provide a clearer visualization of the four images (**a**–**d**) using different pixel colorings.

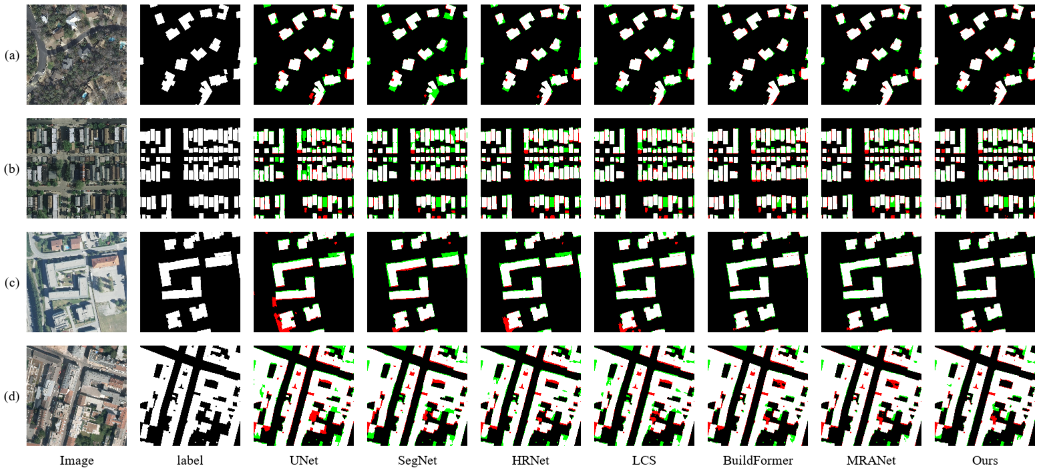

Figure 5. The qualitative comparison results of the Inria data set illustrate the visual outcomes of four images (**a**–**d**).

In Figure 4, partial visualization results of the WHU building data set are displayed. We selected two remote sensing images depicting dense urban structures, one image featuring large buildings, and another showcasing dense vegetation to comprehensively validate the model's performance. In Figure 4a, where buildings have irregular shapes, our model produces smoother predictions along the boundaries and accurately identifies regions with pixel features akin to the buildings. Likewise, in Figure 4c, there is an area at the bottom of the image whose features closely resemble those of buildings, other models encountered certain errors during testing, while MFSANet made better decisions during prediction. Figure 4b includes dense tree cover, with some buildings were obscured by the trees. Compared to other networks, MFSANet demonstrates superior overall performance. In Figure 4d, the buildings on both sides exhibit different styles, other networks fail to clearly delineate the boundaries, whereas MFSANet predicts smoother boundary information. Additionally, in this figure, there is a small "H"-shaped building, which UNet predicts as a single structure along with the surrounding small buildings, while the other three models exhibit prediction omission issues. In contrast, the visualization results of building extraction by MFSANet indicate that this model is capable of better integrating contextual information, enabling more accurate segmentation of buildings from the background. The extracted building outlines are smoother, effectively optimizing the issues of false positives and missed detections.

Figure 5 presents the visual outcomes on select test images of the Inria data set. The remote sensing images in this data set are sourced from five different regions, which exhibit certain variations in building styles and lighting conditions, presenting challenges for the segmentation task. In Figure 5a, the vegetation is dense, and the buildings are largely obscured by trees, leading to significant missed detections by other networks during testing. In contrast, MFSANet is able to reduce false detections. In Figure 5b, the buildings are densely arranged, the lighting is weaker compared to other test images, and trees obstruct the view between the buildings, causing networks like UNet and LCS to encounter missed detections during segmentation. Additionally, some small shadow areas are misclassified as buildings by other networks, while MFSANet can more accurately segment building boundaries, avoiding over-segmentation of shadow areas. Figure 5c also demonstrates the issue of UNet and LCS predicting shadow areas as buildings. In Figure 5d, the buildings are closely connected, and for the background areas surrounding the buildings, other models exhibit misclassification issues. MFSANet shows improved segmentation performance. The effectiveness and robustness of MFSANet, particularly in handling irregular boundaries, high-density small buildings, and severe shadow scenarios, are validated through visual analysis on the more complex Inria data set.

Although MFSANet was designed specifically for building extraction in remote sensing images, its architecture and methodology provide a potential foundation for adaptation to other remote sensing tasks. The ability of MFSANet to effectively integrate multi-scale features and attention mechanisms suggests that it can handle the diverse challenges posed by other remote sensing tasks, which often require capturing both global context and fine-grained details. With appropriate adjustments and optimizations, we believe that MFSANet could be extended to tasks such as land cover classification and object detection.

4.4. Ablation Study

4.4.1. Quantitative Comparison Results

Ablation analysis was performed on the WHU building data set to measure the individual impact of the two modules on the model's performance. We used a classic semantic segmentation network as the baseline, and sequentially validated the effectiveness of the FSAFM and the AGMUM. The outcomes of the experiment are displayed within Table 3.

Table 3. Results of ablation experiments.

Method	IoU	Recall	Precision	F1
Baseline	88.28	93.81	94.74	93.77
Baseline + FSAFM	89.64	94.31	94.76	94.53
Baseline + AGMUM	90.51	94.66	95.38	95.01
Baseline + FSAFM + AGMUM	91.01	95.12	95.47	95.29
SegNet	82.83	87.59	93.90	90.60
SegNet + FSAFM	83.67	89.09	93.22	91.10
SegNet + AGMUM	84.08	89.73	93.03	91.35

This paper introduces a frequency-spatial domain fusion attention method in the encoder part, which helps the model to concentrate on relevant areas. As shown in the table above, after introducing FSAFM, the four metrics improved by 1.36%, 0.50%, 0.02%, and 0.76%, respectively. In the decoder part, AGMUM is introduced, which enables the model to better learn contextual information through multi-scale fusion in the first half. Then, in the upsampling phase, it combines the advantages of two upsampling methods, and, guided by attention, significantly enhances prediction accuracy. After introducing AGMUM, the four metrics improved by 2.23%, 0.85%, 0.64%, and 1.24%, respectively. It is noticeable that both modules independently contribute to the model's enhanced performance. To further validate the performance of the two modules, we incorporated them into the SegNet architecture. The results, shown in Table 3, demonstrate that both FSAFM and AGMUM continue to improve the model's performance.

Table 4 shows that although the proposed model increases both computational cost and parameter count overall, the increase in parameters due to FSAFM is relatively small, demonstrating the efficiency of 2D-DCT channel compression in reducing the computational load. While the introduction of FSAFM and AGMUM leads to an increase in runtime and memory usage, this increase is reasonable considering the significant performance improvements.

Table 4. Computational efficiency and resource usage.

Method	FLOP (G)	Parameters (M)	Runtime (Min)	Memory (MiB)
Baseline	225.66	24.89	4:34	10,165
Baseline + FSAFM	305.42	28.00	6:48	14,033
Baseline + AGMUM	364.79	33.72	7:50	14,725
Baseline + FSAFM + AGMUM	364.85	33.76	8:01	15,267
SegNet	160.83	29.44	4:30	13,227
SegNet + FSAFM	160.89	29.52	5:18	21,499
SegNet + AGMUM	175.67	35.21	5:26	13,281

The input size is set to 512 × 512 × 3, and the runtime time is the duration required for one iteration over the WHU training data set.

4.4.2. Qualitative Results

We present a comparative visualization of four models to intuitively demonstrate the performance of the two modules: Baseline, Baseline + FSAFM, Baseline + AGMUM, and Baseline + FSAFM + AGMUM.

From Figure 6, it can be observed that introducing FSAFM and AGMUM in the Baseline markedly refines the model's perception of building information.

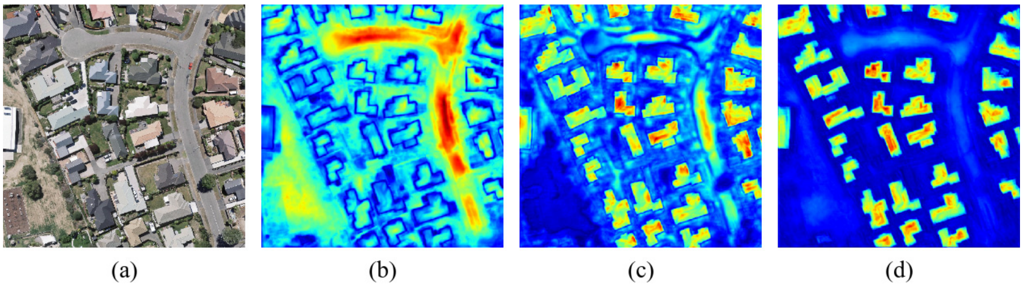

Figure 6. Heatmaps of each module from ablation experiments. (**a**) Original remote sensing image, (**b**) Heatmap of Baseline, (**c**) Heatmap of Baseline + FSAFM, (**d**) Heatmap of Baseline + AGMUM.

As shown in Figure 7, dual-domain attention effectively lowers the model's false positive rate by increasing the weight of building pixels. Specifically, in Figure 7b, the building is surrounded by trees, and FSAFM concentrates on the features of the building, reallocating weights through attention fusion to help the model more clearly distinguish between the building and the background. Similarly, in Figure 7c, several containers are arranged together, and FSAFM, by weighting important features, enables the model to focus more on recognizing buildings, especially demonstrating excellent performance in distinguishing between easily confused objects and buildings.

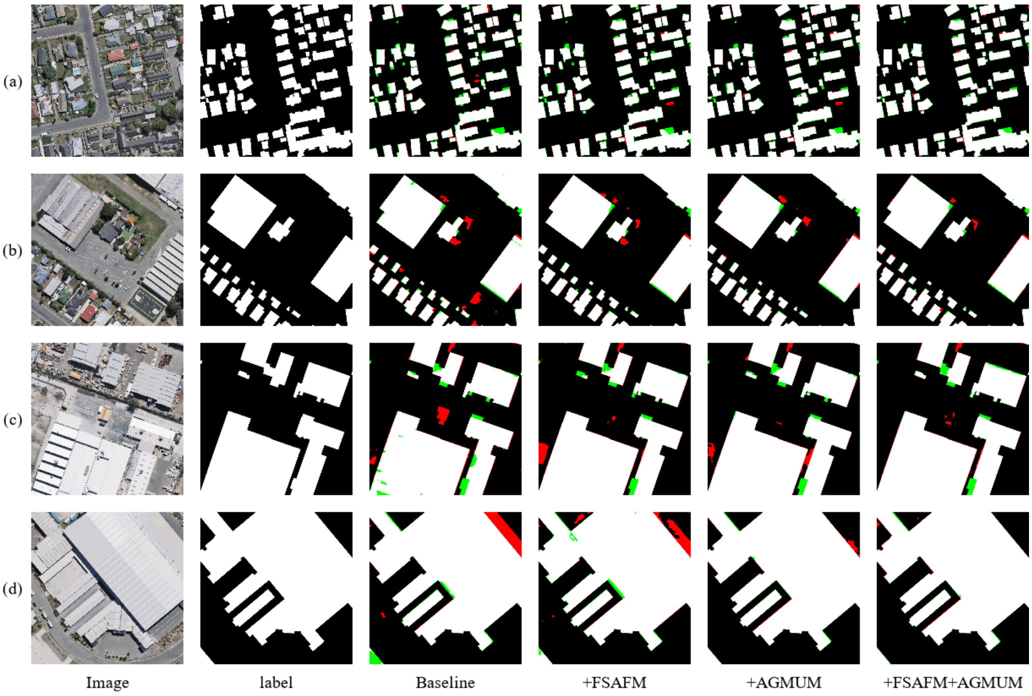

Figure 7. Ablation experiment comparison results for the WHU building data set illustrate the visualization results of four remote sensing images (**a**–**d**).

AGMUM integrates feature information from different levels, allowing the model to simultaneously capture high-level semantics and low-level details, thus facilitating smoother boundary processing when predicting building outlines. Additionally, through further guidance from attention, it improves the model's capacity to identify small targets. As illustrated in Figure 7a, AGMUM comprehensively focuses on buildings of various sizes, effectively alleviating the issue of missed detections by the model, ensuring the accuracy of the predicted number of buildings.

Through the introduction of FSAFM and AGMUM, the feature processing and fusion strategies are optimized, significantly reducing the occurrence of missed detections, ensuring the integrity of building boundaries. Thanks to the multi-scale features fusion, the network's predictions of building boundaries are smoother, with finer details. These improvements enhance the model's efficacy for building-extraction tasks and increase its robustness in intricate environments.

5. Conclusions

This paper employs MFSANet to address the challenge of achieving high precision in building extraction from high-resolution remote sensing imagery. The network utilizes a symmetric encoder-decoder framework to initially capture multi-tiered image features, followed by a reconstruction to the original pixel resolution. In the encoder part, a dual-domain attention method is introduced for feature enhancement. The decoder phase incorporates a multi-scale feature fusion, and the upsampling process is guided by attention. This network fully considers the frequency and spatial feature information in the images, by fusing the attention from the frequency and spatial domains and redistributing the attention weights. Additionally, the model successfully combines multi-scale feature information. This design helps the model better understand the consistency of pixel label assignments within local areas while maintaining the integrity of larger buildings and reducing the missed detection of small objects. To verify the effectiveness of MFSANet, quantitative experiments, qualitative experiments, and ablation studies were conducted on the public data sets WHU building data set and Inria aerial image data set. Its strong performance affirms the method's effectiveness and robustness in carrying out building extraction.

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

Funding: This research was funded by Henan Province Key R&D Project (241111210400), Henan Provincial Science and Technology Research Project (242102211007 and 242102211020), and Science and Technology Innovation Project of Zhengzhou University of Light Industry (23XNKJTD0205).

Data Availability Statement: The WHU Building Aerial Imagery and Inria aerial image data set used in the experiment can be downloaded at http://gpcv.whu.edu.cn/data/building_dataset.html (accessed on 22 November 2024) and https://project.inria.fr/aerialimagelabeling/ (accessed on 22 November 2024), respectively.

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

References

1. Huang, X.; Cao, Y.X.; Li, J.Y. An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images. *Remote Sens. Environ.* **2020**, *244*, 111802. [CrossRef]
2. Chen, Z.; Wei, Y.; Shi, K.; Zhao, Z.; Wang, C.; Wu, B.; Qiu, B.; Yu, B. The potential of nighttime light remote sensing data to evaluate the development of digital economy: A case study of China at the city level. *Comput. Environ. Urban Syst.* **2022**, *92*, 101749. [CrossRef]
3. Bai, H.; Li, Z.W.; Guo, H.L.; Chen, H.P.; Luo, P.P. Urban Green Space Planning Based on Remote Sensing and Geographic Information Systems. *Remote Sens.* **2022**, *14*, 4213. [CrossRef]
4. Sakellariou, S.; Sfougaris, A.I.; Christopoulou, O.; Tampekis, S. Integrated wildfire risk assessment of natural and anthropogenic ecosystems based on simulation modeling and remotely sensed data fusion. *Int. J. Disaster Risk Reduct.* **2022**, *78*, 103129. [CrossRef]

5. Jiang, X.; Zhang, X.; Xin, Q.; Xi, X.; Zhang, P. Arbitrary-Shaped Building Boundary-Aware Detection with Pixel Aggregation Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 2699–2710. [CrossRef]
6. 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]
7. 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]
8. Shao, P.; Shi, W.; Liu, Z.; Dong, T. Unsupervised Change Detection Using Fuzzy Topology-Based Majority Voting. *Remote Sens.* **2021**, *13*, 3171. [CrossRef]
9. You, S.; Liu, Y.; Lei, B.; Wang, S. Fine Perceptive GANs for Brain MR Image Super-Resolution in Wavelet Domain. *arXiv* **2020**. [CrossRef]
10. Chen, H.; Yokoya, N.; Chini, M. Fourier domain structural relationship analysis for unsupervised multimodal change detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 99–114. [CrossRef]
11. Yu, B.; Yang, A.; Chen, F.; Wang, N.; Wang, L. SNNFD, spiking neural segmentation network in frequency domain using high spatial resolution images for building extraction. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102930. [CrossRef]
12. Sun, H.; Luo, Z.; Ren, D.; Du, B.; Chang, L.; Wan, J. Unsupervised multi-branch network with high-frequency enhancement for image dehazing. *Pattern Recognit.* **2024**, *156*, 110763. [CrossRef]
13. Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.-k.; Ren, F. Learning in the Frequency Domain. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1737–1746.
14. Gupta, A.; Mahobiya, C. Analysis of Image Compression Algorithm Using DCT. *Int. J. Sci. Technol. Eng.* **2016**, *3*, 121–127.
15. Chen, Z.; Liu, T.; Xu, X.; Leng, J.; Chen, Z. DCTC: Fast and Accurate Contour-Based Instance Segmentation with DCT Encoding for High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 8697–8709. [CrossRef]
16. Zhang, C.; Lam, K.-M.; Wang, Q. CoF-Net: A Progressive Coarse-to-Fine Framework for Object Detection in Remote-Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5600617. [CrossRef]
17. Zheng, J.; Shao, A.; Yan, Y.; Wu, J.; Zhang, M. Remote Sensing Semantic Segmentation via Boundary Supervision-Aided Multiscale Channelwise Cross Attention Network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4405814. [CrossRef]
18. 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]
19. 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]
20. Li, E.; Femiani, J.C.; 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]
21. 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]
22. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
23. 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]
24. Ali, S.; Lee, Y.R.; Park, S.Y.; Tak, W.Y.; Jung, S.K. Towards Efficient and Accurate CT Segmentation via Edge-Preserving Probabilistic Downsampling. *arXiv* **2024**, arXiv:2404.03991.
25. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
26. Zhou, X.; Wei, X. Feature Aggregation Network for Building Extraction from High-resolution Remote Sensing Images. In Proceedings of the Pacific Rim International Conference on Artificial Intelligence, Jakarta, Indonesia, 15–19 November 2023; pp. 105–116.
27. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Álvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Neural Information Processing Systems, Online, 6–14 December 2021.
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 ECCV Workshops, Montreal, ON, Canada, 11 October 2021.
29. Zhang, C.; Jiang, W.; Zhang, Y.; Wang, W.; Zhao, Q.; Wang, C. Transformer and CNN Hybrid Deep Neural Network for Semantic Segmentation of Very-high-resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4408820. [CrossRef]
30. 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.* **2021**, *190*, 196–214. [CrossRef]
31. Dong, B.; Wang, P.; Wang, F. Head-Free Lightweight Semantic Segmentation with Linear Transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 516–524. [CrossRef]
32. Huang, J.; Guan, D.; Xiao, A.; Lu, S. FSDR: Frequency Space Domain Randomization for Domain Generalization. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6887–6898.
33. Qin, Z.; Zhang, P.; Wu, F.; Li, X. FcaNet: Frequency Channel Attention Networks. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2020; pp. 763–772.

34. Zhu, Y.; Fan, L.; Li, Q.; Chang, J. Multi-Scale Discrete Cosine Transform Network for Building Change Detection in Very-High-Resolution Remote Sensing Images. *Remote Sens.* **2023**, *15*, 5243. [CrossRef]
35. Fan, J.; Li, J.; Liu, Y.; Zhang, F. Frequency-aware robust multidimensional information fusion framework for remote sensing image segmentation. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107638. [CrossRef]
36. Zhang, J.; Shao, M.; Wan, Y.; Meng, L.; Cao, X.; Wang, S. Boundary-Aware Spatial and Frequency Dual-Domain Transformer for Remote Sensing Urban Images Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5637718. [CrossRef]
37. Ji, S.P.; Wei, S.Q.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction From an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 574–586. [CrossRef]
38. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
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]
40. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-Resolution Representations for Labeling Pixels and Regions. *arXiv* **2019**, arXiv:1904.04514.
41. Liu, Z.Y.; Shi, Q.; Ou, J.P. LCS: A Collaborative Optimization Framework of Vector Extraction and Semantic Segmentation for Building Extraction. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5632615. [CrossRef]
42. Wang, L.B.; Fang, S.H.; Meng, X.L.; Li, R. Building Extraction With Vision Transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5625711. [CrossRef]
43. Jiang, W.X.; Chen, Y.; Wang, X.F.; Kang, M.L.; Wang, M.Y.; Zhang, X.J.; Xu, L.X.; Zhang, C. Multi-branch reverse attention semantic segmentation network for building extraction. *Egypt. J. Remote Sens. Space Sci.* **2024**, *27*, 10–17. [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.