remote sensing

MDPI

Article

SCA-Net: Multiscale Contextual Information Network for Building Extraction Based on High-Resolution Remote Sensing Images

Yuanzhi Wang [1,2,3], Qingzhan Zhao [1,2,3,*], Yuzhen Wu [1,2,3], Wenzhong Tian [1,2,4] and Guoshun Zhang [1,2,3]

[1] College of Information Science and Technology, Shihezi University, Shihezi 832002, China; yzwang@stu.shzu.edu.cn (Y.W.); wuyuzhen@stu.shzu.edu.cn (Y.W.); twz_inf@stu.shu.edu.cn (W.T.); zgs_inf@shzu.edu.cn (G.Z.)
[2] Geospatial Information Engineering Research Center, Xinjiang Production and Construction Crops, Shihezi 832002, China
[3] Xinjiang Production and Construction Corps Industrial Technology Research Institute, Shihezi 832002, China
[4] College of Mechanical and Electrical Engineering, Shihezi University, Shihezi 832002, China
* Correspondence: zqz_inf@shzu.edu.cn

Abstract: Accurately extracting buildings is essential for urbanization rate statistics, urban planning, resource allocation, etc. The high-resolution remote sensing images contain rich building information, which provides an important data source for building extraction. However, the extreme abundance of building types with large differences in size, as well as the extreme complexity of the background environment, result in the accurate extraction of spatial details of multi-scale buildings, which remains a difficult problem worth studying. To this end, this study selects the representative Xinjiang Tumxuk urban area as the study area. A building extraction network (SCA-Net) with feature highlighting, multi-scale sensing, and multi-level feature fusion is proposed, which includes Selective kernel spatial Feature Extraction (SFE), Contextual Information Aggregation (CIA), and Attentional Feature Fusion (AFF) modules. First, Selective kernel spatial Feature Extraction modules are used for cascading composition, highlighting information representation of features, and improving the feature extraction capability. Adding a Contextual Information Aggregation module enables the acquisition of multi-scale contextual information. The Attentional Feature Fusion module bridges the semantic gap between high-level and low-level features to achieve effective fusion between cross-level features. The classical U-Net, Segnet, Deeplab v3+, and HRNet v2 semantic segmentation models are compared on the self-built Tmsk and WHU building datasets. The experimental results show that the algorithm proposed in this paper can effectively extract multi-scale buildings in complex backgrounds with IoUs of 85.98% and 89.90% on the two datasets, respectively. SCA-Net is a suitable method for building extraction from high-resolution remote sensing images with good usability and generalization.

Keywords: high-resolution remote sensing imagery; building extraction; deep learning; semantic segmentation

Citation: Wang, Y.; Zhao, Q.; Wu, Y.; Tian, W.; Zhang, G. SCA-Net: Multiscale Contextual Information Network for Building Extraction Based on High-Resolution Remote Sensing Images. *Remote Sens.* **2023**, *15*, 4466. https://doi.org/10.3390/rs15184466

Academic Editors: Yonas Zewdu Ayele and Wen Liu

Received: 26 July 2023
Revised: 7 September 2023
Accepted: 8 September 2023
Published: 11 September 2023

1. Introduction

Buildings are the main vehicle for human life and development. The building density contains key information for urban development. Accurate building detection data play a vital role in environmentally friendly urban planning, business programming, land use change detection, national defense construction, disaster monitoring, and early warning [1–5]. For instance, analyzing building information through an integrated disaster monitoring system allows for early detection of disaster signs, and after a disaster, detailed building information facilitates the planning of new infrastructure, residential areas, and public facilities [6,7]. With the continuous rapid growth of sensors and space

techniques, the spatial resolution of remote sensing images (RSIs) is becoming higher, and the update period is becoming shorter. Remote sensing data has been used on a large scale to obtain building information [8,9]. Accurate and efficient extraction of buildings from high-resolution RSIs remains a key direction of research due to the spectral differences in buildings, the variety of building types and dimensions, and the influence of the complex background environment [10–12].

Deep learning has seen extensive application in recent years, encompassing target detection, image classification, and image segmentation [13]. Convolutional neural networks (CNNs) can automatically extract and classify hierarchical features within a single model. CNNs can also automate the feature selection process, as conducted by traditional methods [14]. A Fully Convolutional Neural network (FCN) [13] is based on traditional CNN but replaces the fully connected layer of the classification network with a convolutional layer. During down sampling, features are extracted, and up sampling is then employed to recover the original image size. FCN is gradually becoming the basic framework for many semantic segmentation networks [15,16]. Ronneberger et al. [17] propose a U-Net network based on FCN. The encoder part learns low-level features from the input through convolution and pooling operations. The decoder part uses convolution and upsampling to ensure the original image size in the output, and cascading operations are performed between the encoder and decoder. Thanks to the excellent architecture and powerful performance of U-Net, it is widely used in building extraction tasks [18]. Si et al. [19] improve the combination of encoder and decoder based on Deeplab v3+ [20], introduce three parallel channel attention mechanisms to improve the extraction of deep features, and realize high-accuracy building extraction. Seonsyeong et al. [21] realize high-accuracy building extraction by applying HRNet-v2 [22] and combining the channel and spatial attention modules to effectively learn important features, fusing richer cross-layer semantic features, and improving both building extraction accuracy and speed. Shi et al. introduced a spatial channel attention mechanism based on U-Net to improve the feature extraction capability of the model [23]. Aryal incorporates multi-scale feature maps with parts of a Feature Pyramid Network (FPN) into the U-Net framework to obtain higher building extraction accuracy and robustness [24]. Xu proposes that the backbone of the U-Net encoder is replaced by a ResNeXt101 network for feature extraction, and a feature pyramid structure is used to fuse feature maps at different scales to improve the accuracy of building segmentation for small sample sizes [25].

In recent years, Transformer has also been applied to building extraction, where its powerful feature representation and ability to establish long-term dependencies between pixels are important for building extraction. Li et al. embedded a transposed convolutional sampling module incorporating multiple normalized activation layers into a decoder based on the SegFormer network to overcome the problems of loss of detailed information about local buildings and lack of information at a distance [26]. Wei proposed a multi-scale adaptive segmentation network model (MSST-Net) based on the Swin transformer to realize high-precision extraction of buildings [27]. Chen designed an efficient two-channel transformer structure (SST) to realize the reduction of transformer computational complexity [28]. Kirillov et al. used a large number of masks for training and proposed the Segment Anything Model (SAM) that can segment any target object with pre-prompting [29]. Chen et al. proposed the RSPrompter based on the SAM model in combination with prompt learning, which enables SAM to produce semantically recognizable segmentation results for remote sensing images [30]. Although Transformer has many advantages in building extraction, its complexity is still large, and its performance may not be good when the training dataset is small.

According to the above literature, the spectral differences, background complexity, and large scale differences of buildings pose a challenge for extracting buildings in high-resolution RSIs by directly utilizing existing semantic segmentation networks [31,32]. The encoder in the semantic segmentation network lacks the ability to effectively capture low-level feature representations, leading to reduced spatial information in building features

and an abundance of redundant information. Consequently, it fails to convey the precise spatial details of the buildings. The maximum pooling operation reduces computational complexity by integrating global information. However, for feature maps with wide perceptual areas, ordinary convolution can only capture local information. The jump connection improves the utilization of underlying features and facilitates the incorporation of high-level semantic information with low-level features. However, it ignores the effect of superfluous information and the semantic gap between low-level features and high-level features, thus limiting the performance of multiple scales of building extraction.

Deep learning being data-driven, it heavily relies on the diversity and quality of datasets for effective building extraction. The current building extraction datasets mainly include the WHU Building [33] and Massachusetts [34] datasets. The WHU Building dataset includes a large range of satellite and aerial imagery, and the coverage mainly includes Europe, the United States, New Zealand, and East Asia, with a variety of building styles and dense building coverage. The Massachusetts dataset primarily comprises aerial images of the Boston area, encompassing both urban and suburban regions. These two datasets have been used in many studies [35–37], mainly for buildings with architectural styles in foreign regions. Domestic towns, especially those with low urbanization, have obvious differences between building styles and morphology and foreign buildings, which makes the extraction of buildings more difficult, but the number of datasets extracted for buildings in domestic towns and cities is relatively small at present.

To address the issue of inadequate spatial details in multi-scale building extraction, this paper adopts the U-Net structure as the fundamental framework. The SFE module is employed for accurate feature extraction, while the CIA module captures global information from multi-scale perceptual fields, enabling precise building extraction in high-resolution remote sensing images. The feature fusion is enhanced by adding the AFF mechanism, and the imbalance between foreground and background categories of buildings is addressed using a hybrid loss function. We selected Tumxuk City as the research subject and utilized a UAV equipped with a visible camera to capture high-resolution images and label building targets. This approach results in a diverse building dataset for Tumxuk City, comprising various building types and challenging extraction scenarios. The dataset introduces new requirements for the building extraction network and enriches the diversity of building extraction datasets. The main contributions of this study can be summarized as follows:

(1) We use UAV remote sensing technology to construct the Tmsk building high-resolution remote sensing building extraction dataset, which covers multiple types and scales of buildings;
(2) We propose an effective building extraction model, SCA-Net, that can accurately extract buildings at different scales. We introduced SFE to enhance the feature extraction capability of the network; by introducing CIA, we can improve the ability to detect multi-scale buildings; and applying AFF increases the network's capability to perceive the details of buildings in complex environments;
(3) Our network, SCA-Net, is evaluated on two remote sensing building datasets, demonstrating its robustness and superior accuracy compared to other building extraction methods.

The main chapters of this paper are organized as follows: Section 2 describes the status of the research area of this paper, including the general architecture of the proposed network, the design of the feature extraction module, the design of the contextual information aggregation module, and the design of the Attentional Feature Fusion module, in terms of the main technical methods, data acquisition and dataset construction, the experimental setup, the selection of evaluation metrics, and the introduction of the loss function. Section 3 presents the main experimental results. Section 4 discusses the work of this paper. Section 5 summarizes the work of this paper as well as the outlook for future work.

2. Materials and Methods

2.1. Study Area

In this study, the urban area of Tumxuk, located in southwestern Xinjiang, China, was selected as the study area, as shown in Figure 1. It is located at the southern foot of the Tianshan Mountains and the northwestern edge of the Tarim Basin, with a warm-temperate continental arid climate and a total area of 3664 km^2. Located in the border zone between China and Central Asia, there are some Central Asian influences in the architectural style, and the buildings are usually in the form of flat-roofed, square, or rectangular buildings. With the development of the times, Tumxuk has also seen many modern buildings appear, especially in the commercial and downtown areas.

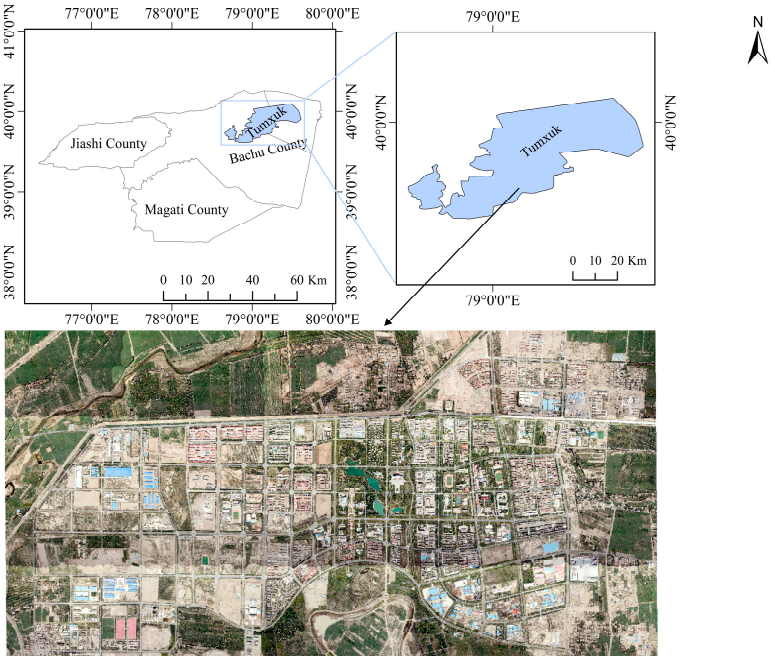

Figure 1. Study area. A location map of the study area and a schematic diagram of some types of buildings.

In summary, the architectural styles in the study area vary significantly, with a variety of geometric attributes such as building shapes, sizes, and structures, as shown in Figure 2. Specifically including high-rise buildings, mid-rise regular residential areas, mid-rise irregular residential areas, low-rise scattered houses, and low-rise continuous houses, as well as the existence of a significant proportion of self-built houses such as bungalows and buildings in scale, due to the local climate and building habits, some buildings have spectral characteristics that are highly similar to the background (roads and land), and some houses are partially covered by vegetation.

Compared with the WHU building dataset in Figure 2d, the building environment in the Tmsk building dataset is more complex, with some buildings being shaded by trees more, as shown by the red circles in Figure 2. Many of the building roofs are chosen to be made of materials such as cement boards, which are similar to the roads, they are difficult to distinguish between foreground and background. They fit more into the Chinese architectural style, as shown by the red boxes in Figure 2. Therefore, the test area selected in this study is representative and enriches the diversity of building extraction datasets in China, which is more difficult for the building extraction algorithm and has higher accuracy, generalization, and usability requirements.

Figure 2. Examples of typical buildings in the Tumk building dataset and the WHU building dataset: (**a**) Low-rise buildings are scattered, similar in textural character to the background, and heavily obscured by vegetation; (**b**) middle-rise and high-rise buildings have a wide range of size types; (**c**) large building footprint and complex environment with many trucks and sheds in the vicinity; (**d**) WHU building dataset typical building.

2.2. Methodology

2.2.1. Architecture Overview

The proposed network in this paper follows a classical end-to-end structure, comprising the encoder module, contextual information aggregation module, and decoder module, as illustrated in Figure 3. The RSIs containing the building are fed into the encoder, which automatically extracts high-level semantic features using the SFE feature extraction module, which comprises multiple-level cascaded SFE units. The CIA module utilizes the ASPP-HDC module to continuously aggregate semantic information about the building. Subsequently, during upsampling, the decoder module employs the AFF module to merge various levels of building features, generating the final building segmentation map.

The encoder module extracts low-level features using cascaded SFE units as the backbone, and the input data are processed in four stages of repeated convolutional layers. Each group of convolutional layers contains three SFE units, and different levels of low-level features are generated using residual connections between each unit. The convolution step of the first SFE unit in each group is set to 2, the feature map space resolution is reduced by 1/2, and the number of channels doubles. In contrast, the resolution and number of channels are adjusted using 1×1 convolution, as shown in Figure 4a, and the second and third units are schematically shown in Figure 4b.

The CIA module captures multi-scale image contexts and semantic information about architectural features by employing parallel multiple cavity convolutions with appropriate expansion rates. Additionally, in stage 4, it aggregates low-level features with large perceptual fields at different scales.

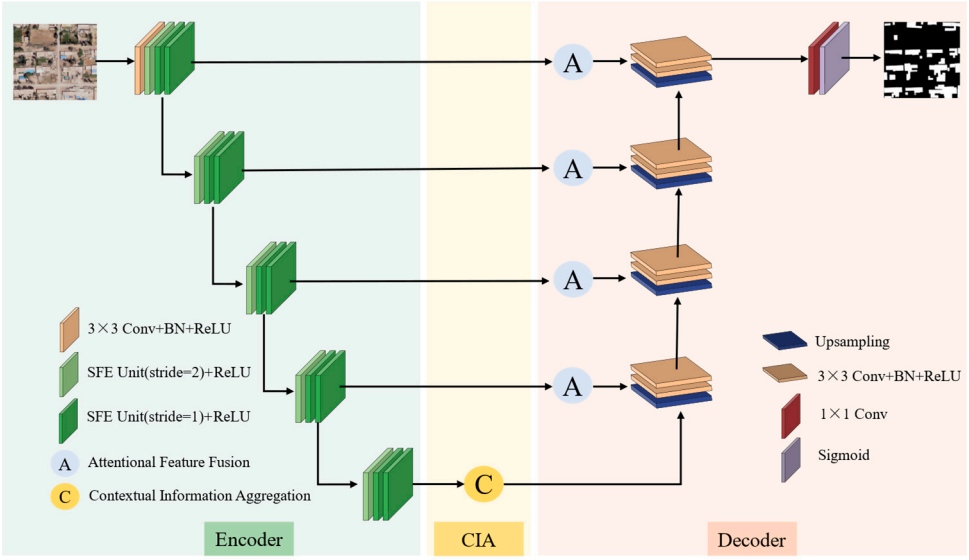

Figure 3. SCA-Net overview architecture.

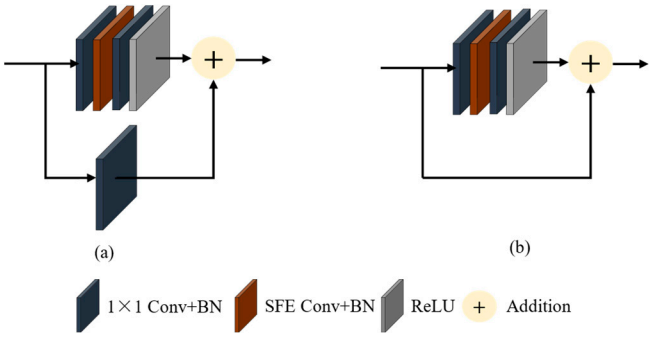

Figure 4. SFE unit architecture: (**a**) denotes an SFE unit with stride = 2; (**b**) denotes an SFE unit with stride = 1.

The decoder module employs bilinear interpolation and 3 × 3 convolution to restore the feature map resolution. Despite the increased perceptual field due to down sampling, valuable spatial information is lost, making it challenging to fully recover detailed global semantic information through up sampling and standard convolution operations. So we utilize the AFF module to address the semantic gap between low-level and high-level features during the jump connection operation. This approach reduces noisy information interference and enhances the utilization of relevant feature information. We apply the AFF module to fuse low-level features from stages 1, 2, and 3 with high-level features after upsampling, thereby recovering spatial information and highlighting building-related details such as space, shape, and edge features while suppressing irrelevant backgrounds like roads, trees, and farmland. The decoder restores the feature map to the original image size by four times upsampling and finally outputs the building extraction results. To prevent overfitting and improve the training speed, dropout [38] and batch normalized (BN) [39] are applied after each convolution operation, respectively.

2.2.2. Selective Kernel Spatial Feature Extraction

Buildings have complex natural attributes and backgrounds, such as roofs with various color, size, and shape features, and standard convolutional operations use a fixed perceptual field while focusing on neighboring pixels, which cannot accurately obtain multiple scale features and pixel distributions and provide a limited exploration of overall spatial and channel relationships. Our study proposes an SFE convolution inspired by [40], as shown in

Figure 5. The SFE convolution aims to extract building features while exploring the spatial distribution patterns of pixels and the relationships between channels to highlight building feature representations. The SFE convolution comprises three main parts: separation, fusion, and selection. SFE convolution facilitates adaptive learning of feature expressions in channel and spatial dimensions, allowing different neurons in the model to acquire channel and spatial weights for each feature through three key steps.

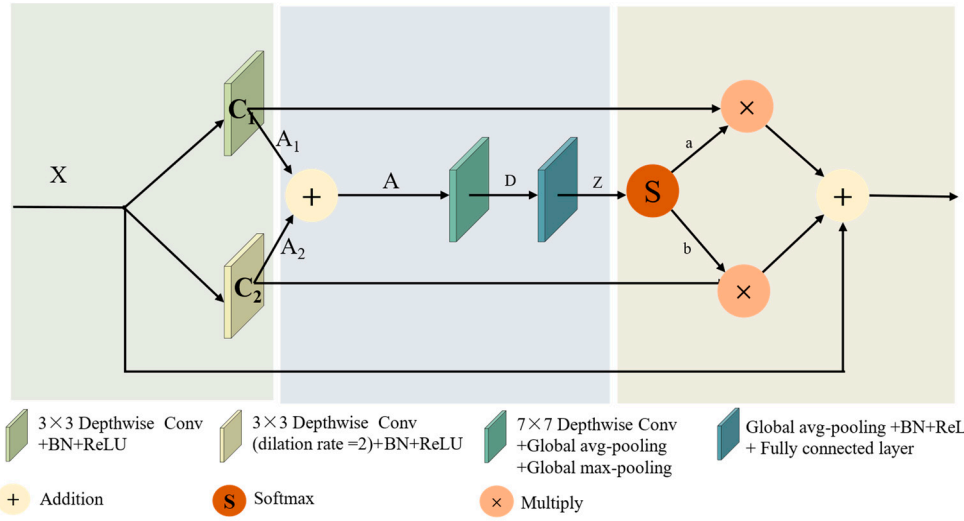

Figure 5. The architecture of the selective kernel spatial feature extraction module.

1. Separation: for any given feature map $X \in \mathbb{R}^{H\times W\times C}$, first two transformations with kernel sizes 3 and 5 $C_1 : X \to A_1 \in \mathbb{R}^{H\times W\times C}$ and $C_2 : X \to A_2 \in \mathbb{R}^{H\times W\times C}$. where both C_1 and C_2 are composed of deep group convolution, batch normalization, and ReLU [41] activation function sequences. To further improve the efficiency, the original 5×5 convolution kernel is replaced by an inflated convolution of size 3×3 and a dilation rate of 2;
2. Fusion: First, the information from different branches is integrated, and the feature maps A_1 and A_2 obtained through different-sized sensory fields are summed element by element;

The fused features $A \in \mathbb{R}^{H\times W\times C}$ undergo averaging pooling over channel dimensions and global maximum pooling to optimize the spatial distribution information of each feature, resulting in pooling results a1 and b1, respectively. a1 and b1 are fused for a 7×7 convolution operation to obtain the feature map F. Feature map F is obtained by the sigmoid [42] activation function to generate the spatial attention feature map S. S can effectively highlight the distribution of feature points. Then, feature map S is dotted with feature map A to obtain feature map $D \in \mathbb{R}^{H\times W\times C}$. The calculation procedure is as follows:

$$D = \sigma(f_{7\times 7}([\text{AvgPool}(A); \text{MaxPool}(A)])) \cdot A \tag{1}$$

where σ denotes the sigmoid activation function and $f_{7\times 7}$ denotes the 7×7 convolution operation.

After that, the spatial dimensionality of compressed feature map $D \in \mathbb{R}^{H\times W\times C}$ is compressed using global average pooling, embedding global information to generate the channel vector $E \in \mathbb{R}^{C}$. The reduced dimensionality features are constructed by a simple fully connected layer $Z \in \mathbb{R}^{d\times 1}$, which guides the adaptive selection of features while reducing the dimensionality to achieve better efficiency. The computational procedure is as follows:

$$Z = C_{fc}(E) = \delta(\mathcal{B}(C_{gp}(D))) \tag{2}$$

where δ denotes the ReLU activation function and $\mathcal{B}$ denotes the regularization.

3. Selection: The features after dimensionality reduction are selected adaptively at different spatial scales using the channel attention mechanism and convolved with the convolution kernels $A, B \in \mathbb{R}^{C \times d}$, respectively, and then processed by Softmax to obtain the channel attention information corresponding to each convolution kernel. The computation process is as follows:

$$a = \frac{e^{AZ}}{e^{AZ} + e^{BZ}}, b = \frac{e^{BZ}}{e^{AZ} + e^{BZ}} \tag{3}$$

where a, b denote the attention vectors corresponding to A_1 and A_2, respectively.

The final feature map $Y \in \mathbb{R}^{H \times W \times C}$ is obtained by multiplying a, b, A_1, and A_2 channel by channel and adding them element by element with original feature map X. The computational procedure is as follows:

$$Y = a \cdot A_1 + b \cdot A_2 + X \tag{4}$$

The SFE unit is constructed using SFE convolution and residual structure. The number of channels is first reduced using a 1×1 convolution kernel. Then, feature extraction is performed using SFE convolution. Finally, the number of channels is recovered using a 1×1 convolution kernel, and the feature maps are output by element-by-element summation of the input features connected by residuals.

2.2.3. Contextual Information Aggregation

After the feature extraction in the encoding stage, the Atrous Spatial Pyramid Pooling module (ASPP) with an expansion rate set by the design principles of HDC [43] is introduced. ASPP-HDC combines dilation convolution and spatial pyramid pooling, utilizing multiple dilation convolutions with different expansion rates in parallel. This approach increases the perceptual field of the convolution kernel without adding parameters, enabling the network to preserve more image features and capture multi-scale contextual information for more accurate segmentation. The scale variability among buildings is large, containing small buildings with rich local details and larger targets such as public places and factories. Therefore, this study uses ASPP-HDC to extract multi-scale features of buildings.

The ASPP-HDC module used in this paper consists of six branches, including a 1×1 convolutional kernel and four parallel 3×3 convolutional kernels with different expansion rates. The expansion rates are set to 2, 4, 8, and 16 to prevent the influence of the gridding effect on the building extraction while satisfying the acquisition of multi-scale information above. After the global mean pooling of the input feature map, the number of feature map channels is adjusted using a 1×1 convolution kernel and upsampling to the same size as the other five branches using bilinear interpolation. Then the multi-scale feature maps of the fused six branches are output through a 1×1 convolution kernel. The ASPP-HDC is applied to the feature maps generated in the encoder section. The generated feature maps are sent to the decoder section, as shown in Figure 6.

2.2.4. Attentional Feature Fusion

Most semantic segmentation networks directly use channel cascading or pixel summation for low-level feature and high-level feature fusion, which may result in the network failing to learn adequate complementary information among cross-level features or even having noisy and redundant information. Low-level features tend to contain more details and local information during feature fusion, while high-level features have higher semantic information. In this paper, the introduction of AFF allows the model to dynamically adjust the level of attention given to high- and low-level features. Consequently, the attention module enables the model to focus more on detailed information in the low-level features and seamlessly integrate them with the high-level features, effectively bridging the gap between them. The AFF module is incorporated at the jump connection to fuse low-level and

high-level features, resulting in the elimination of redundant information and a semantic gap between different features.

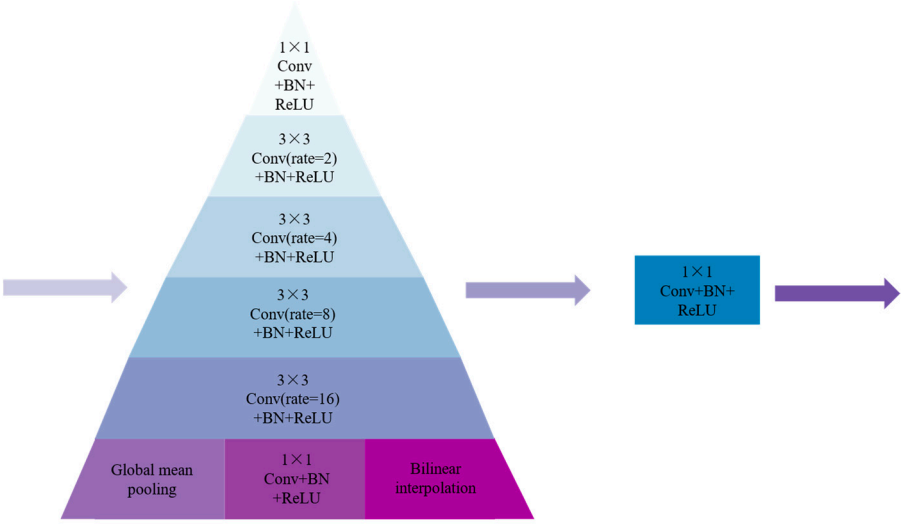

Figure 6. The architecture of ASPP-HDC.

The structure of the AFF module is shown in Figure 7. The inputs to the AFF module are the low-level features from the encoder $G \in \mathbb{R}^{H \times W \times C_1}$ and the high-level features from the decoder sampled by the decoder $X \in \mathbb{R}^{H \times W \times C_2}$, use 1×1 convolution kernel to adjust the channel number of high-level features and low-level features; perform element-by-element summing to generate fusion features; apply global average pooling operation and 3×3 convolution kernel to generate features $I \in \mathbb{R}^{H \times W}$ with channel number 1; use Sigmoid function to generate attention weight coefficients in the range of 0–1, where the closer the value is to 1, the more valuable the feature is. The closer the value is to 1 means, the more valuable the feature is, and finally, the generated weight coefficient vector is multiplied by the feature $X \in \mathbb{R}^{H \times W \times C_2}$ to obtain the final feature map $H \in \mathbb{R}^{H \times W \times C}$. The calculation process is as follows:

$$I = f_{3\times3}(p(f_{1\times1}(G);\ f_{1\times1}(X))) \tag{5}$$

$$H = \sigma(I) \cdot X \tag{6}$$

where f denotes the convolution operation, p denotes the global average pooling operation, and σ denotes the sigmoid function.

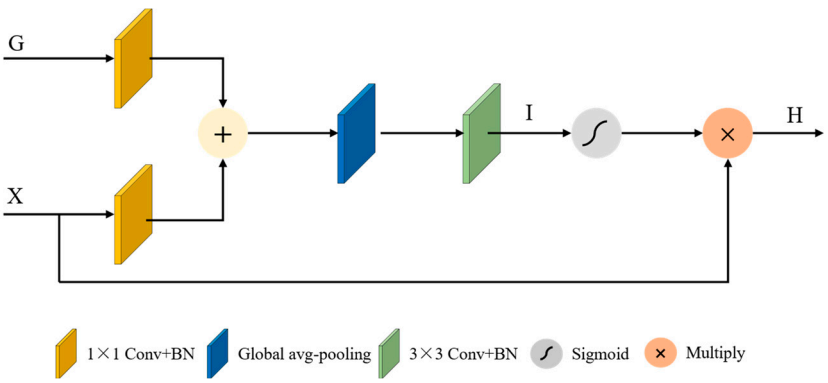

Figure 7. The architecture of Attentional Feature Fusion.

2.3. Data Acquisition and Dataset Construction

The mission used CW-20 composite with a UAV with a SONY-A7RII camera to obtain visible images with 0.09 m ground resolution, 700 m relative flight height, 1791 m route elevation, and 75% overlap, covering the whole Tumushuk urban area. The obtained raw data were stitched and radiometrically corrected, and 2206 images of 2048 × 2048 were generated using the sliding window method. The buildings were labeled by manual decoding, and the building area of less than 5% was removed. The final Tmsk building dataset is 1323, randomly divided into 1058 training sets, 132 validation sets, and 133 test sets. To avoid overfitting the network, the data are enhanced by random inversion, image noise addition, and increasing image brightness.

To validate the generalization performance of the network, we use an aerial image dataset from the WHU Aerial Building Imagery dataset, which is constructed from Christchurch, New Zealand, covering an area of 450 square kilometers. It contains 8189 images of 512 × 512 pixels with a ground resolution of 0.3 m. The training set contains 4736 images, and the validation and test sets contain 1036 and 2417 images, respectively.

2.4. Implementation Setting and Evaluation Indicators

All experiments in this paper are based on a Linux system equipped with 24 GB of video memory, a NVIDIA GeForce RTX 3090, implemented using Python 3.9 and the PyTorch deep learning framework, using the Adam optimizer [44], with an initial learning rate of 0.0001, a sample size of 4 selected for each training, and 100 iterations for all experimental networks.

To objectively evaluate the performance of the proposed network, four metrics commonly used for semantic segmentation are used in this paper, including Precision, Recall, F1-score, and Intersection-over-Union (IoU), with the following equations:

$$\text{Precision} = \frac{\text{TP}}{\text{TF} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{F1} - \text{score} = \frac{2 \times \text{ Precision } \times \text{Recall}}{\text{Precision } + \text{Recal}} \tag{9}$$

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

where TP is the number of actual building pixels classified as building pixels; FN is the number of actual building pixels classified as background pixels; FP is the number of background pixels classified as building pixels; and TN is the number of background pixels classified as background pixels.

2.5. Loss Function

Since the pixel regions of buildings in remote sensing images vary in size, especially the extremely high proportion of background pixels, this usually leads to a more background-biased prediction of the network, which results in poor segmentation accuracy for buildings. Therefore, this paper uses dice loss [45] and focal loss [46] to calculate the difference between the actual and predicted values to train the model. Dice loss is a region-based loss that calculates the entire image region and focuses more on mining the foreground, but the training is quickly unstable. Focal loss is a variation of cross-entropy loss based on cross-entropy loss. The focal loss is a variation of cross-entropy loss that introduces adjustable factors on top of cross-entropy loss to reduce the weight of easily classified samples and increase the weight of hard-to-classify samples, focusing on pixel-level loss calculation. The fusion of the two loss functions helps the model learn the foreground knowledge better and accelerates the convergence of the model simultaneously.

Dice loss is defined as:

$$L_d = 1 - \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \tag{11}$$

where p_i denotes the predicted result of pixel i, g_i denotes the true labeling result of pixel i, and N denotes the number of all pixels.

Focal loss is defined as:

$$L_f = -\alpha(1 - p_t)^\gamma \log(p_t) \tag{12}$$

$$\alpha_i = 1 - r_i \tag{13}$$

where the closer p_t is to 1, the more accurate the pixel prediction is, the easier it is to distinguish, and the less weight it takes up, allowing the network to focus on hard-to-classify pixels. r_i is the percentage of pixels in each category, α_i is the weight of each category, γ is the parameter controlling the degree of weight, and γ = 2 is taken in this paper.

Therefore, the total loss function L_b in this paper is defined as:

$$L_b = L_f + L_d \tag{14}$$

where L_f is the focal loss function and L_d is the dice loss function.

3. Results

3.1. Comparative Experimental Results on the Tmsk Building

To verify the effectiveness of the proposed SCA-Net, building extraction comparison experiments are conducted with several mainstream semantic segmentation networks on the Tmsk Building dataset, including U-Net, Segnet, Deeplab v3+, HRNet v2, and SegFormer-B2 [47]. Building extraction results are shown in Figure 8.

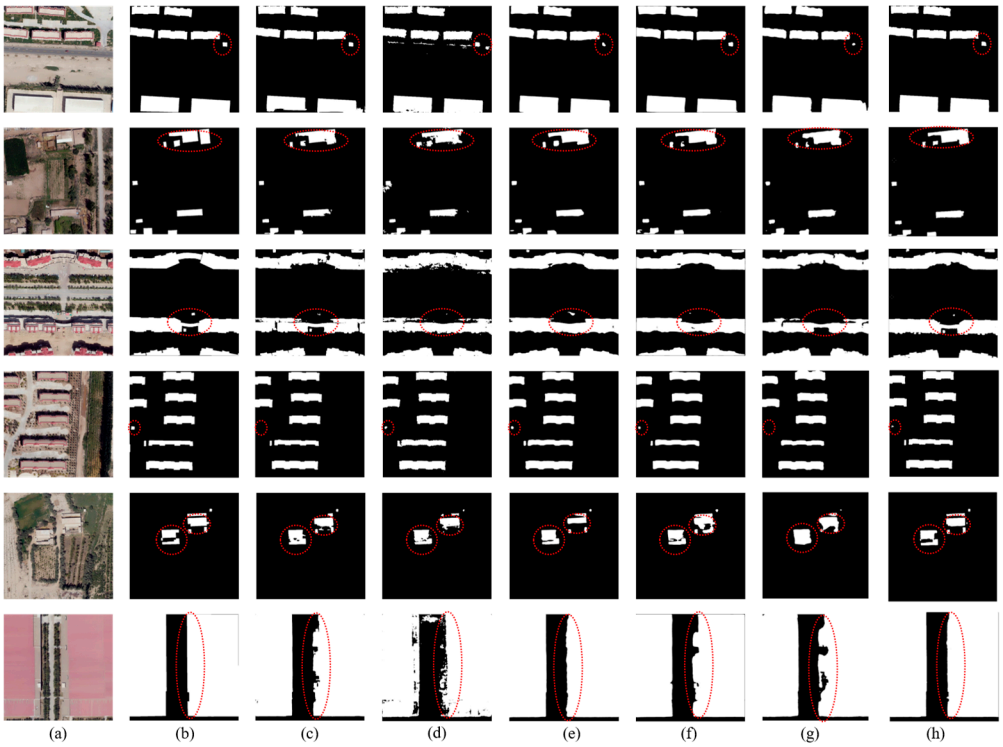

Figure 8. Visualization of building extraction results on the Tmsk building dataset: (**a**) original images; (**b**) ground truth; (**c**) U-Net; (**d**) Segnet; (**e**) Deeplab v3+; (**f**) HRNet v2; (**g**) SegFormer-B2; (**h**) SCA-Net.

As shown in Figure 8, U-Net, Segnet, Deeplab v3+, and HRNet v2 algorithms extracted most of the buildings, but there are still problems with extraction errors and missing extractions. U-Net and SegNet have a large number of voids due to the lack of multi-scale feature extraction capability, and Deeplab v3+ directly quadruples down sampling the images at the beginning of the network. HRNet v2 uses direct pixel summation or merging on channel dimensions for feature fusion, which loses the details and distinction of the original features. The extracted results cannot focus on the edges of buildings or the shape of the region, resulting in boundary adhesion and false detection. Due to the similar spectral features of buildings and backgrounds in the first row and the second row of red circles, building extraction is tough and challenging. SegFormer-B2 can extract the shape information of buildings better, but the extraction results may occasionally show noise and blurred boundary information.

The method proposed in this paper has a more complete extraction of small buildings, sharper building edges, and can better overcome the interference of similar spectral features. From the red circles in the fourth and sixth rows, we can see that for large-scale buildings, the extraction results of other methods have serious voids and building edge error phenomena. However, the method proposed in this paper solves the above problems by using the CIA module to extract multi-scale contextual information about buildings, which helps the network better understand building boundaries, shapes, and details. In other challenging building scenarios, such as building shadows (third row) and complex country courtyard (fifth row) buildings, other methods suffer from incomplete extraction results and inaccurate building outer boundary locations. The SCA-Net proposed in this paper has the best extraction results through SFE and AFF, retaining more feature details while suppressing the expression of redundant noise, integrating multiple scale of contextual information, and accomplishing effective fusion of cross-layer information.

The quantitative analysis of the segmentation results of different algorithms on the Tmsk Building dataset is shown in Table 1, and the metrics used for the quantitative analysis are Precision, Recall, IoU, and F1-Score. Table 1 shows that the proposed SCA-Net algorithm has 93.89% Precision, 90.95% Recall, 85.98% IoU, and 92.40% F1-Score on the Tmsk Building dataset, achieving an outstanding level in all assessment indicators. Rows 1~5 show the semantic segmentation quantifiers of U-Net, Segnet, Deeplab v3+, HRNet v2, and SegFormer-B2 comparison algorithms. Compared with the U-Net, Segnet, Deeplab v3+, HRNet v2, and SegFormer-B2 comparison algorithms, the precision of the comparison metrics improved by 1.46%, 4.66%, 1.33%, 1.32%, and 0.68%, respectively; the recall of the metrics improved by 4.46%, 8.48%, 4.58%, 2.12%, and 1.08%, respectively; the IoU improved by 3.69%, 10.09%, 4.27%, 3.67%, and 0.63%, respectively; the F1-Score improved by 3.05%, 6.68%, 3.04%, 1.72%, and 0.89%, respectively.

Table 1. Quantitative comparison of Precision, Recall, F1-score, and IoU in the Tmsk building dataset.

Method	Precision (%)	Recall (%)	IoU (%)	F1-Score (%)
U-Net	92.43	86.49	82.29	89.35
Segnet	89.23	82.47	75.89	85.72
Deeplab v3+	92.56	86.37	81.71	89.36
HRNet v2	92.57	88.83	82.31	90.68
SegFormer-B2	93.21	89.87	85.35	91.51
SCA-Net	93.89	90.95	85.98	92.40

In summary, it can be seen from the segmentation indexes and comparison algorithms that the SCA-Net algorithm proposed in this paper has a better segmentation effect in the Tmsk Building dataset and can perform higher quality segmentation in remote sensing building images in various complex scenes.

3.2. Comparative Experimental Results on the WHU Building

To verify the generalization of the proposed SCA-Net, a building segmentation comparison experiment is conducted with several mainstream semantic segmentation networks on the building segmentation public dataset WHU Building, including U-Net, Segnet, Deeplab v3+, HRNet v2, and SegFormer-B2. The segmentation results are shown in Figure 9.

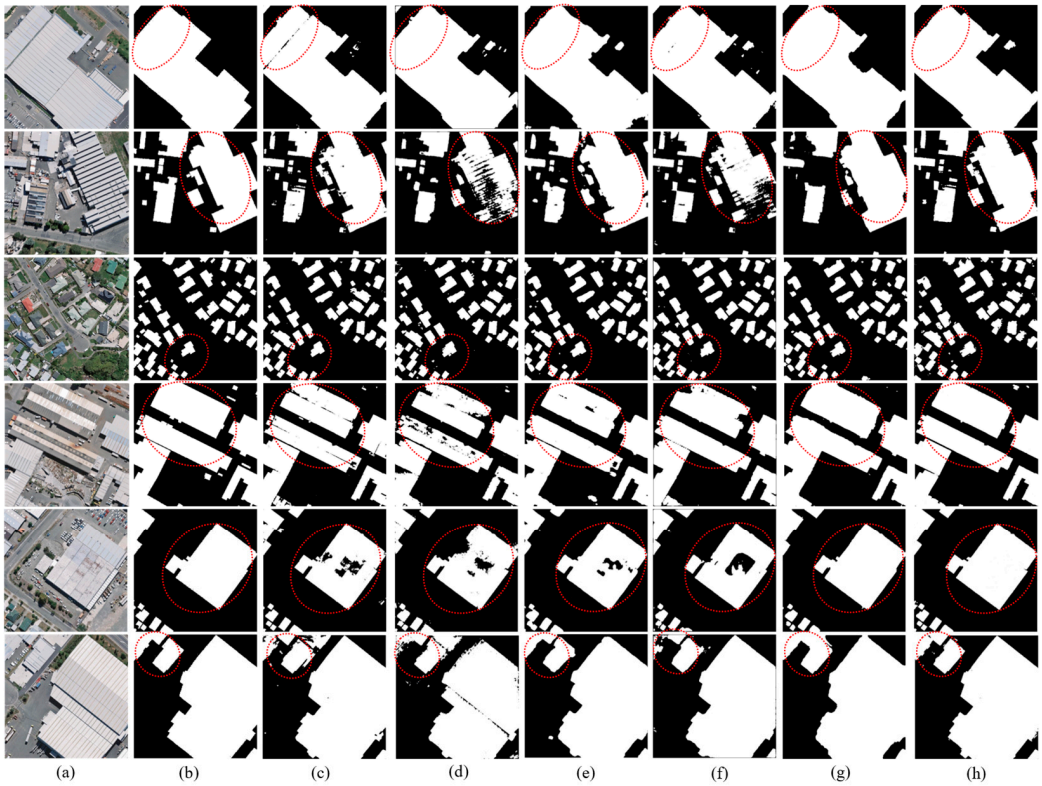

Figure 9. Visualization of building extraction results on the WHU building dataset: (**a**) original images; (**b**) ground truth; (**c**) U-Net; (**d**) Segnet; (**e**) Deeplab v3+; (**f**) HRNet v2; (**g**) SegFormer-B2; (**h**) SCA-Net.

From Figure 9, it can be seen that the SCA-Net algorithm proposed in this paper can better distinguish the background and foreground, which ensures the visualization of the extraction results in a variety of situations, such as tree shadow occlusion, dense building clusters, more significant buildings, small buildings, complex backgrounds, etc. The extraction results of the U-Net, Segnet, Deeplab v3+, HRNet v2, and SegFormer-B2 algorithms. The edge integrity of the U-Net, Segnet, Deeplab v3+, HRNet v2, and SegFormer-B2 algorithms is poor, as shown by the red circle in Figure 9. In addition, the U-Net, Segnet, Deeplab v3+, and HRNet v2 algorithms also have large building extraction results in the void due to the insufficient feature extraction capability of the encoder part and the insufficient sensory field size, as shown in Figure 9. The Segnet algorithm also has the phenomenon of omission in the extraction of the small buildings as well as the phenomenon of incomplete extraction due to the lack of feature information in the lower layers.

As shown in Figure 9, the algorithm proposed in this paper benefits from the introduction of SFE to better extract the feature information of the building, the addition of the CIA module to obtain the contextual information of the larger and more scaled buildings, and the introduction of AFF to better guide the feature fusion at the same time to eliminate the interference of redundant noise, so the SCA-Net algorithm can effectively retain the boundary information of the building in complex environments. The extraction results are precise contours, and the extraction of large buildings is incomplete. The result is a

clear outline; the extraction of large buildings is complete, and small buildings are also effectively extracted.

The results of the quantitative analysis of the segmentation results of different algorithms on the WHU Building dataset are shown in Table 2. The metrics used in the quantitative analysis are Precision, Recall, IoU, and F1-Score. As shown in Table 2, the proposed SCA-Net algorithm has 95.18% Precision, 92.59% Recall, 89.90% IoU, and 93.87% F1-Score on the WHU Building dataset. Rows 1~7 show the semantic segmentation quantifiers of U-Net, Segnet, Deeplab v3+, HRNet v2, SST, MSST-Net, and SegFormer-B2 comparison algorithms. Compared with the U-Net, Segnet, Deeplab v3+, HRNet v2, SST, MSST-Net, and SegFormer-B2 comparison algorithms, the comparison metrics Precision improved by 2.13%, 5.34%, 2.73%, 1.36%, and −1.06%, respectively; the metrics Recall improved by 2.53%, 5.81%, 3.62%, 0.61%, and 1.23%, respectively; the IoU improved by 2.51%, 6.87%, 3.31%, 1.65%, −0.58%, 1.9%, and −0.13%, and the F1-Score improved by 2.23%, 5.59%, 3.19%, 0.98%, −1.1%, 5.67%, and 0.13%, respectively. It is again verified that the proposed method in this paper is robust enough to handle buildings in multiple scenarios with strong robustness. Although the SCA-Net proposed in this paper is slightly lower than the SST (R18, S4) algorithm in terms of evaluation metrics, the model complexity of SST (R18, S4) is the highest among the SST family of algorithms.

Table 2. Quantitative comparison of Precision, Recall, F1-score, and IoU in the WHU building dataset.

Method	Precision (%)	Recall (%)	IoU (%)	F1-Score (%)
U-Net	93.05	90.06	87.39	91.53
Segnet	89.84	86.78	83.03	88.28
Deeplab v3+	92.45	88.97	86.59	90.68
HRNet v2	93.82	91.98	88.25	92.89
SST	-	-	90.48	94.97
MSST-Net	-	-	88.00	88.20
SegFormer-B2	96.24	91.36	90.03	93.74
SCA-Net	95.18	92.59	89.90	93.87

In summary, it can be seen from the segmentation metrics and comparison algorithms that the proposed SCA-Net algorithm has a better segmentation effect in the WHU Building dataset, and it is also verified that SCA-Net has good extraction ability and generalization performance in different scenarios.

3.3. Ablation Study

To verify the effectiveness of each module of SCA-Net, ablation experiments are conducted on the Tmsk Building dataset based on SCA-Net. Using the semantic segmentation model U-Net as the baseline model (B), we successively added Selective kernel spatial Feature Extraction (SFE), Contextual Information Aggregation (CIA), and Attentional Feature Fusion (AFF), mainly from the segmentation results of Precision, Recall, IoU, and F1-score four indicators for comparison, as shown in Table 1. The segmentation results are shown in Figure 10.

After adding AFF based on baseline, the network pays better attention to the target region, suppresses the image of redundant information on segmentation results, and effectively solves the problem of misclassification in noisy backgrounds, as shown in the red circle in the first row in Figure 10. The use of SFE for feature extraction significantly improves the completeness of the network for small building extraction. It solves the problem of mis-segmentation for courtyard buildings with similar background colors, as shown in the red circle in the second row of Figure 10. The complete extraction of large buildings is achieved by using cavity convolution with a more significant cavity rate to reduce the cavity phenomenon of large building extraction results, as shown in the red circle in the third row of Figure 10. The SCA-Net proposed in this paper realizes the accurate extraction of multi-scale buildings against complex backgrounds.

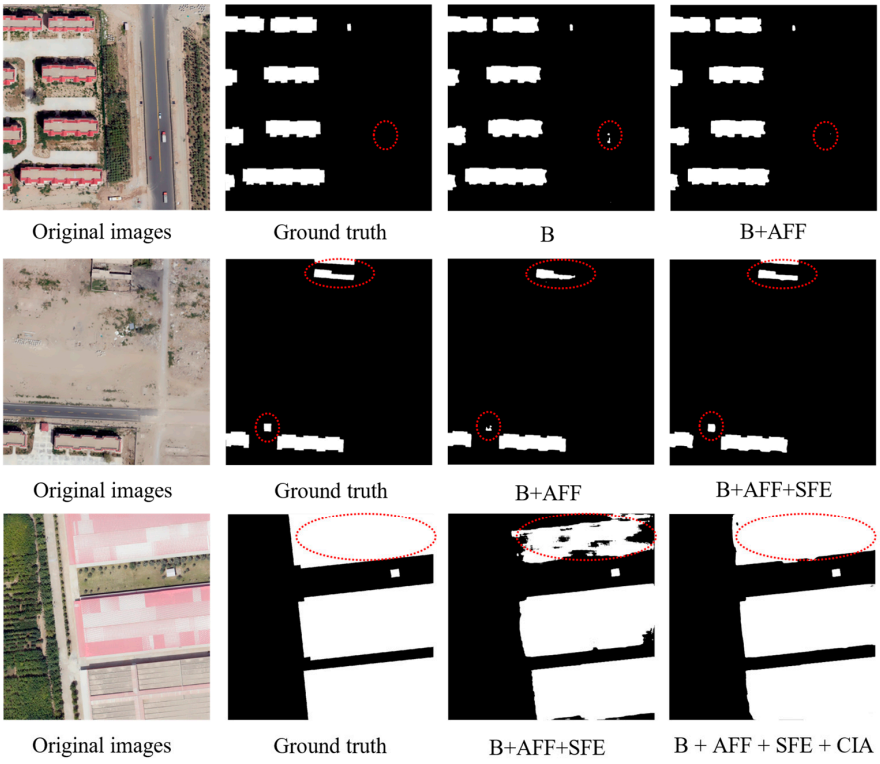

Figure 10. Results of the ablation experimental building extraction.

As shown in Figure 11, after Baseline introduces the AFF module, Precision, Recall, IoU, and F1-score are improved by 0.74%, 2.53%, 1.28%, and 1.69%, respectively, which proves that the Attention Fusion Module removes the disparity between the high- and low-level features during the feature fusion process, makes full use of the low-level features featuring diverse spatial information, and guides the high-level features to focus on the spatial location of the foreground. Precision, Recall, IoU, and F1-score improved after using the SFE module as an encoder by 0.89%, 4.16%, 2.18%, and 2.61%, respectively, which verified the feature extraction ability of the SFE conv. After adding the CIA module, Baseline obtains an improvement of 1.46%, 4.46%, 2.69%, and 3.04% in Precision, Recall, IoU, and F1-score, respectively, demonstrating the usefulness of the HDC-compliant nulling-rate ASPP design for feature capture and global feature integration of buildings at various scales.

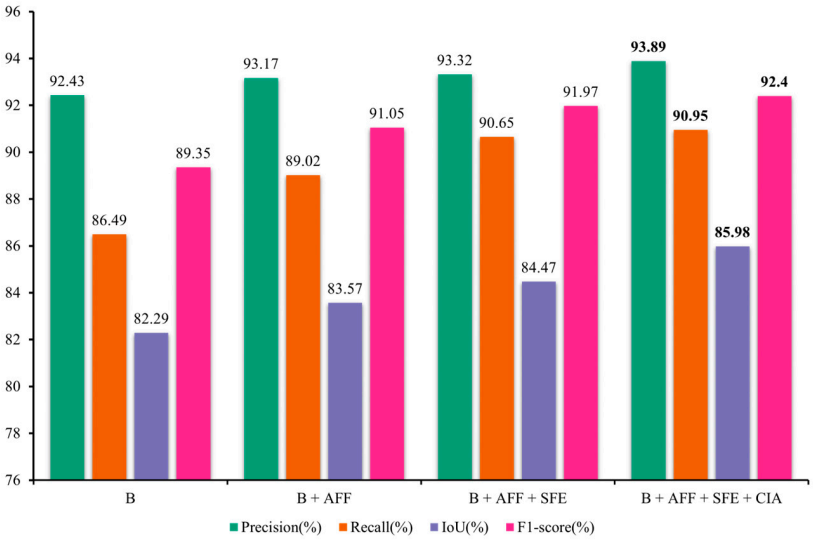

Figure 11. Evaluation results of the ablation experiments.

4. Discussion

Buildings are one of the most important man-made features in remote sensing imagery, and researchers have proposed various methods to accurately extract building information. Huang et al. [48] proposed to use the Morphological Building Index (MBI) method to extract buildings in high-resolution imagery, and then Yuan [49] used a CNN network-based method to apply it to the accurate extraction of buildings.

In the building extraction task, the different materials (concrete, asphalt, and metal) between the buildings lead to large differences in the spectral features, but some of the buildings are extremely similar to the spectral features of the roads in the background, which puts forward strict requirements on the feature extraction method. Deepening the network depth and expanding the scale can have a certain effect, but there will be high network complexity and overfitting phenomena. This paper proposes that the SFE conv module uses the attention mechanism at the same time as the feature extraction so that the network has a good generalization performance while focusing on the building itself during the feature extraction.

The size difference between the buildings is huge. For small buildings, the low-level features are very important as the target itself has a limited width of only a few pixels, and the spatial location information of the target becomes unstable after multiple down sampling, so the spatial location information of the low-level features is more accurate. The fusion of bottom layer features and high layer features can solve the problem of accurate extraction of small buildings. In this paper, the low-level features are used to guide the high-level features through the AFF module to add more spatial location information while eliminating the interference of redundant information from the low-level features. In addition, the extraction of multi-scale buildings relies on multi-level receptive fields and global information; using hollow convolution to expand the receptive field or using deep-level features can play a role. This paper uses the CIA module in the deepest part of the deepest network to realize the acquisition of multi-scale receptive fields and global information, which plays an important role in the network.

Although the method proposed in this paper obtains good results in terms of usability and robustness, it still has some limitations. For example, in the extraction of low-rise bungalows and relatively old buildings in the suburbs of the city, although the extraction method in this paper is indeed better than other methods, the visualization effect is still somewhat different from that of high-rise buildings and other buildings with distinctive features, and the roofs of neighboring buildings are not well recognized as separate individuals. Shadows are an important factor that has to be considered when extracting buildings from remote sensing images. Shadows can have an impact on the shape and location information of a building, which may lead to blurring of the building outline, making it difficult to accurately determine the boundary information of the building, and incorrectly identifying the shadowed area as part of the building. Therefore, in future research, we will study the morphological processing method of the neighbor relationship between buildings, introduce building edge information, use image enhancement techniques to improve the effect of shadows in the impact, and develop the multi-sensor complex building extraction method to further improve the effect of neighboring building extraction.

5. Conclusions

Aiming at the problems of low extraction accuracy and unclear building boundaries of the existing building extraction methods. Firstly, we select the Tumxuk urban area, which has various types of buildings in the town and is challenging to extract, as the study area. Data acquisition is carried out using remote sensing technology to produce the Tmsk Building dataset, which enriches the diversity of the building extraction types and puts forward new challenges to the existing building extraction algorithms. Secondly, the SCA-Net algorithm is proposed, where the encoder part consists of a cascade of SFE units to improve the extraction of building features; the CIA module is used to obtain a multi-scale sensory field for more comprehensive semantic information and to enhance the

completeness of the network for the extraction of multi-sized buildings; and the decoder part uses the AF for the effective fusion of cross-level features to achieve a fine-scale inter-scalar transfer of information, which improves the focus on building targets while eliminating redundant information. Beyond that, we are using an improved loss function to balance the building-to-background ratio to improve the prediction accuracy of building edges and the ability to retain details. Finally, the ablation experiments are conducted on the Tmsk building dataset to validate the effectiveness of each module. In comparison with U-Net, Segnet, Deeplab v3+, HRNet v2, SST, MSST-Net, and SegFormer-B2 comparison algorithms on the Tmsk building dataset and the WHU building dataset, the experimental results show that the algorithm proposed in this paper overcomes the building shadows and tree occlusions, dramatically reduces the phenomenon of omission of extraction for small buildings and extraction of voids for large buildings, and that the extraction effect is outstanding with high extraction accuracy and good generalization performance.

Author Contributions: Y.W. (Yuanzhi Wang): methodology, software, writing, editing, data analysis, result verification. Q.Z.: methodology, supervision, funding acquisition. Y.W. (Yuzhen Wu): methodology, review, draw. W.T.: data analysis, supervision. G.Z.: data analysis, supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China (32260388) and the Xinjiang Production and Construction Corps Key Field Science and Technology Tackling Program Project (2023CB008-22).

Data Availability Statement: The WHU building dataset can be downloaded from http://gpcv.whu.edu.cn/data/building_dataset.html (accessed on 16 March 2022). The Tmsk building dataset can be downloaded from https://doi.org/10.6084/m9.figshare.24063777.

Acknowledgments: The authors appreciate Wuhan University for sharing the building datasets for free.

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

References

1. Güneralp, B.; Zhou, Y.; Ürge-Vorsatz, D.; Gupta, M.; Yu, S.; Patel, P.L.; Fragkias, M.; Li, X.; Seto, K.C. Global Scenarios of Urban Density and Its Impacts on Building Energy Use through 2050. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 8945–8950. [CrossRef] [PubMed]
2. Claassens, J.; Koomen, E.; Rouwendal, J. Urban Density and Spatial Planning: The Unforeseen Impacts of Dutch Devolution. *PLoS ONE* **2020**, *15*, e0240738. [CrossRef] [PubMed]
3. Li, X.; Ying, Y.; Xu, X.; Wang, Y.; Hussain, S.A.; Hong, T.; Wang, W. Identifying Key Determinants for Building Energy Analysis from Urban Building Datasets. *Build. Environ.* **2020**, *181*, 107114. [CrossRef]
4. Yuan, P.; Zhao, Q.; Zhao, X.; Wang, X.; Long, X.; Zheng, Y. A Transformer-Based Siamese Network and an Open Optical Dataset for Semantic Change Detection of Remote Sensing Images. *Int. J. Digit. Earth* **2022**, *15*, 1506–1525. [CrossRef]
5. Rafiei-Sardooi, E.; Azareh, A.; Choubin, B.; Mosavi, A.H.; Clague, J.J. Evaluating Urban Flood Risk Using Hybrid Method of TOPSIS and Machine Learning. *Int. J. Disaster Risk Reduct.* **2021**, *66*, 102614. [CrossRef]
6. Shugar, D.H.; Jacquemart, M.; Shean, D.; Bhushan, S.; Upadhyay, K.; Sattar, A.; Schwanghart, W.; McBride, S.; De Vries, M.V.W.; Mergili, M.; et al. A Massive Rock and Ice Avalanche Caused the 2021 Disaster at Chamoli, Indian Himalaya. *Science* **2021**, *373*, 300–306. [CrossRef]
7. Li, D.; Lu, X.; Walling, D.E. High Mountain Asia Hydropower Systems Threatened by Climate-Driven Landscape Instability. *Nat. Geosci.* **2022**, *15*, 520–530. [CrossRef]
8. 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]
9. Yuan, W.; Wang, J.; Xu, W. Shift Pooling PSPNet: Rethinking Pspnet for Building Extraction in Remote Sensing Images from Entire Local Feature Pooling. *Remote Sens.* **2022**, *14*, 4889. [CrossRef]
10. 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]
11. Ran, S.; Gao, X.; Yang, Y.; Li, S.; Zhang, G.; Wang, P. Building Multi-Feature Fusion Refined Network for Building Extraction from High-Resolution Remote Sensing Images. *Remote Sens.* **2021**, *13*, 2794. [CrossRef]
12. Yuan, X.; Shi, J.; Gu, L. A Review of Deep Learning Methods for Semantic Segmentation of Remote Sensing Imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
13. Dong, S.; Wang, P.; Abbas, K. A Survey on Deep Learning and Its Applications. *Comput. Sci. Rev.* **2021**, *40*, 100379. [CrossRef]

14. Hao, S.; Zhou, Y.; Guo, Y. A Brief Survey on Semantic Segmentation with Deep Learning. *Neurocomputing* **2020**, *406*, 302–321. [CrossRef]
15. Zuo, T.; Feng, J.; Chen, X. HF-FCN: Hierarchically Fused Fully Convolutional Network for Robust Building Extraction. In Proceedings of the Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; Revised Selected Papers, Part I 13. Springer: Berlin/Heidelberg, Germany, 2017; pp. 291–302.
16. Schuegraf, P.; Bittner, K. Automatic Building Footprint Extraction from Multi-Resolution Remote Sensing Images Using a Hybrid FCN. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 191. [CrossRef]
17. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18. Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
18. Hosseinpoor, H.; Samadzadegan, F. Convolutional Neural Network for Building Extraction from High-Resolution Remote Sensing Images. In Proceedings of the 2020 International Conference on Machine Vision and Image Processing (MVIP), Qom, Iran, 18–20 February 2020; pp. 1–5.
19. Si, Z.; Zhou, B.; Wang, B.; Wang, X.; Zhu, L. High-Resolution Remote Sensing Building Extraction Based on Attention Mechanism and DeepLabv3+. In Proceedings of the 5th International Conference on Computer Information Science and Application Technology (CISAT 2022), Chongqing, China, 29–31 July 2022; Volume 12451, pp. 122–126.
20. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
21. Seong, S.; Choi, J. Semantic Segmentation of Urban Buildings Using a High-Resolution Network (HRNet) with Channel and Spatial Attention Gates. *Remote Sens.* **2021**, *13*, 3087. [CrossRef]
22. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.
23. Shi, X.; Huang, H.; Pu, C.; Yang, Y.; Xue, J. CSA-UNet: Channel-Spatial Attention-Based Encoder–Decoder Network for Rural Blue-Roofed Building Extraction From UAV Imagery. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6514405. [CrossRef]
24. Aryal, J.; Neupane, B. Multi-Scale Feature Map Aggregation and Supervised Domain Adaptation of Fully Convolutional Networks for Urban Building Footprint Extraction. *Remote Sens.* **2023**, *15*, 488. [CrossRef]
25. Xu, X.; Zhang, H.; Ran, Y.; Tan, Z. High-Precision Segmentation of Buildings with Small Sample Sizes Based on Transfer Learning and Multi-Scale Fusion. *Remote Sens.* **2023**, *15*, 2436. [CrossRef]
26. Li, M.; Rui, J.; Yang, S.; Liu, Z.; Ren, L.; Ma, L.; Li, Q.; Su, X.; Zuo, X. Method of Building Detection in Optical Remote Sensing Images Based on SegFormer. *Sensors* **2023**, *23*, 1258. [CrossRef] [PubMed]
27. Yuan, W.; Xu, W. MSST-Net: A Multi-Scale Adaptive Network for Building Extraction from Remote Sensing Images Based on Swin Transformer. *Remote Sens.* **2021**, *13*, 4743. [CrossRef]
28. Chen, K.; Zou, Z.; Shi, Z. Building Extraction from Remote Sensing Images with Sparse Token Transformers. *Remote Sens.* **2021**, *13*, 4441. [CrossRef]
29. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y. Segment Anything. *arXiv* **2023**, arXiv:2304.02643.
30. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z. RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model. *arXiv* **2023**, arXiv:2306.16269.
31. Chen, X.; Qiu, C.; Guo, W.; Yu, A.; Tong, X.; Schmitt, M. Multiscale Feature Learning by Transformer for Building Extraction from Satellite Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 2503605. [CrossRef]
32. Hu, Q.; Zhen, L.; Mao, Y.; Zhou, X.; Zhou, G. Automated Building Extraction Using Satellite Remote Sensing Imagery. *Autom. Constr.* **2021**, *123*, 103509. [CrossRef]
33. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586. [CrossRef]
34. 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.
35. Wang, Y.; Zeng, X.; Liao, X.; Zhuang, D. B-FGC-Net: A Building Extraction Network from High Resolution Remote Sensing Imagery. *Remote Sens.* **2022**, *14*, 269. [CrossRef]
36. Chen, M.; Wu, J.; Liu, L.; Zhao, W.; Tian, F.; Shen, Q.; Zhao, B.; Du, R. DR-Net: An Improved Network for Building Extraction from High Resolution Remote Sensing Image. *Remote Sens.* **2021**, *13*, 294. [CrossRef]
37. Chen, Z.; Li, D.; Fan, W.; Guan, H.; Wang, C.; Li, J. Self-Attention in Reconstruction Bias U-Net for Semantic Segmentation of Building Rooftops in Optical Remote Sensing Images. *Remote Sens.* **2021**, *13*, 2524. [CrossRef]
38. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
39. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pmlr; pp. 448–456.

40. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
41. Agarap, A.F. Deep Learning Using Rectified Linear Units (Relu). *arXiv* **2018**, arXiv:1803.08375.
42. Han, J.; Moraga, C. The Influence of the Sigmoid Function Parameters on the Speed of Backpropagation Learning. In *Proceedings of the International Workshop on Artificial Neural Networks*; Springer: Berlin/Heidelberg, Germany, 1995; pp. 195–201.
43. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding Convolution for Semantic Segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1451–1460.
44. Kinga, D.; Adam, J.B. A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; Volume 5, p. 6.
45. Abdollahi, A.; Pradhan, B.; Alamri, A. VNet: An End-to-End Fully Convolutional Neural Network for Road Extraction from High-Resolution Remote Sensing Data. *IEEE Access* **2020**, *8*, 179424–179436. [CrossRef]
46. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
47. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Advances in Neural Information Processing Systems, Curran Associates, Inc., virtual, 6–14 December 2021; Volume 34, pp. 12077–12090.
48. Huang, X.; Zhang, L. Morphological Building/Shadow Index for Building Extraction From High-Resolution Imagery Over Urban Areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 161–172. [CrossRef]
49. Yuan, J. Learning Building Extraction in Aerial Scenes with Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2793–2798. [CrossRef] [PubMed]