remote sensing

MDPI

Article

Multilevel Feature Aggregated Network with Instance Contrastive Learning Constraint for Building Extraction

Shiming Li [1,2,3,*], Tingrui Bao [1], Hui Liu [1,2,3], Rongxin Deng [1] and Hui Zhang [1]

[1] College of Surveying and Geo-informatics, North China University of Water Resources and Electric Power, Zhengzhou 450046, China; dengrongxin@ncwu.edu.cn (R.D.); zhanghui2019@ncwu.edu.cn (H.Z.)
[2] Collaborative Innovation Center of Geo-Information Technology for Smart Central Plains, Zhengzhou 450045, China
[3] Key Laboratory of Spatiotemporal Perception and Intelligent Processing, Ministry of Natural Resources, Zhengzhou 450045, China
* Correspondence: lishiming@ncwu.edu.cn; Tel.: +86-166-3808-0102

Abstract: Building footprint extraction from remotely sensed imagery is a critical task in the field of illegal building discovery, urban dynamic monitoring, and disaster emergency response. Recent research has made significant progress in this area by utilizing deep learning techniques. However, it remains difficult to efficiently balance the spatial detail and rich semantic features. In particular, the extracted building edge is often inaccurate, especially in areas where the buildings are densely distributed, and the boundary of adjacent building instances is difficult to distinguish accurately. Additionally, identifying buildings with varying scales remains a challenging problem. To address the above problems, we designed a novel framework that aggregated multilevel contextual information extracted from multiple encoders. Furthermore, we introduced an instance constraint into contrastive learning to enhance the robustness of the feature representation. Experimental results demonstrated that our proposed method achieved 91.07% and 74.58% on the intersection over union metric on the WHU and Massachusetts datasets, respectively, outperforming the most recent related methods. Notably, our method significantly improved the accuracy of building boundaries, especially at the building instance level, and the integrity of multi-scale buildings.

Keywords: contrastive learning; feature aggregation; building extraction; remote sensing image; semantic segmentation

Citation: Li, S.; Bao, T.; Liu, H.; Deng, R.; Zhang, H. Multilevel Feature Aggregated Network with Instance Contrastive Learning Constraint for Building Extraction. *Remote Sens.* **2023**, *15*, 2585. https://doi.org/10.3390/rs15102585

Academic Editor: Oktay Baysal

Received: 23 March 2023
Revised: 16 April 2023
Accepted: 12 May 2023
Published: 15 May 2023

1. Introduction

In recent years, the advancement of remote sensing technology has led to the availability of vast amounts of high-resolution remotely sensed imagery, which serves as a reliable data source for regular building extraction. Since buildings are the primary object of land cover and are closely related to human activity, efficiently and accurately extracting buildings from large amounts of high-resolution remote sensing images has significant implications in the fields of land use investigation and illegal building detection [1,2]. However, due to the varied appearance of buildings and the complex backgrounds present in remote sensing images, along with other influencing factors such as sensor differences, view angle, and climate, buildings exhibit clear interclass variability in remote sensing images. Thus, the challenge of extracting buildings from a large range of remotely sensed images with high accuracy and efficiency remains a major one in the field [3].

The development of Deep Convolutional Neural Networks (DCNN) in recent years has led to the proposal of various algorithms for processing remote sensing images [4–6]. It has been demonstrated that data-driven DCNNs can automatically identify specific objects in remote sensing imagery by being trained on large numbers of labeled samples. This technology benefits from Fully Convolutional Networks (FCN) [7], which replace the fully

connected layers with convolutional layers, making dense prediction possible on large-scale remotely sensed images.

The FCN-based method [8] efficiently extracts buildings from remotely sensed imagery, but the boundaries of the buildings may still be blurry due to the loss of spatial information during repeated downsampling operations in the encoder stage. This loss of information can be challenging to recover through upsampling in the decoder stage. To overcome this issue, encoder–decoder-based methods such as U-Net [9] have introduced skip-connections to fuse multi-scale shallow features extracted from the encoder stage, which helps to recover detailed spatial information.

Current building extraction methods primarily rely on encoder–decoder networks, which extract multi-scale features and integrate them to restore detailed spatial information. Advanced network architectures, such as attention mechanisms and transformer networks, have been successfully utilized in remote sensing and have resulted in significant improvements. However, these methods still exhibit certain limitations that hinder their ability to accurately extract buildings.

Firstly, existing methods lack strong generalization ability. They learn features automatically through labeled samples with pixel-wise supervision but fail to consider building-specific global characteristics such as geometry and texture. Consequently, the extracted features may be insufficient for accurately characterizing buildings and differentiating them from their surroundings. Secondly, current methods suffer from blurred building boundaries due to the underutilization of high-resolution features and an unbalanced representation of spatial and semantic information. This leads to inaccuracies in distinguishing adjacent building boundaries, as illustrated in Figure 1.

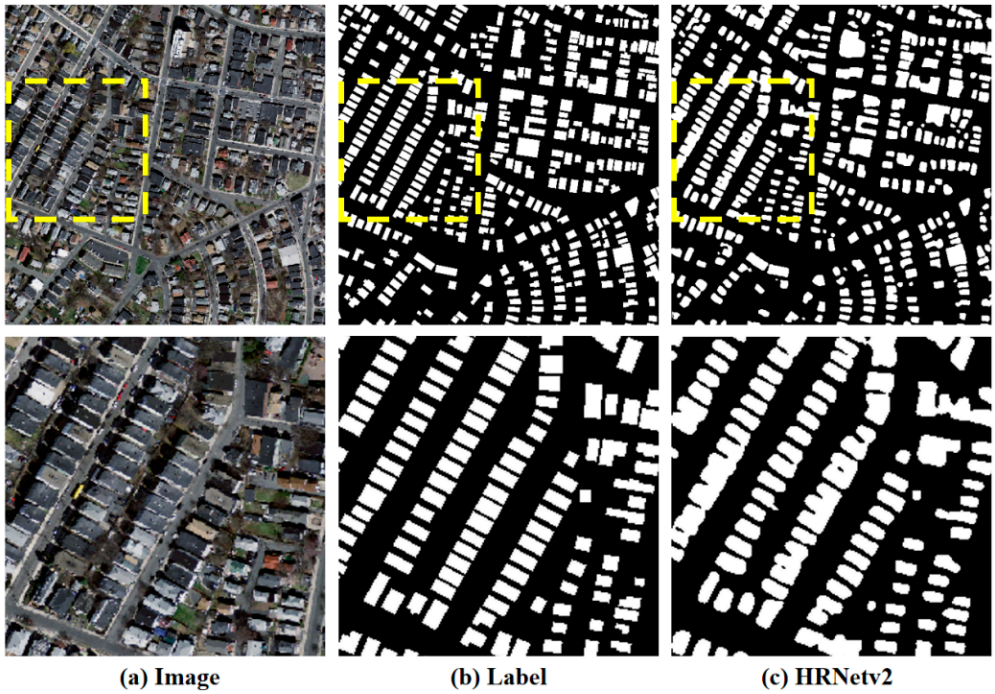


Figure 1. An example of extracted blurred boundaries, especially in densely distributed buildings. Columns (**a**–**c**) represent the image, label, and extracted results by HRNetv2. The second row illustrates the details of the yellow box marked in the first row.

The different architectures of the encoders have varying capabilities in capturing semantic features, each with its advantages. A well-designed strategy for fusing multi-scale features extracted from multiple encoders can combine these advantages and enhance the representation of features. To address the above challenges, we propose an MFA-Net that aggregates multilevel features extracted from two lightweight encoders to maintain multi-scale spatial and semantic context. Our approach adaptively screens and optimizes

the multi-scale features from different backbones, enabling the model to better capture building-specific features. Additionally, we introduce an instance-level contrastive loss based on the multi-scale features with prior mask constraints, which increases the discrimination between buildings and their backgrounds and significantly improves the model's generalization ability.

The main contributions of this work are as follows:

(1) We proposed a channel attention-based multilevel feature aggregation framework that adaptively fuses the spatial and semantic information, resulting in refined multi-scale building extraction;
(2) We devised an instance-level contrastive learning strategy with global consistency constraints among buildings and backgrounds, which leads to robust feature representation;
(3) We conducted extensive experimentation on the WHU and Massachusetts datasets to validate the performance and generalization of the proposed method.

The paper is organized as follows: In Section 2, we explore the existing literature on building extraction. The details of the proposed method are illustrated in Section 3, followed by Section 4, which describes the experiments and analyzes the results obtained from them. Lastly, in Sections 5 and 6, we present a comprehensive discussion and conclusions of this paper.

2. Related Works

In the past few decades, lots of algorithms have been developed for extracting buildings from remotely sensed imagery. These methods can generally be categorized into two groups: traditional image processing-based methods and deep convolutional neural network (DCNN)-based methods. Traditional methods for building extraction rely on designing features that capture the texture, geometry, and shadow characteristics of the buildings in the imagery [10–12]. However, traditional methods tend to not perform well on a task with large-scale remote sensing imagery, since these specific features vary under different illuminations, sensor types, and building architecture. Many researchers combine optical imagery with multi-source data containing elevation information to distinguish non-building areas that are similar to buildings [4,13]. This significantly increases the robustness of building recognition; however, it is costly to obtain such a wide range of corresponding multi-source data. In summary, traditional methods are heavily based on prior knowledge to design appropriate features according to the specific data. It is unfeasible to extract buildings from large-scale remote sensing images efficiently due to the complex background and the diversity in appearance and scale of buildings in remote sensing imagery.

DCNN-based building extraction methods have progressed with the development of the fully convolutional network (FCN), realizing efficient pixel-to-pixel classification of input images with any size. FCN-based methods [8,14–18] automatically extract building footprints from the original images pixel to pixel and achieve high-quality results. Due to the repeated downsample operation in the stage of feature extraction, detailed spatial information was also weakened. To restore the potential spatial information that gets lost in the feature extraction process, encoder–decoder-based semantic segmentation methods have been developed, represented by the U-Net [9], gradually fusing high-resolution features by skip-connection on the decoder stage.

Building extraction frameworks based on the encoder–decoder architecture [19–23] have demonstrated significant advancements in accurately extracting buildings compared to FCN-based methods, particularly for tiny buildings and boundaries. To further enhance the quality of extracted building boundaries, recent studies [24–27] have incorporated the structural information of buildings to improve the recognition and optimization of building edges, by utilizing potential a priori forms of contours and employing post-processing strategies. Additionally, learning-based regularized algorithms have been introduced to refine the extracted building boundaries and generate building polygons, as demonstrated in [28–31]. These approaches have shown promising results in improving the accuracy and

efficiency of building extraction and are expected to have a significant impact on future applications in remote sensing and geospatial analysis.

To tackle the problem of gradient explosion during backpropagation in deeper neural network architectures that hinders further performance improvements, ResNet [32] introduced a novel residual block that employs identity mapping. This breakthrough allowed for the design of deeper and more complex deep convolutional neural networks (DCNNs), facilitating richer semantic feature extraction in image recognition tasks. As a result, numerous frameworks [5,23,33] have utilized pre-trained ResNet models on large datasets to encode features. These methods have significantly surpassed others that lack transferred knowledge in tasks such as semantic segmentation and building extraction, among others.

The issue of multi-scale building extraction has been addressed in previous works such as those proposed by [34–36]. These authors introduced a specific architecture that integrated multi-scale input to enhance the precision of extracting multi-scale buildings. However, these methods significantly increased computational complexity, which is a significant challenge in practical applications. In response to this challenge, ref. [37] proposed the atrous spatial pyramid pooling module (ASPP), which effectively extracted multi-scale features through pooling layers with various kernel sizes. This approach captured the global context without introducing additional computational complexity. Moreover, ref. [38] presented the SRI-Net, which incorporated multi-scale features by means of a spatial residual inception module, further improving the efficacy of multi-scale building extraction. Additionally, for efficiently refined building extraction, MA-FCN [28] utilized a feature pyramid network for multi-scale feature extraction and a boundary regularization strategy for refined building extraction. The DeepLabv3 series networks [20,39] implemented dilated convolution to enlarge the receptive field and efficiently fuse multi-scale features, thereby improving the accuracy and consistency of object extraction without increasing the computational complexity. It efficiently fused the multi-scale features, which benefited multi-scale object extraction and improved the integrity of the objects.

The attention mechanism is an efficient technique for capturing long-range dependencies in a spatial by exploiting self-relations and optimizing features from a channel perspective. It has been shown to enhance the feature representation of tiny objects and improve the overall integrity by leveraging intra-class similarities, as evidenced by numerous studies [40–42]. To further retain the spatial details, HRNetv2 [43] proposed a high-resolution parallel network to deal with multi-scale feature extraction for more precise semantic segmentation. Based on the above idea, MAP-Net [44] proposed a multi-path parallel feature extraction network that preserves detailed spatial information in high-resolution features while maintaining rich semantic representation in deeper features. Furthermore, an attention-based feature optimization module was introduced to adaptively fuse the multi-scale features, resulting in refined building footprint extraction.

In summary, extant methodologies have emphasized multi-scale building extraction via the incorporation of particular modules and post-processing techniques to refine boundary recognition. Nevertheless, the majority of these methodologies continue to rely on pixel-level supervision for model training and optimization, disregarding instance-level consistency, which denotes the global similarity between buildings and their differentiation from surrounding backgrounds. In addition, the potential of combining various encoders for the extraction of multi-scale features has been infrequently explored.

3. Methods

3.1. Architecture Overview

To alleviate the problems above, we designed a novel instance-level contrastive learning constraint network for accurate multi-scale building extraction, which aggregated two encoders to optimize the multilevel features with rich semantic and detailed spatial information. The proposed MFA-Net is illustrated in Figure 2 and is mainly composed

of two primary modules: (1) a multilevel feature aggregation module (AFFM) and (2) a building instance-level constraint contrastive learning module (ICL).

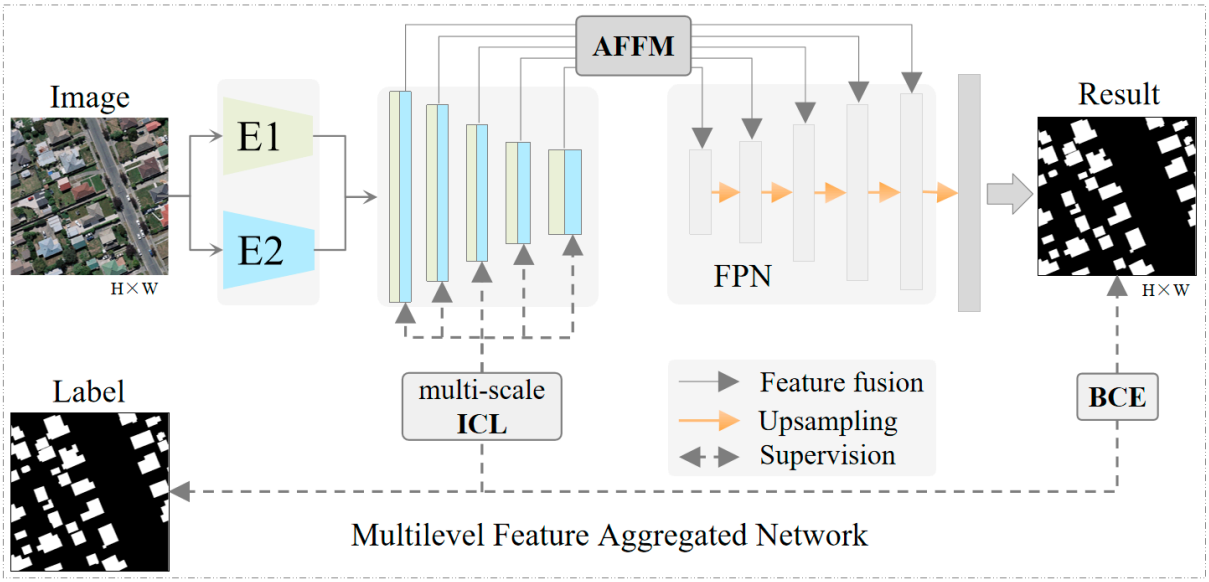


Figure 2. Architecture of the proposed MFA-Net, which includes an attention-based multilevel feature fusion module (AFFM), an instance-level contrastive learning loss (ICL), and a binary cross-entropy loss (BCE). The AFFM fuses multilevel features through an attention mechanism, while the ICL constrains building instances to enhance feature representation robustness. BCE is used as a standard loss function for pixel-wise semantic segmentation.

Specifically, we extracted five distinct scales of features from two pre-trained encoders, HRNet18 and Resnest26, which had differing architectures. To effectively fuse and combine the advantage of each scale of features, we introduced a channel attention-based aggregation module, which utilized channel-wise weighting. The resulting multi-scale features were then fused and decoded using a Feature Pyramid Network (FPN) head, ensuring that spatial details were maintained across the different scales. To optimize the results, we employed a binary cross-entropy loss function. Unlike the existing methods, we designed a building instance constraint contrastive loss to learn the global representation of buildings. This helps to differentiate between buildings and non-buildings through object-level consistency. Further details of these modules are described below.

3.2. Aggregation of Multilevel Features

Considering that the scale varies for buildings, our approach utilizes two lightweight encoders, HRNet18 and Resnest26, to extract multi-scale features. HRNet18 is specifically designed to maintain high-resolution representations throughout its encoder. The encoder consists of four stages, each processing the input at a different scale. In the first stage, the input image is taken and high-resolution feature maps are produced. Subsequently, lower-resolution feature maps are combined with the initial set of feature maps in the subsequent stages to generate even higher-resolution feature maps. The final output of the encoder is a set of feature maps that possess high resolution and are rich in spatial details. On the other hand, Resnest26 is an improved version of ResNet that includes a nested-scale attention mechanism consisting of 26 layers. Similarly to HRNet18, it is composed of multiple stages, with each stage consisting of multiple residual blocks. The attention mechanism allows the network to attend to features at multiple scales, enabling it to capture both local and global contexts. The output of each stage of Resnest26 is then fed into the next stage, ultimately producing a set of multi-scale features, making it a suitable choice for the proposed task.

Compared with the existing methods, we maintained the five scales of features to preserve the spatial details, especially for small buildings. Since different encoders have their particular advantages, we designed an attention-based multilevel feature fusion module (AFFM), as shown in Figure 3a, to optimize and channel-wise squeeze the features extracted from two pre-trained encoders. It adaptively learned a weight for each channel to select and enhance each scale of the fused features without increasing the dimensions.

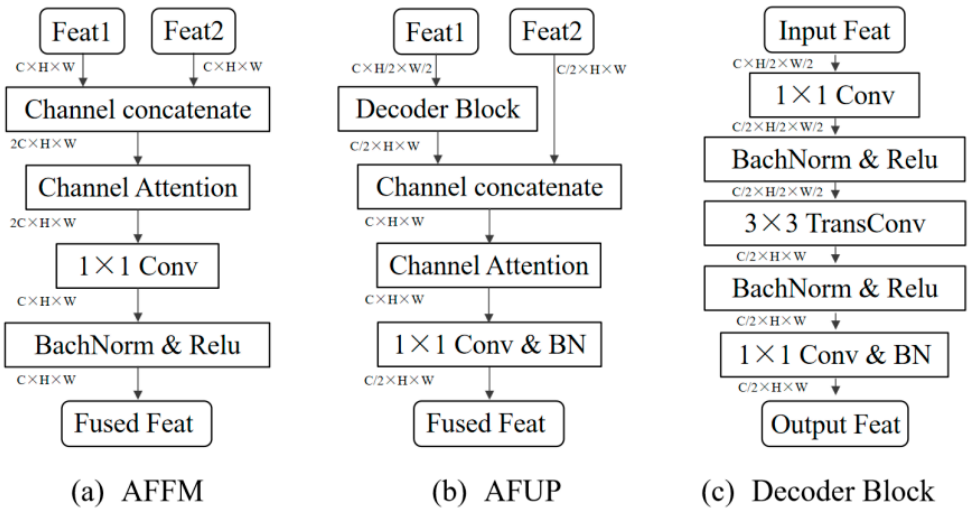


Figure 3. Details of the Attention-based Feature Fusion Module.

The channel attention module was based on the SE-Module with a channel reduction in the middle fully connected layers, which significantly improved the effectiveness, as shown in Figure 4. Apart from the feature aggregation during the encoder stage, we also introduced the attention-based feature fusion layer into the FPN-based upsampling stage (AFUP). Unlike the skip-connection for recovering spatial details, we upsampled the lower resolution features through use of a transconvolution layer, as shown in Figure 3c, and efficiently concatenated them with channel-attention fusion without increasing the dimensions of the feature. Compared with existing methods, which directly upsample the features by linear interpolation, the AFUP module upsampling features take the nearby pixels into account in space and channel-wise to select the optimal features. Additionally, the proposed multilevel feature fusion module balances spatial details and semantic information, which helped to extract multi-scale buildings accurately.

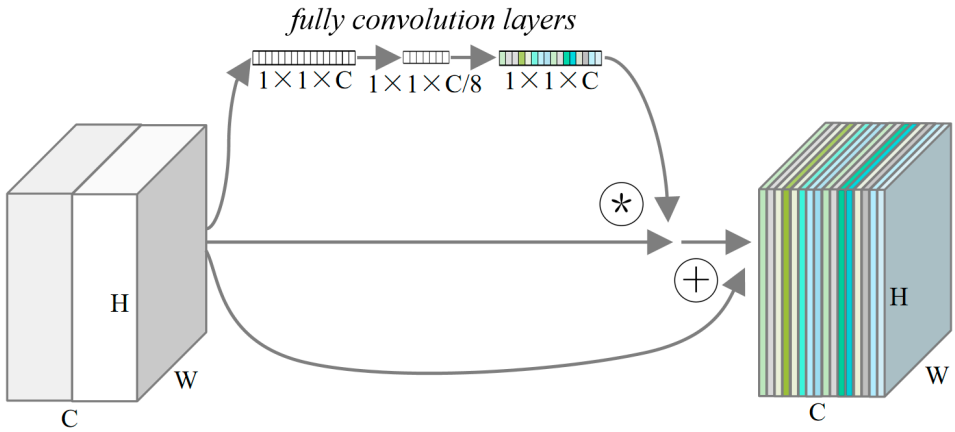


Figure 4. The channel attention module. The W, H, and C represent the dimensions of the features.

3.3. Instance-Level Constraint Contrastive Loss

Pixel-wise contrastive learning improves the robustness and quality of feature representation by increasing the discrimination between positive and negative samples and

decreasing the distance within pairs of positive or negative samples [45–47]. However, the independent pixel-wise features ignore the building's global consistency representations such as structure and texture. Similar to the pixel-wise contrastive loss, we designed an instance-level contrastive loss L_c by increasing the similarity of the paired global building representation mapping to the multilevel features F_a, and F_b extracted from the two encoders, respectively. To learn the discriminated features between the buildings and backgrounds, we randomly sampled a negative sample F^g from the background for each building mask and enlarged the distance of the feature presentation. The distance D was evaluated by cosine similarity. The presentation of the instance-level contrastive loss was shown in Formula (1). The M represents the number of building instances, and the i represents each building instance.

$$L_c = \frac{1}{M}\sum_{i=1}^{M} D\left(F_a^i, F_b^i\right) + \frac{1}{M}\sum_{i=1}^{M}\left(\left[1 - D\left(F_a^i, F_a^g\right)\right] + \left[1 - D\left(F_b^i, F_b^g\right)\right]\right) \quad (1)$$

Apart from the instance-level contrastive loss, we also introduced the binary cross-entropy loss L_{bce} jointly with the Dice loss L_{dice} to optimize the segmentation results. The p_i and y_i represented the pixel-wise predicted results and labels, respectively. To make the model converge better [48], we balanced the L_c, L_{bce}, and L_{dice} by factors α, β, and γ, which were set as 0.6, 0.1, and 0.3, respectively, as shown in Formulas (2)–(4).

$$L_{bce} = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -[y_i \cdot log(p_i) + (1 - y_i) \cdot log(1 - p_i)] \quad (2)$$

$$L_{dice} = 1 - \frac{2\sum_i^N y_i \cdot p_i + 1}{\sum_i^N y_i^2 + \sum_i^N p_i^2 + 1} \quad (3)$$

$$Loss = \alpha \cdot L_{bce} + \beta \cdot L_{dice} + \gamma \cdot L_c \quad (4)$$

3.4. Evaluation Metrics and Training Strategies

The proposed method for building extraction from remote sensing images relies on semantic segmentation, where the objective is to classify each pixel as belonging to either the building or the background. To evaluate the effectiveness of the proposed method, we utilized standard binary segmentation metrics, including recall (R), precision (P), F1-score (F1), and intersection over union (IoU). These metrics are commonly used to evaluate the accuracy of binary segmentation methods and provide a quantitative measure of the performance of the proposed method.

There were three classification conditions to consider: true prediction on the positive sample (TP), false prediction on the positive sample (FP), and false prediction on the negative sample (FN). The recall indicates the percentage of TP among all positive labels, while the precision measures the percentage of TP among all positive predictions. The IoU calculates the overlap between the predicted and labeled regions for buildings, relative to their union. The F1-score is a weighted average of precision and recall, taking both metrics into account. The formulas of these metrics are as follows:

$$R = \frac{TP}{TP + FN} \quad (5)$$

$$P = \frac{TP}{TP + FP} \quad (6)$$

$$IoU = \frac{P * R}{P + R - P * R} \quad (7)$$

$$F1 = \frac{2 * P * R}{P + R} \quad (8)$$

We implemented our method using PyTorch on an 11 Gigabyte memory 2080Ti GPU. To optimize the model, we used the stochastic gradient descent optimizer with an initial learning rate of 0.01, which was weakened adaptively based on the validation accuracy. To ensure fair comparisons, all compared methods were trained for 100 epochs with data augmentation, including random scaling, rotation, and flipping, as detailed in Section 4.1. We also maintained a consistent batch size and other hyperparameters to compare the performance of the proposed MFA-Net against the other methods on the above metrics.

4. Experiments and Results

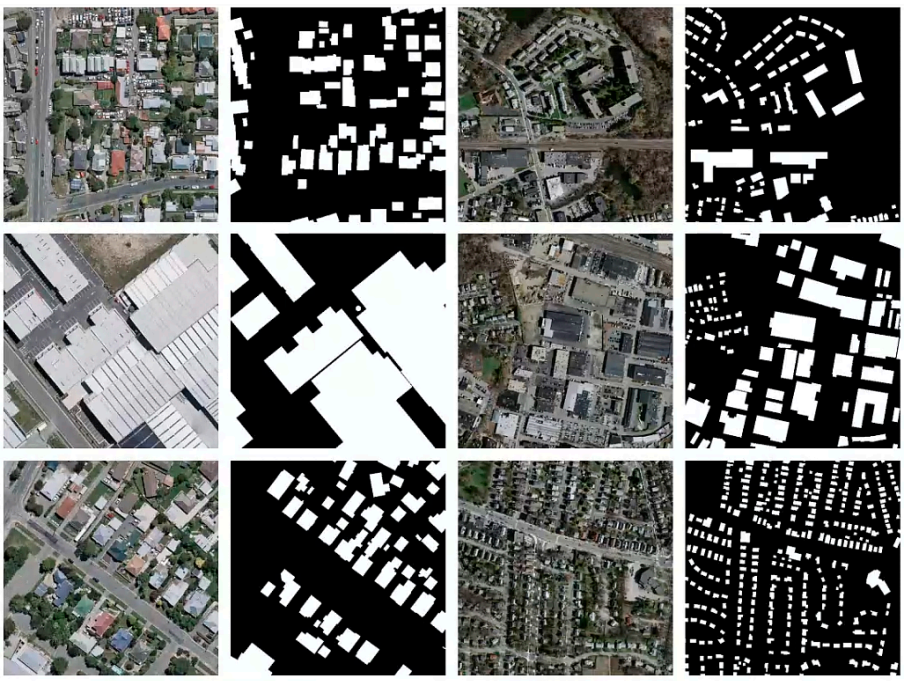
4.1. Datasets Description

We evaluate the effectiveness of the proposed method using two publicly available datasets, the WHU building dataset [14], and the Massachusetts dataset [49].

The WHU building dataset contains a vast array of buildings with varying appearances and sizes, allowing us to evaluate the robustness of the proposed algorithm. This dataset spans over 450 km^2 and comprises more than 187,000 buildings with a resolution of 0.3 m. The images consist of three color bands, each with a size of 512 $\times$ 512 pixels. The dataset is split into training, validation, and test subsets, containing 4736, 1036, and 2416 samples, respectively. We employed the official dataset division for our experiments.

The Massachusetts dataset, on the other hand, covers about 340 km^2 of the Boston area and contains 151 tiles of aerial images and corresponding single-channel labels. The resolution of this dataset is 1 m, and all images have a size of 1500 $\times$ 1500 pixels. The dataset is divided into training, validation, and test subsets, comprising 137, 4, and 10 tiles, respectively. For our experiments, we clipped the dataset into 512 $\times$ 512 pixels with an overlap of 12 pixels.

We illustrated sample images and corresponding building labels on the WHU and Massachusetts datasets in Figure 5. The WHU dataset consists of higher-resolution aerial images, while the Massachusetts dataset consists of lower-resolution satellite images. Buildings in both datasets exhibit significant scale variations, which effectively validate the feasibility and sophistication of our proposed method in accurately identifying fine-grained boundaries of buildings at multiple scales.



(a) Samples of WHU dataset (b) Samples of Massachusetts dataset

Figure 5. Sample images and corresponding labels from the WHU and Massachusetts building extraction datasets.

4.2. Performance Comparison with Semantic Segmentation Methods

To assess the effectiveness of our proposed approach, we conducted a comparative analysis with several classical semantic segmentation methods, namely PSPNet, DeepLabv3+, U-NetPlus, and HRNetv2, on the aforementioned datasets. The PSPNet introduced a different-region-based context aggregation module for gathering global context information to improve the quality of the results. The DeepLabv3+ explored the depth-wise atrous separable convolution on the decoder, which effectively refined the segmentation results, especially for the object boundary. The U-NetPlus introduced the pre-trained encoder and replaced the transposed convolution operation with upsampling based on U-Net which significantly improved the performance. The HRNetv2 designed a multiple-path network for multi-scale feature extraction to recover detailed spatial information. We implemented these compared methods according to the public source code. The experimental results on the WHU dataset are shown in Table 1.

Table 1. Performance Comparison of Classical Semantic Segmentation Methods and Our Proposed Approach on the WHU Datasets.

Method	IoU (%)	Precision (%)	Recall (%)	F1-Score (%)
DeepLabv3+	88.16	94.64	92.79	93.71
PSPNet	88.81	94.33	93.82	94.07
U-NetPlus	89.35	94.73	94.02	94.37
HRNetv2	90.09	93.60	96.01	94.79
MFA-Net (Ours)	91.07	94.64	96.02	95.33

The proposed method demonstrated superior performance in comparison to other semantic segmentation methods, as evidenced by its higher IoU and F1 metrics. In fact, our method achieved a 0.98% IoU improvement when compared to the HRNetv2. To facilitate the comparison of results obtained from our method with those from other techniques, we included visual examples of results on the WHU dataset, which can be seen in Figure 6. The figure presents two sample images alongside their corresponding extracted results. The details of the areas marked with the red dashed box are shown in the right-hand column for comparison.

To further evaluate the performance and generalization of the proposed method, we conducted a comparative experiment on the Massachusetts building extraction dataset, which with a lower image resolution. We re-implemented the above classical methods for semantic segmentation including PSPNet, DeepLabv3+, U-NetPlus, and HRNetv2 on the Massachusetts dataset. The experimental results, as shown in Table 2, demonstrate that our method outperformed the other methods in terms of IoU and F1 metrics, achieving a 2.26% improvement in IoU compared to HRNetv2. Furthermore, we provide visual comparison results on the Massachusetts dataset in Figure 7. The right-hand column of the figure highlights the detailed part, marked with a red box for easier comparison, between the results obtained by our method and those obtained by the other methods.

Table 2. Performance Comparison of the Classical Semantic Segmentation Methods and Our Proposed Approach on the Massachusetts Datasets.

Method	IoU (%)	Precision (%)	Recall (%)	F1-Score (%)
DeepLabv3+	67.36	82.76	78.35	80.49
PSPNet	71.64	86.46	80.69	83.48
U-Net	71.05	84.02	82.15	83.07
HRNetv2	72.32	85.34	82.58	83.94
MFA-Net (Ours)	74.58	87.11	83.84	85.44

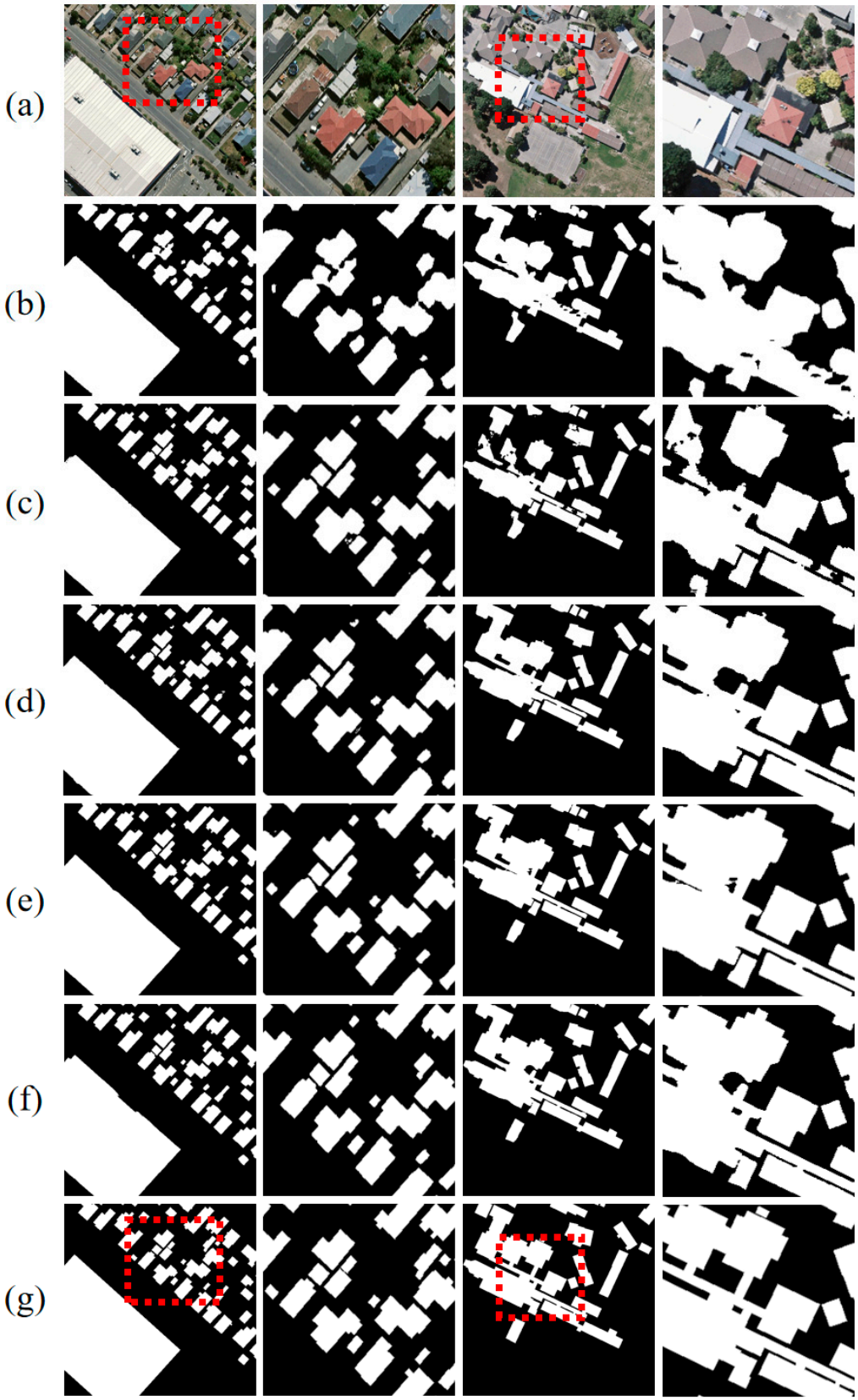


Figure 6. Comparison of the results obtained from the WHU test dataset by various segmentation methods. The rows (**a**,**g**) represent the sampled images and labels, respectively. The rows (**b**–**f**) illustrate the results extracted by DeepLabv3+, PSPNet, U-NetPlus, HRNetv2, and our method. The second and last columns are the local areas marked by red boxes.

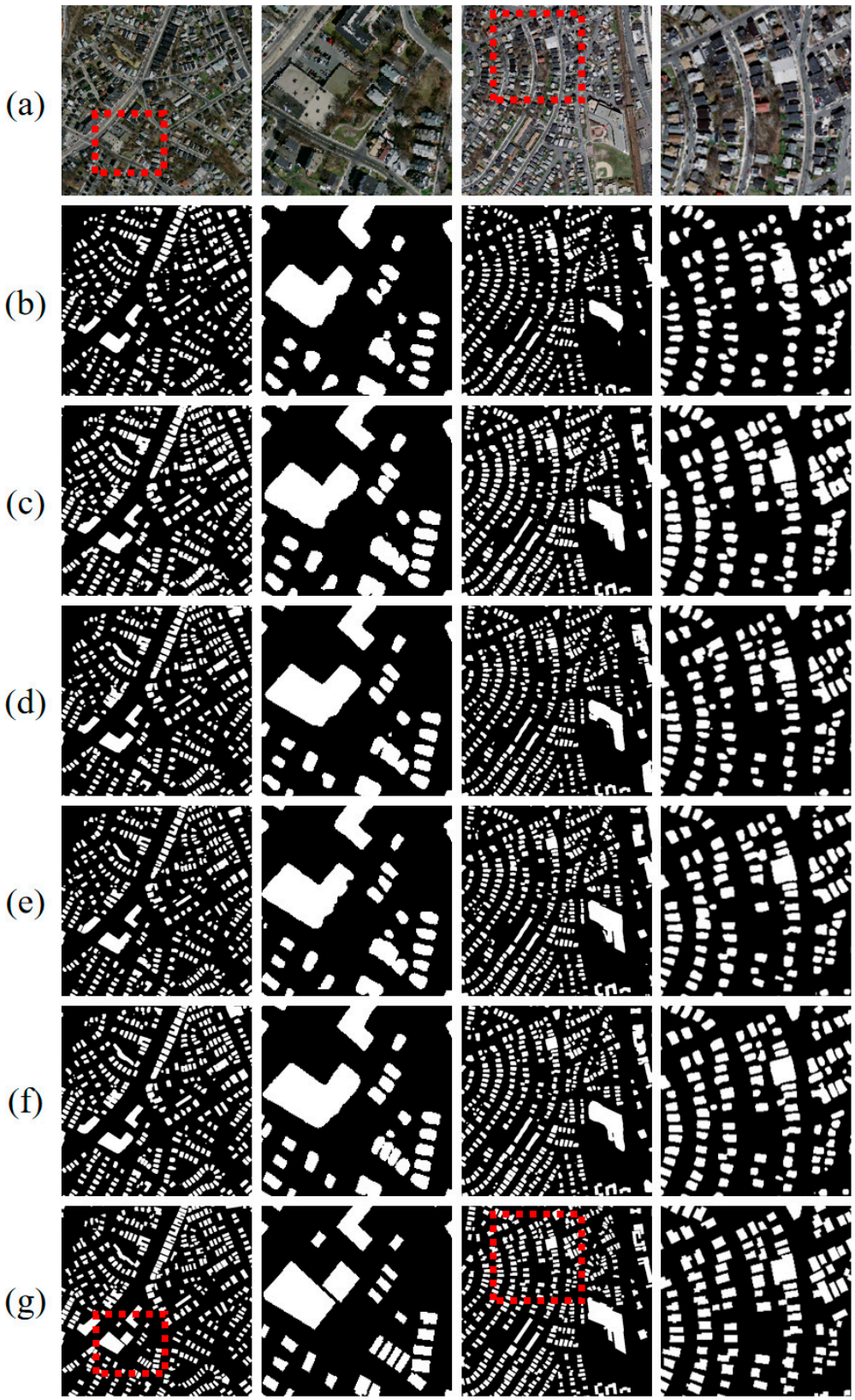


Figure 7. Comparison of the results obtained from the Massachusetts test dataset by various segmentation methods. The rows (**a**,**g**) represent the sampled images and labels, respectively. The rows (**b**–**f**) illustrate the results extracted by DeepLabv3+, PSPNet, U-NetPlus, HRNetv2, and our method. The second and last columns are the local areas marked by red boxes.

As can be seen from the results of the two datasets, our methods extracted a more refined boundary than that of the compared methods, especially for distinguishing the dense and close buildings. It benefited from the multi-scale features aggregation module with an attention-based adaptive fusion, which maintained the multilevel information contained in the multi-scale features and balanced the tradeoff between the semantic and spatial details. Additionally, the instance-level contrastive loss enhanced the difference between the buildings and backgrounds through global representation, which helped to extract the discriminable features.

4.3. Comparison of Recent Related Methods

To fairly evaluate the significance of our proposed method, we compared the proposed method with the most recently related building extraction methods including MC-FCN [18], RSR-Net [15], SiU-Net [14], SRI-Net [38], BOMSC-Net [50], HD-Net [51], EU-Net [52], CBR-Net [25], MA-FCN [28], MAP-Net [44], BRRNet [17], and BP-Net [25] on the above two datasets. The experimental outcomes are displayed in Table 3.

Table 3. Performance Comparison of the Most Recently Related Building Extraction Methods with Ours on the WHU and Massachusetts Datasets.

Datasets	Methods	IoU	Precision	Recall	F1-Score
WHU	MC-FCN	87.10	94.60	91.70	93.13
	RSR-Net	88.32	94.92	92.63	93.76
	SiU-Net	88.40	93.80	93.90	93.85
	SRI-Net	89.23	95.67	93.69	94.51
	BOMSC-Net	90.15	94.80	95.14	94.50
	HDNet	90.50	95.20	94.80	95.00
	EU-Net	90.56	94.98	95.10	95.04
	MA-FCN	90.70	95.20	95.10	95.15
	MAP-Net	90.86	95.62	94.81	95.21
	MFA-Net (Ours)	91.07	94.64	96.02	95.33
Massachusetts	MAP-Net	73.34	85.49	83.76	84.62
	EU-Net	73.93	86.70	83.40	85.01
	BRRNet	74.46	–	–	85.36
	BP-Net	74.51	85.44	85.34	85.39
	CBR-Net	74.55	86.50	84.36	85.42
	BOMSC-Net	74.71	86.64	84.68	85.13
	MFA-Net (Ours)	74.58	87.11	83.84	85.44

For the WHU building extraction dataset, the MC-FCN designed a multi-constraint on the multi-scale features based on FCN for accurate multi-scale building extraction. The SiU-Net proposed a Siamese U-Net with two branches which significantly improved the segmentation accuracy, especially for large buildings. The SRI-Net extracted multi-scale features based on the ResNet101 backbone and aggregated the multilevel contexts, achieving 89.23% in the IoU metric. The EU-Net captured multi-scale features by introducing the DSPP module and a focal loss to alleviate the unbalanced building label issue. This method outperforms previous approaches, achieving a remarkable 90.56% in the IoU metric. Similarly, the MA-FCN extracted multi-scale features based on the FPN, combined with a postprocessing module to refine the extracted building boundary, obtaining comparable results to the EU-Net. The MAP-Net constructed a parallel multi-scale feature extraction network with an attention enhancement module, which outperformed the other methods and achieved an outstanding 90.86% in IoU.

In the case of the Massachusetts dataset, MAP-Net and EU-Net achieved 73.34% and 73.93%, respectively, in the IoU metric, since the image resolution was much lower than the WHU. The BRRNet, on the other hand, designed a refinement module for building-boundary optimization based on the encoder–decoder framework, which enabled it to outperform previous methods with a score of 74.46% in IoU. The BP-Net introduced a

structural constraint module based on the MAP-Net, which significantly improved the performance of building boundaries and achieved 74.51% in IoU.

In summary, our method surpassed most recently related building extraction methods under comparison, both on the WHU and Massachusetts datasets. This infers the importance of context aggregation from multiple encoders as well as the significance of instance-level global building representation constraints for more robust building extraction from remote sensing imagery.

4.4. Ablation Experiments on the WHU Dataset

To explore the significance of each proposed module, we conducted an ablation experiment on the WHU dataset. Our backbone included a lightweight HRNetv2 with a depth of 18 convolution layers and an improved ResNet with a depth of 26 convolution layers represented as HRNet18 and Resnest26, respectively. The experimental results are shown in Table 4.

Table 4. Comparison of the Performance of each Module for the Proposed Method on the WHU Dataset. The √ and × represent the corresponding module included in the method or.

Methods	MEA	ICL	AUP	IoU	Precision	Recall	F1-Score
HRNet18	×	×	×	90.08	93.59	96.01	94.79
Resnest26	×	×	×	90.03	93.72	95.81	94.76
+ MEA	√	×	×	90.36	93.90	96.00	94.94
+ ICL	√	√	×	90.77	94.71	95.62	95.16
+ AUP	√	√	√	91.07	94.63	96.02	95.33

First, we evaluated these two baselines with an FPN-based decoder which fused the multi-scale features with skip-connection. The HRNet18 and Resnest26 achieved 90.08% and 90.03% on the IoU metric, respectively. Second, we evaluated the multiple encoder aggregation module (MEA) which fused the multilevel features extracted from those two encoders with a channel attention-based squeeze. The MEA module achieved 90.36% in the IoU metric, because the fused feature adaptively combined the optimal features. Then, we added the instance-level contrastive loss (ICL) which captured the global consistency constraint. It gained about a 0.41% boost compared to the MEA in the IoU metric, which was the most effective module in our method. Finally, we introduced the spatial and channel attention module during the upsampling decoder stage (AUP). With this, the proposed method achieved 91.07% in the IoU metric, which obtained a 0.3% performance improvement.

To summarize, the proposed modules achieved an obvious improvement compared to the baseline, especially benefiting from the aggregated multiple encoders and object-level contrastive constraints. Since the multiple encoders were lightweight, they did not cause a heavy computational complexity. To compare the different methods, we show the extracted samples in Figure 8.

Figure 8. Visualization of the extracted results on the WHU test dataset. The rows (**a**,**g**) represent the images and labels, respectively. The rows (**b**–**f**) illustrate the results extracted by our ablation experiment.

5. Discussion

In this article, we investigated the efficiency of different encoders for multi-scale feature extraction focusing on building extraction methods. Existing methods predominantly rely on local pixel-wise feature representation while neglecting the importance of global constraints, especially in the context of building-specific characteristics such as structure and texture. Therefore, we proposed the multi-encoder combination, multi-scale feature fusion, and instance-level contrastive loss optimization method that takes advantage of multiple feature encoders and instance-level global constraints, which are under-utilized by methods for building extraction. Since we utilized lightweight encoders to extract multi-scale features while integrating their advantages and improving the global discriminability of building and non-building, we ensured low model complexity. In terms of the loss function, considering the significant differences between the feature representations of buildings and their environments, we introduced instance-level constraint information to guide the model to focus on pixel-level classification while also considering the global building features, thereby improving the generalization of the model. Due to the higher resolution of the aerial imagery in the WHU dataset compared to that of the Massachusetts dataset, there is a significant difference in building extraction accuracy between the two datasets. Nevertheless, the experimental results validate the effectiveness of the proposed MFA-Net. Additionally, our method combines features from different encoders and improves the foreground–background differences by utilizing object-level contrastive learning. This approach can be applied effectively to the extraction of other land covers as well.

6. Conclusions

Our study presents a novel building extraction framework MFA-Net, which aggregates multilevel features from different encoders with an attention-based feature adaptive fusion module. In addition, we propose an instance-level contrastive loss with the constraint of building masks. This is beneficial to enhance the discrimination of features between buildings and backgrounds. We evaluate the proposed MFA-Net through comparative experiments on two publicly available building extraction datasets. Our experimental results indicate the superiority of our approach over existing semantic segmentation methods, and even outperform the most recent related building extraction works, demonstrating its efficacy in accurately extracting buildings while overcoming the limitations of previous methods. Furthermore, we conducted an ablation study to assess the significance of each proposed module, and our findings demonstrate the potential of the proposed framework, particularly in capturing building boundaries.

Author Contributions: Conceptualization, S.L.; Methodology, S.L. and H.L.; Software, S.L., T.B. and H.Z.; Validation, S.L. and T.B.; Formal analysis, S.L. and H.L.; Investigation, S.L. and H.L.; Data curation, S.L., T.B. and R.D.; Writing—original draft preparation, S.L.; Writing—review and editing, S.L. and H.Z.; Visualization, S.L. and T.B.; Project administration, S.L. and H.Z.; Funding acquisition, S.L. and R.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by China Henan Province Science and technology project, No. 232102320068 and funded by Joint Fund of Collaborative Innovation Center of Geo-Information Technology for Smart Central Plains, Henan Province and Key Laboratory of Spatiotemporal Perception and Intelligent processing, Ministry of Natural Resources, No. 212110, and funded by the National Natural Science Foundation of China, No. 31971723.

Data Availability Statement: The data presented in this study are openly available in reference number [14,49].

Acknowledgments: The authors would like to thank the editor and reviewers for their contributions on the paper.

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

References

1. Mishra, A.; Pandey, A.; Baghel, A.S. Building detection and extraction techniques: A review. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016; pp. 3816–3821.
2. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
3. 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]
4. Awrangjeb, M.; Zhang, C.; Fraser, C.S. Automatic extraction of building roofs using LIDAR data and multispectral imagery. *ISPRS J. Photogramm. Remote Sens.* **2013**, *83*, 1–18. [CrossRef]
5. Li, S.; Liao, C.; Ding, Y.; Hu, H.; Jia, Y.; Chen, M.; Xu, B.; Ge, X.; Liu, T.; Wu, D. Cascaded Residual Attention Enhanced Road Extraction from Remote Sensing Images. *ISPRS Int. J. Geo-Inform.* **2022**, *11*, 9. [CrossRef]
6. Afaq, Y.; Manocha, A. Analysis on change detection techniques for remote sensing applications: A review. *Ecol. Inform.* **2021**, *63*, 101310. [CrossRef]
7. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. Available online: https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html (accessed on 25 July 2022).
8. Bittner, K.; Cui, S.; Reinartz, P. Building extraction from remote sensing data using fully convolutional networks. *ISPRS—Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *XLII-1/W1*, 481–486. [CrossRef]
9. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. [CrossRef]
10. 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]
11. Cote, M.; Saeedi, P. Automatic Rooftop Extraction in Nadir Aerial Imagery of Suburban Regions Using Corners and Variational Level Set Evolution. *IEEE Trans. Geosci. Remote Sens.* **2013**, *51*, 313–328. [CrossRef]
12. Li, Q.; Wang, Y.; Liu, Q.; Wang, W. Hough Transform Guided Deep Feature Extraction for Dense Building Detection in Remote Sensing Images. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 1872–1876. [CrossRef]
13. Huang, Z.; Cheng, G.; Wang, H.; Li, H.; Shi, L.; Pan, C. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. In Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 1835–1838. [CrossRef]
14. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 574–586. [CrossRef]
15. 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]
16. Shrestha, S.; Vanneschi, L. Improved Fully Convolutional Network with Conditional Random Fields for Building Extraction. *Remote Sens.* **2018**, *10*, 1135. [CrossRef]
17. 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]
18. Wu, G.; Shao, X.; Guo, Z.; Chen, Q.; Yuan, W.; Shi, X.; Xu, Y.; Shibasaki, R. Automatic Building Segmentation of Aerial Imagery Using Multi-Constraint Fully Convolutional Networks. *Remote Sens.* **2018**, *10*, 407. [CrossRef]
19. Liu, H.; Luo, J.; Huang, B.; Hu, X.; Sun, Y.; Yang, Y.; Xu, N.; Zhou, N. DE-Net: Deep Encoding Network for Building Extraction from High-Resolution Remote Sensing Imagery. *Remote Sens.* **2019**, *11*, 2380. [CrossRef]
20. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11211, pp. 833–851. [CrossRef]
21. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
22. Hasan, S.M.K.; Linte, C.A. U-NetPlus: A Modified Encoder-Decoder U-Net Architecture for Semantic and Instance Segmentation of Surgical Instruments from Laparoscopic Images. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin/Heidelberg, Germany, 23–27 July 2019; pp. 7205–7211. [CrossRef]
23. Diakogiannis, F.I.; Waldner, F.; Caccetta, P.; Wu, C. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *162*, 94–114. [CrossRef]
24. Liao, C.; Hu, H.; Li, H.; Ge, X.; Chen, M.; Li, C.; Zhu, Q. Joint Learning of Contour and Structure for Boundary-Preserved Building Extraction. *Remote Sens.* **2021**, *13*, 1049. [CrossRef]

25. 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]
26. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 242–2424. [CrossRef]
27. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
28. Wei, S.; Ji, S.; Lu, M. Toward Automatic Building Footprint Delineation from Aerial Images Using CNN and Regularization. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2178–2189. [CrossRef]
29. Zorzi, S.; Bittner, K.; Fraundorfer, F. Machine-learned Regularization and Polygonization of Building Segmentation Masks. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 3098–3105. [CrossRef]
30. Chen, Q.; Wang, L.; Waslander, S.L.; Liu, X. An end-to-end shape modeling framework for vectorized building outline generation from aerial images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *170*, 114–126. [CrossRef]
31. Zhao, W.; Persello, C.; Stein, A. Building outline delineation: From aerial images to polygons with an improved end-to-end learning framework. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 119–131. [CrossRef]
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
33. Wang, H.; Miao, F. Building extraction from remote sensing images using deep residual U-Net. *Eur. J. Remote Sens.* **2022**, *55*, 71–85. [CrossRef]
34. Ji, S.; Wei, S.; Lu, M. A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery. *Int. J. Remote Sens.* **2019**, *40*, 3308–3322. [CrossRef]
35. Li, L.; Liang, J.; Weng, M.; Zhu, H. A Multiple-Feature Reuse Network to Extract Buildings from Remote Sensing Imagery. *Remote Sens.* **2018**, *10*, 1350. [CrossRef]
36. Sun, G.; Huang, H.; Zhang, A.; Li, F.; Zhao, H.; Fu, H. Fusion of Multiscale Convolutional Neural Networks for Building Extraction in Very High-Resolution Images. *Remote Sens.* **2019**, *11*, 227. [CrossRef]
37. 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. Available online: https://openaccess.thecvf.com/content_cvpr_2017/html/Zhao_Pyramid_Scene_Parsing_CVPR_2017_paper.html (accessed on 17 October 2022).
38. Liu, P.; Liu, X.; Liu, M.; Shi, Q.; Yang, J.; Xu, X.; Zhang, Y. Building Footprint Extraction from High-Resolution Images via Spatial Residual Inception Convolutional Neural Network. *Remote Sens.* **2019**, *11*, 830. [CrossRef]
39. 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.* **2018**, *40*, 834–848. [CrossRef]
40. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149. [CrossRef]
41. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
42. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803. [CrossRef]
43. 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.* **2021**, *43*, 3349–3364. [CrossRef]
44. Zhu, Q.; Liao, C.; Hu, H.; Mei, X.; Li, H. MAP-Net: Multiple Attending Path Neural Network for Building Footprint Extraction From Remote Sensed Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6169–6181. [CrossRef]
45. Mekhazni, D.; Dufau, M.; Desrosiers, C.; Pedersoli, M.; Granger, E. Camera Alignment and Weighted Contrastive Learning for Domain Adaptation in Video Person ReID. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 1624–1633.
46. Thota, M.; Leontidis, G. Contrastive domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2209–2218.
47. Kang, G.; Jiang, L.; Yang, Y.; Hauptmann, A.G. Contrastive adaptation network for unsupervised domain adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4893–4902.
48. Vandenhende, S.; Georgoulis, S.; Van Gansbeke, W.; Proesmans, M.; Dai, D.; Van Gool, L. Multi-Task Learning for Dense Prediction Tasks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3614–3633. [CrossRef] [PubMed]
49. Mnih, V. *Machine Learning for Aerial Image Labeling*; University of Toronto: Ottawa, ON, Canada, 2013.
50. 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]

51. Li, J.; Zhuang, Y.; Dong, S.; Gao, P.; Dong, H.; Chen, H.; Chen, L.; Li, L. Hierarchical Disentangling Network for Building Extraction from Very High Resolution Optical Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 1767. [CrossRef]
52. 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]