remote sensing

MDPI

Article

RDAU-Net: A U-Shaped Semantic Segmentation Network for Buildings near Rivers and Lakes Based on a Fusion Approach

Yipeng Wang [1,2], Dongmei Wang [2,*], Teng Xu [1], Yifan Shi [2], Wenguang Liang [2], Yihong Wang [2], George P. Petropoulos [3] and Yansong Bao [4]

[1] College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China; 221602010081@hhu.edu.cn (Y.W.); teng.xu@hhu.edu.cn (T.X.)
[2] Jiangsu Hydraulic Research Institute, Nanjing 210017, China; 20171103081@nuist.edu.cn (Y.S.); liangwg04@163.com (W.L.); njnuyhwang@163.com (Y.W.)
[3] Department of Geography, Harokopio University of Athens, EI. Venizelou 70, 17671 Athens, Greece; gpetropoulos@hua.gr
[4] School of Atmospheric Physics, Nanjing University of Information Science and Technology, Nanjing 210044, China; ysbao@nuist.edu.cn
* Correspondence: skywdm2024@163.com

Abstract: The encroachment of buildings into the waters of rivers and lakes can lead to increased safety hazards, but current semantic segmentation algorithms have difficulty accurately segmenting buildings in such environments. The specular reflection of the water and boats with similar features to the buildings in the environment can greatly affect the performance of the algorithm. Effectively eliminating their influence on the model and further improving the segmentation accuracy of buildings near water will be of great help to the management of river and lake waters. To address the above issues, the present study proposes the design of a U-shaped segmentation network of buildings called RDAU-Net that works through extraction and fuses a convolutional neural network and a transformer to segment buildings. First, we designed a residual dynamic short-cut down-sampling (RDSC) module to minimize the interference of complex building shapes and building scale differences on the segmentation results; second, we reduced the semantic and resolution gaps between multi-scale features using a multi-channel cross fusion transformer module (MCCT); finally, a double-feature channel-wise fusion attention (DCF) was designed to improve the model's ability to depict building edge details and to reduce the influence of similar features on the model. Additionally, an HRI Building dataset was constructed, comprising water-edge buildings situated in a riverine and lacustrine regulatory context. This dataset encompasses a plethora of water-edge building sample scenarios, offering a comprehensive representation of the subject matter. The experimental results indicated that the statistical metrics achieved by RDAU-Net using the HRI and WHU Building datasets are better than those of others, and that it can effectively solve the building segmentation problems in the management of river and lake waters.

Keywords: semantic segmentation; remote sensing; buildings extraction; convolutional neural network; transformer; RDAU-Net

check for updates

Academic Editor: Javier Marcello

Received: 26 September 2024
Revised: 26 November 2024
Accepted: 20 December 2024
Published: 24 December 2024

Citation: Wang, Y.; Wang, D.; Xu, T.; Shi, Y.; Liang, W.; Wang, Y.; Petropoulos, G.P.; Bao, Y. RDAU-Net: A U-Shaped Semantic Segmentation Network for Buildings near Rivers and Lakes Based on a Fusion Approach. *Remote Sens.* **2025**, *17*, 2. https://doi.org/10.3390/rs17010002

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Rivers and lakes are an important source of support for the development of human civilization and an important part of the global hydrological cycle [1]. Due to rapid social development, the space encroachment of humankind's activities on rivers and lakes has become increasingly serious, and it is especially common to construct buildings that hinder

flooding within the management areas of rivers and lakes. The resulting lack of water space will lead to a reduction in the self-regulating ability of the water body, which will result in difficulty in conducting flooding and drainage, ecological deterioration, and other problems, and will threaten the safety of people's lives and properties [2,3]. Therefore, it is crucial to urgently address the ineffective river and lake regulatory efforts to reduce the number of buildings encroaching on water spaces. Due to the numerous rivers and lakes in the study area, it is difficult to count the buildings within the relevant management area, and it is also difficult to detect their changes over time. In the current supervision of river and lake waters, the buildings in remote-sensing images are usually extracted by manual labeling, and with limited human resources, it is difficult to meet the daily supervision needs of rivers and lakes by manually labeling a large number of building samples [4]. In recent years, aerial remote-sensing technology has developed rapidly, such as unmanned aerial vehicle (UAV) remote sensing and deep learning (DL) algorithms, and it is now possible to obtain images of the daily supervision areas of rivers and lakes through aerial remote sensing. The use of DL algorithms instead of manual labor to effectively and accurately extract buildings for the management of rivers and lakes can significantly enhance the efficiency of daily supervision [5]. Currently, few scholars have studied building segmentation DL algorithms for river and lake regulation scenarios, and the problem of how to effectively minimize the influence of environmental factors on DL algorithms needs to be solved urgently. This includes the texture characteristics of water specular reflection and the characteristics of boats along rivers and lakes which are similar to the characteristics of neighboring buildings, as well as the question of how to simultaneously improve the accuracy of extracting complex and irregular buildings.

In the early 21st century, machine learning methods such as support vector machines, artificial neural networks, and adaboost and random forest algorithms have been applied to extract architectural objects from remote-sensing images [6,7]. These methods can extract the edges and local information of buildings from remote-sensing images. But with the development of remote-sensing technology and the resulting increased image resolution, these methods are unable to meet the application requirements of pixel-level building extraction to the extent that DL algorithms can [8]. The earliest DL model for the semantic segmentation of images was a fully convolutional network (FCN) proposed by Shelhamer et al. [9], which can obtain the categories of pixels from the abstract features of an image for semantic level image segmentation. Meanwhile, Ronneberger et al. [10] designed U-Net, with an encoder-decoder structure, for semantic segmentation, which captures contextual information and precise localization through a U-shaped structure containing completely symmetric compressing paths and expansive paths. Thereafter, encoder-decoder-based architectures have been widely used for architectural image segmentation using full convolutional neural networks. In 2016, SegNet, proposed by Badrinarayanan et al. [11], was used to perform non-linear up-sampling utilizing pooling indexes computed in the maximum pooling step, which improves the quality of boundary segmentation while reducing the memory requirement. Since the so-called transformer algorithm was proposed by Google in 2017 [12], combining the transformer with CNN model represented by U-Net and its variants has become a hot spot of semantic segmentation research. Its hybrid model often outperforms the single transformer and CNN model. Liu et al. [13] applied a joint attention block to the encoder section of U-Net, and CCT as a skip connection. The transpose cross attention module (TCA) was proposed as a decoder for UJAT-Net, which can effectively segment edge details at different scales. Lin et al. [14] designed DS-TransUnet by merging the hierarchical Swin transformer into U-shape net, which can efficiently capture multi-scale information. Although the hybrid model combining CNN and a transformer has

better performance in semantic segmentation, the segmentation accuracy of multi-scale targets and detailed features still needs to be further improved.

Deep learning-based semantic segmentation algorithms for buildings in remote-sensing images are a data-driven approach, which means that such algorithms often require abundant remote-sensing images of labeled buildings for training and validation. The use of a large number of high-quality labeled images for training can largely improve the segmentation effect of the algorithm. Currently, the most commonly used remote-sensing building datasets in related research mainly include the Inria Aerial Image Labeling Dataset [15], the Massachusetts Buildings Dataset [16], and the WHU Building dataset [17]. Liu et al. [18] proposed SU-Unet, based on U-Net, which used the feature compression activation module within the encoder to increase the network feature learning capability, which provides better recognition of irregular and different-sized buildings in the Massachusetts Buildings Dataset. Lu et al. [19] proposed a dual-resolution U-Net using pairs of images as the input to capture multi-scale resolution features and achieved high-precision segmentation on the Inria Aerial Image Labeling Dataset. Wang et al. [20] designed ENRU-Net by combining a U-shaped coder-decoder structure and APNB. This network adopts the encoder-decoder structure for the fine extraction and reduction of feature mapping, and captures global contextual information through the self-attention mechanism, which has yielded high accuracy in the WHU Building Dataset. Although all the above building datasets can be used to test the segmentation accuracy and detail feature extraction effect of a semantic segmentation algorithm on multi-scale building targets, there is still a lack of building types in river and lake supervision work scenarios in China [21]. For example, temporary buildings with a colored steel plate structure and permanent buildings with brick and concrete structure or complex segmentation scenarios faced in daily supervision work, such as the specular reflection of water bodies to different degrees and boats with similar features to buildings. At present, only a few scholars [22] have studied such complex scenarios and there is a lack of datasets for this scenario. The latter limits the application and development of building segmentation algorithms for remote-sensing images in real scenarios. The current research on building segmentation has focused on solving the multi-scale problem and improving the accuracy of building segmentation. However, few studies have focused on how to minimize the negative effects of similar features (e.g., specular reflections from water bodies and boats along rivers and lakes) on building segmentation in practice while improving the accuracy of building segmentation.

In purview of the above, this study proposes a U-shaped network for the semantic segmentation of buildings, RDAU-Net, that fuses a CNN and a transformer. In order to make the reader better understand RDAU-Net, this study first introduces the application of a CNN and transformer in remote sensing in Section 2, then describes the proposed method and its working principle in Section 3, gives the specific setup of the experiments in Section 4, and, finally, analyzes and discusses the experimental results. Overall, the present study aims at making the following contributions:

- It proposes RDAU-Net, consisting of a residual dynamic short-cut down-sampling module, RDSC, and two attention mechanism modules, MCCT and DCF; this model is capable of accurately segmenting buildings and eliminating the effects of specular reflections from water bodies as well as complex backgrounds such as boats;
- The residual dynamic shortcut down-sampling (RDSC) module is also proposed herein, aiming at solving the problem of the high loss of feature information during down-sampling in the decoding stage. This module can enhance the model's representation without adding depth or width to the network and facilitate the integrity of building segmentation and model detail feature extraction;

- The multi-channel cross fusion transformer (MCCT) module reduces the semantic differences that exist in the encoder-decoder features, greatly reduces the negative impact of boat targets on building segmentation, and makes it less likely that independent buildings at multiple scales will be missed during segmentation;
- The double-feature channel-wise fusion attention (DCF) module is proposed for fusing MCCT input features with decoder features. It can reduce the semantic differences between decoded features at the same scale, which is beneficial in reducing the effects of specular reflections from water bodies and the impact of boats along rivers and lakes on building segmentation;
- The HRI Building dataset was constructed under the guidance of river and lake regulation staff and is the first building dataset to include multiple water-edge building types to represent complex scenarios of river and lake regulation. It fills the gaps regarding river and lake regulation scenarios in the existing building semantic segmentation datasets.

2. Related Works

2.1. CNN for Semantic Segmentation in Remote Sensing

Convolutional neural networks (CNNs) represent the main algorithm type used for the semantic segmentation of remote-sensing images, named as such because the networks before this type of network generally used matrix multiplication, while this network type uses the mathematical operation of convolution. The mainstream CNN algorithms currently applied in the field of remote-sensing semantic segmentation mainly include U-Net, DeepLab, ResNet, their related improved algorithms, etc. U-Net is the first CNN architecture with an encoder-decoder structure [10], which learns the high-level information of an image by down-sampling the semantic information and then classifies images at the pixel level. DeepLab, proposed by Chen et al. [23] in 2015, improves the model's ability to acquire semantic information as well as its detail capturing by combining probabilistic graphical models and deep convolutional neural networks (DCNNs). Improvements to CNNs have focused on introducing an attention mechanism, improving the convolutional form of CNNs, and introducing residual connections. Since the U-Net network cannot deal with multi-scale problems in remote-sensing images well, U-Net++ [24] forms a more efficient and flexible feature fusion method through the use of a nested U-Net structure and an attention gates mechanism to improve the multi-scale information processing ability. DeeplabV3+ [25] is an improved model based on Deeplab. Deeplab is improved due to combining the ASPP module and an encoder-decoder structure and applying depth separable convolution to the above two structures, as a result of which the model's performance has been significantly enhanced. The improved semantic segmentation algorithm based on U-Net is a hot spot in remote-sensing image processing research. For example, Yu et al. [26] designed AGs-Unet, which enhances the feature learning ability of U-Net while focusing on the feature information of small-scale buildings. Zhong et al. [27] combined the improved ASPP module with skip connections based on U-Net, which overcomes the limitation of U-Net in small-target detection. Liu et al. [28] proposed the EMU-CNN framework by combining MFUN, RPN, and ECMN based on U-Net and achieved good results and strong robustness on the WHU aerial dataset.

2.2. Transformer for Semantic Segmentation in Remote Sensing

A transformer [12] is a network model that combines the encoder-decoder structure with multi-head self-attention, and was first applied in the field of natural language processing (NLP). In 2020, Dosovitskiy et al. [29] proposed a vision transformer (ViT), which is the first time that a transformer was directly applied to image processing by splitting the

image into patches of prior embedding sequences as the input of the transformer to obtain an output suitable for image classification. Based on previous studies, the Swin transformer was further proposed by Liu et al. [30], who designed the shifted windows scheme and improved the model's efficiency through cross-window connections. Its powerful global modeling capability and local feature capturing capability enable the transformer model to handle remote-sensing imagery with rich multi-scale features. Since then, scholars refer to the Swin transformer and continuously improve the use of transformers in remote-sensing image segmentation. Meng et al. [31] designed the class-guided transformer for the decoder part and introduced the Swin transformer as an encoder, which achieved good segmentation performance on the ISPRS Vaihingen and Potsdam datasets. Xu et al. [32] proposed a novel transformer called a mixed-mask transformer (MMT). Through the MSL strategy, the transformer model can be used effectively to segment large-scale changing targets in remote-sensing images. Chen et al. [33] fused local feature mapping and global dependencies using the channel-spatial transformer module and embedded hybrid attention in a transformer. This allowed them to obtain more comprehensive semantic information through a global cross-fusion module, which improved the traditional transformer's ability to segment the global contextual information of remote-sensing images.

2.3. Combining CNN and Transformer for Semantic Segmentation in Remote Sensing

The sole utilization of either a CNN or transformer fails to satisfy the requirement for high precision in remote-sensing image segmentation. Only by combining the respective advantages of CNNs and transformers can their segmentation performance for remote-sensing images be further enhanced. Xiao et al. [34] proposed a coding enhancer based on a displacement window (Swin) transformer and combined it with the U-shaped network to form STEB-UNet, which is able to recognize and extract buildings of different sizes. Wang et al. [35] used a CSW transformer and MSCA to construct the U-Net network and proposed the MCAT-UNet network, which has a stronger multi-scale segmentation capability and remote spatial information capturing capability. Liu et al. [36] proposed MBT-UNet, based on a pyramid vision transformer, and designed a multi-branch encoder to capture features of different scales and refine the model's remote-sensing semantic segmentation results using a multi-scale up-sampling module (MSUM). Ding et al. [37] designed a spatial enhancement module (SEM) to enhance the Swin transformer in the encoding part, and its feature extraction capability was improved as a result. In addition, a LEM was designed to extract local semantic information in the decoding part. Dimitrovski et al. [38] designed a multi-axis vision transformer, EfficientNet, and ConvFormer, three U-Net networks, and integrated geometric mean generation into them to enhance their performance in evaluating remote-sensing images. Generally speaking, techniques such as designing the encoder, decoder, or hopping connection and combining them with a new transformer based on vision transformers is very suitable for the semantic segmentation of remote-sensing images. It effectively fuses the local feature extraction capability of CNNs and the multi-scale feature processing capability of transformers. Although the fusion can improve the model's segmentation ability for complex remote-sensing scenes, it needs to be further verified in edge detail processing or special application scenarios.

3. Methods

3.1. Over Architecture of RDAU-Net

In this study, an improved U-shaped attention network, RDAU-Net, is proposed based on the traditional U-Net, which is optimized for the deficiencies of the U-Net in the encoder, decoder, and skip connections parts, respectively, as shown in Figure 1. In the encoder part, the residual dynamic short-cut down-sampling (RDSC) module is proposed

to replace the convolution operation of the traditional down-sampling process. The RDSC increases the information correlation between feature maps of different scales, and reduces the interference of the complex building shapes and the differences in building scales on the segmentation results; the multi-scale channel cross-fertilization module (MCCT) is adopted to replace the traditional skip connections to further increase the information correlation between feature maps of different scales, which reduces the resolution and semantic gaps between multi-scale features and improves the wholeness and overall accuracy of building segmentation; the multi-feature channel fusion module (DCF) is proposed for use in the decoder part to replace the simple splicing of different features in the decoding process, to better fuse the semantically inconsistent same-scale features, to minimize the negative impact of similar features, and to reduce the segmentation redundancy of the edge details.

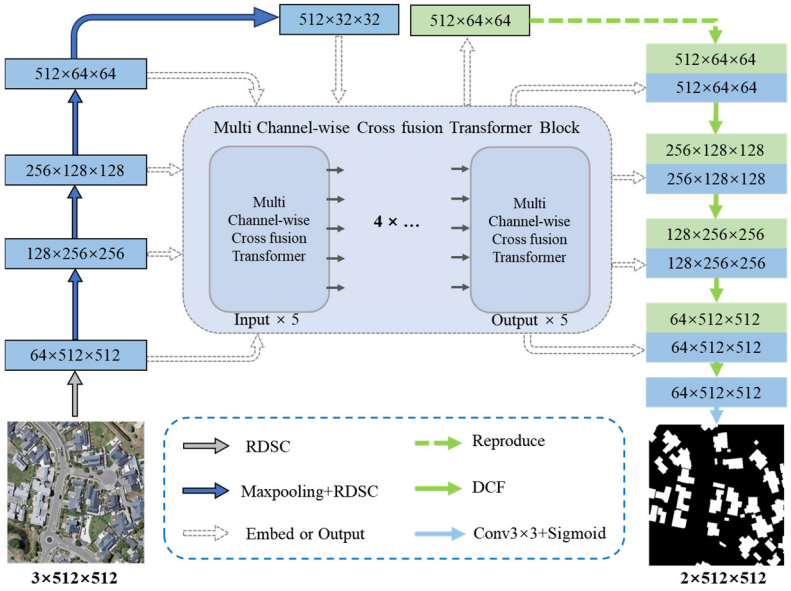


Figure 1. The architecture of RDAU-Net.

3.2. Residual Dynamic Short-Cut Down-Sampling (RDSC)

3.2.1. Dynamic Convolution

First proposed by Chen et al. [39], dynamic convolution distinguishes itself from traditional convolution operations that share a single convolution kernel, as it is based on the dynamic aggregation of attention across multiple parallel convolution kernels. Compared with traditional convolution, it aggregates in a non-linear manner through the attention mechanism, which results in stronger feature representation capabilities and can increase the complexity of semantic segmentation network models without adding depth or width to networks. As shown in Formulas (1)–(3) and Figure 2, after x is input, k attention weights are generated through the attention block, and then multiplied with the corresponding k linear models $W_k^T x + b_k$, after which a non-linear ensemble model $W^T(x)x + b(x)$ can finally be obtained. Therefore, different inputs will generate different attention weights, and the resulting non-linear model will have more representation power than the traditional convolutional linear model.

$$y = \text{Relu}(W^T(x)x + b(x)) \tag{1}$$

$$W(x) = \sum_{k=1}^{K} \pi_k(x) W_k \tag{2}$$

$$b(x) = \sum_{k=1}^{K} \pi_k(x) b_k \tag{3}$$

where π_k is the weight of attention, its value range is [0, 1], and the total weight is always 1; W is the weight matrix and b is the bias vector.

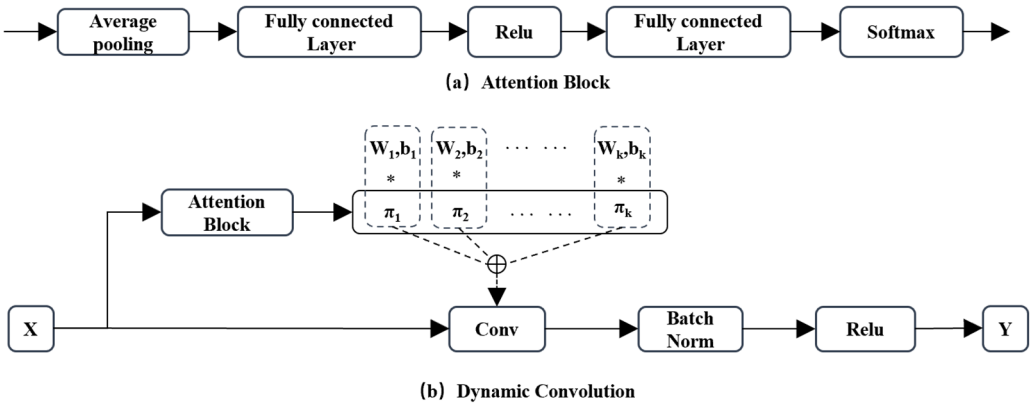


Figure 2. The running flow of dynamic convolution.

This technique is not limited to the direct replacement of traditional convolution with dynamic convolution, and its application is beneficial to the performance of remote-sensing image segmentation models. For example, Feng et al. [40] proposed Gaussian dynamic convolution (GDC) based on dynamic convolution combined with the Gaussian distribution of offsets to randomly select a spatial adoptive region, which can effectively aggregate the contextual information of the model. Hou et al. [41] proposed dynamic hybrid convolution (DHGC) to enhance the ability to gather boundary information. Wang et al. [42] utilized self-concerned mechanisms and dynamic convolution to design LDCM, which can reduce the computational cost while enhancing the model's feature extraction ability. Chen et al. [43] combined the convolutional operation to design a multistep down-sampling structure in residuals, improving the model's adaptability to different scales of inputs.

3.2.2. Specific Structure of the RDSC

In the process of down-sampling of the encoder, the linear transmission of traditional convolution is not conducive to the transmission of complex feature information. In addition, simple and multiple convolution operations will lose part of the original upper feature information, which is not conducive to effectively transferring the upper feature information to the lower layer. Moreover, such loss will accumulate continuously among the features of different scales in multiple layers, thus affecting the model's capture of multi-scale feature information.

Therefore, in the present study, a residual dynamic short-cut down-sampling (RDSC) module is proposed, as shown in Figure 3. Dynamic convolution is introduced to improve the convolution operation. This non-linear model with stronger representation power is more conducive to the transfer of complex features in the upper layer. Since the convolution kernel of the dynamic convolution is related to its input, it is beneficial for different features to better convey relevant information. The down-sampling process with the residual structure can not only extract image detail features through convolution operation, but can also retain part of the upper-layer feature information, which reduces the loss of feature information in the transmission process and is conducive to the acquisition of multi-scale feature information. With features of different scales as initial inputs, the output is obtained after dynamic convolution with a convolution kernel size of 3, batch normalization, and

ReLU activation function operations. The output is combined with the initial input of the dynamic convolution and a batch normalization operation is conducted with the convolution kernel set to 1, the final output of which is the subordinate feature after the activation function.

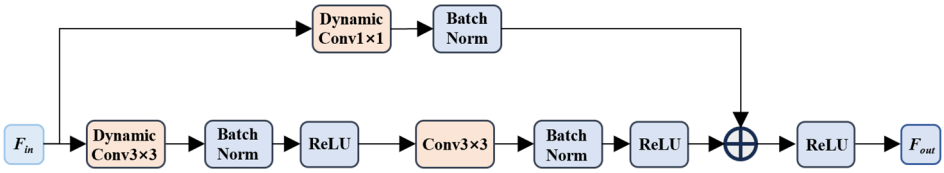


Figure 3. The structure of RDSC module.

3.3. Multi-Channel Cross Fusion Transformer (MCCT)

In the U-Net with an encoder-decoder structure, there exists a certain degree of a semantic gap between encoder-decoder features at the same level, which cannot be bridged by only the independent simple replication of skip connections, and the direct replication of encoder features into the decoder may even lead to degradation of the model's performance, especially for shallow encoder features with little semantic information [44]. Therefore, reducing the semantic gap between multi-scale features helps to improve the model's segmentation. The multi-head cross-attention [45] model can fuse multi-scale context and cross-attention from the perspective of the feature channel, and it combines the characteristics of the multi-scale encoder and the advantages of the transformer's long sequence modeling. Multi-head cross-attention can effectively fuse features from different scales and channels, and then fuse information from different levels and angles to help the model understand and analyze the image.

In this study, the multi-channel cross fusion transformer (MCCT) module is proposed, as shown in Figure 4. It takes five encoder features of different scales as inputs and performs channel-level feature fusion on the inputs through the use of multi-head cross-attention to reduce the semantic and resolution gaps between the multi-scale features, and finally obtains five outputs with the same scale as the inputs. The next two subsections describe the two main parts of the module and the execution process.

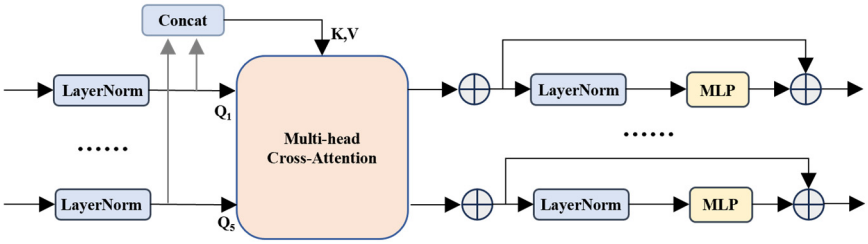


Figure 4. The structure of the MCCT module.

3.3.1. Multi-Scale Feature Embedding

The five layers of encoder features with different scales, $E_i \in \mathbb{R}^{\frac{H\times W}{2^{i-1}}\times C_i}$, ($i$ = 1, 2, 3, 4, 5), are used as inputs to the MCCT module, respectively, where C_1 = 64, C_2 = 128, C_3 = 256, C_4 = 512, and C_5 = 512. Since the transformer needs a series of token embeddings as input, here, E_i is reconstructed with features to form a 2D patch sequence $L_i \in \mathbb{R}^{l\times C_i}$ of size $\frac{P}{2^{i-1}}$, where P = 16 and l is the number of patches. Then, the five processed patch sequences are fused into S as Key and Value, and the computational formula is expressed as follows:

$$S = Concat[\delta(L_1), \delta(L_2), \delta(L_3), \delta(L_4), \delta(L_5)] \quad (4)$$

where *Concat* is the feature fusion operation and $\delta(\cdot)$ denotes the normalization operation.

3.3.2. Multi-Head Cross-Attention

To fully obtain the multi-scale features of each layer of the encoder, the normalized sequences L_i and S are used as inputs to the multi-head channel attention mechanism, and then the channels and their associated terms are encoded using a residual-structured multilayer perceptron (MLP).

The multi-head attention mechanism used in this study consists of six inputs (shown in Figure 5). There are five different input sequences L_i and one integrated embedding layer S, defined as K and V. The inputs are then transformed into Q, K, and V matrices by the weights W_{Qi}, W_K, $W_V \in \mathbb{R}^{l \times C_i}$ of the different inputs as shown in the following equation:

$$Q_i = \delta(L_i) \times W_{Qi} \tag{5}$$

$$K = S \times W_K \tag{6}$$

$$V = S \times W_V \tag{7}$$

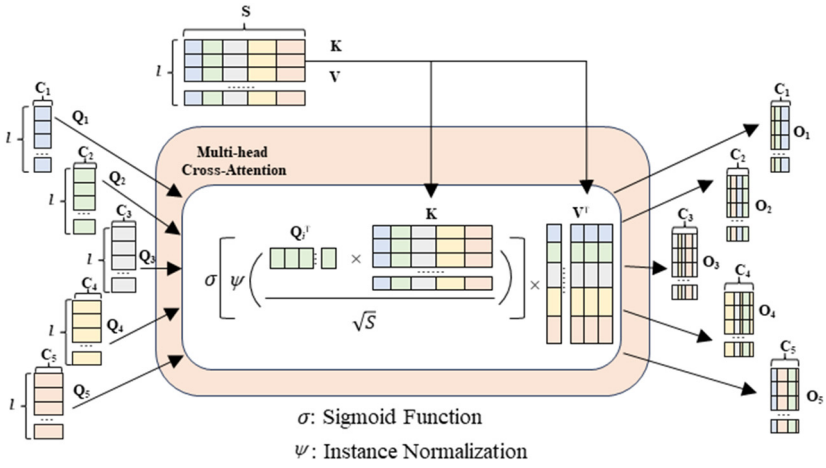


Figure 5. Multi-channel attention mechanism.

Further, in the case of n-head attention, the input matrix is further processed through Equation (9) to obtain the output matrix MHCA of the channel attention mechanism:

$$Attention_i = \sigma\left[\psi\left(\frac{Q_i^T \times K}{\sqrt{C_\Sigma}}\right)\right] \times V^T \tag{8}$$

$$MHCA_i = \frac{Attention_i^1 + Attention_i^2 + \cdots + Attention_i^n}{n} \tag{9}$$

where $\psi(\cdot)$ is the instance normalization and $\sigma(\cdot)$ is the original sigmoid function.

Finally, the input of the attention mechanism is regularized, and the multilayer perceptron is residually processed through Equation (10) to obtain the final output. In this study, the operation of Equation (10) is performed a total of four times.

$$O_i = MHCA_i + MLP[\delta(Q_i + MHCA_i)] \tag{10}$$

3.4. Double-Feature Channel-Wise Fusion Attention (DCF)

The main purpose of up-sampling is to fuse and re-amplify the output features of the skip connections with the corresponding same-scale features in the decoder to retain more detailed information. However, no matter whether the output features are processed by traditional skip connection or the MCCT module, there will be some semantic gap with the same-scale features of the decoder. If two same-scale features are directly spliced, it

does not reduce the semantic difference between them, and it also leads to ambiguity in the up-sampling process, which affects the retention of detailed information of the semantic segmentation target. Therefore, this study designs the double-feature channel-wise fusion attention (DCF) module. By filtering the channel information of the output features of the MCCT module, the semantic differences between them and the decoder features are reduced. This enhances the detailed representation of the original feature map while better fusing the two same-scale features by short-cut. The principle of operation is shown in Figure 6.

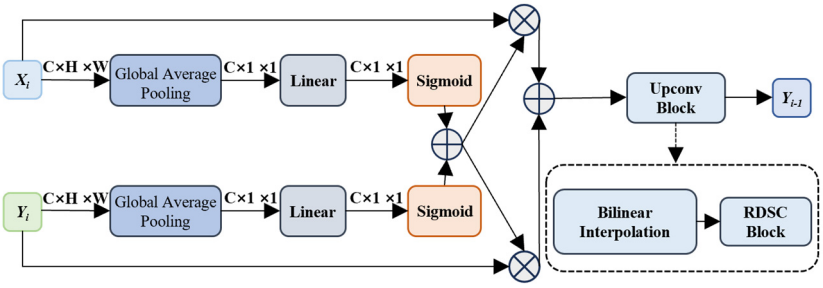


Figure 6. The structure of the DCF module.

The two feature vectors in layer i of the decoding part are set as X_i and $Y_i \in \mathbb{R}^{C\times H\times W}$, respectively. To avoid the negative impact of dimensionality reduction on the channel attention [28], the vectors $f(X_i)$ and $f(Y_i)$, shown in Equation (11), are formed by embedding the global spatial information through the global average pooling (GAP), compressing the scales of the input features, followed by using a single linear layer and a sigmoid function to construct the channel feature maps, and finally obtaining the vectors $\zeta(X_i)$ and $\zeta(Y_i)$. Further, after fusing $\zeta(X_i)$ and $\zeta(Y_i)$, the initial features X_i and Y_i are calibrated for activation and re-fused according to Equation (12), respectively, after which, finally, up-sampling is performed to obtain the feature vector Y_{i-1} of the i-1th layer.

$$f(\kappa) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \kappa^k(i,j) \tag{11}$$

$$\zeta(\kappa) = \alpha \times f(X), G(\kappa) = \sigma[\zeta(X_i) + \zeta(Y_i)] \times (X_i + Y_i) \tag{12}$$

where k is the kth channel of the feature; the activation function $\sigma(\cdot)$ characterizes the importance of the channel; and $\alpha \in \mathbb{R}^{C\times C}$ is the weight of the linear layer and the sigmoid operator.

4. Experimental Set Up

4.1. Datasets

The HRI Building dataset, which consists of remote-sensing building images taken within the management areas of some rivers and lakes in Jiangsu Province, China, and the high-resolution public remotely sensed building image dataset, the WHU Building dataset, were used in this study, respectively. The characteristics of these datasets are briefly outlined below.

4.1.1. HRI Building Dataset

The HRI Building dataset was mainly collected in two cities in Jiangsu Province in the second half of 2023, using a DJI Warp M30 series drone equipped with an M30T camera, and all of them are three-channel RGB images of 3840 × 2160 pixels. In addition, the dataset also contains UAV aerial RGB images of some river regulation work scenarios

in Jiangsu Province in 2023, with rich samples of river and lake water-edge buildings (3419 water-edge buildings). As shown in Figure 7, this dataset contains rich samples of river and lake supervision work scenarios, such as water-edge buildings, specular reflections of different degrees of water bodies, different numbers of shoreline vessels, different densities of permanent and temporary buildings, and building images with different lighting conditions. In order to fully utilize the source images, they were cropped to 512 × 512 pixels, and for the rows and columns with less than 512 edge positions, the inverse 512 was used to form the cropping region; finally, the images that do not contain buildings were eliminated, and 817 images were retained as the dataset.

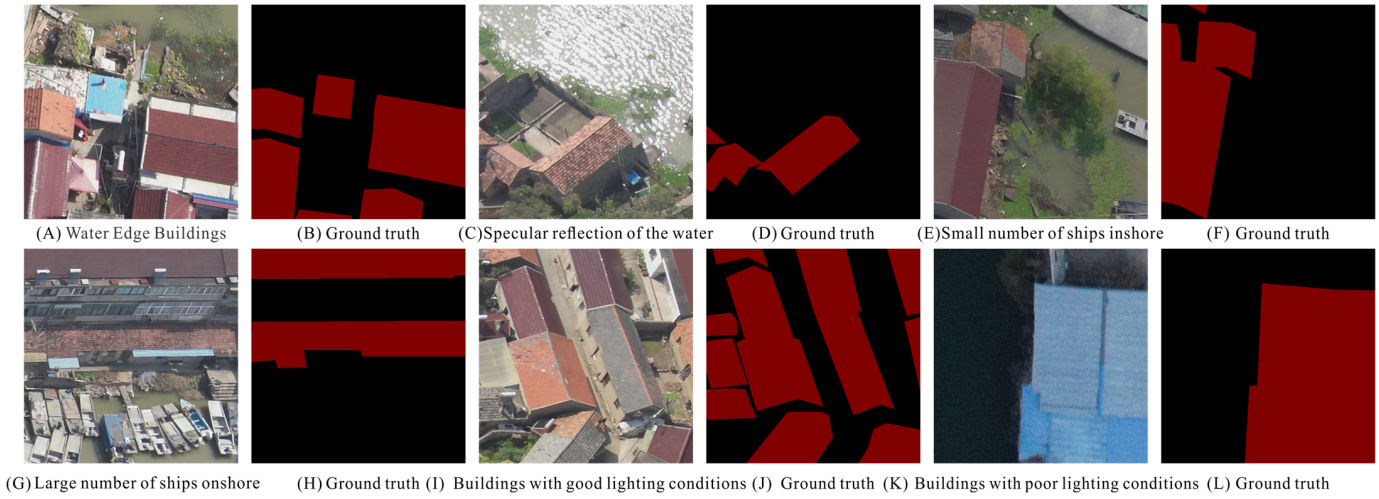


Figure 7. Example of a typical sample of the HRI Building dataset.

4.1.2. WHU Building Dataset

The WHU Building dataset is a commonly used public dataset of building samples provided by Wuhan University, and is widely used in building segmentation research at home and abroad. The dataset consists of 18,700 individual buildings extracted from aerial images covering Christchurch, New Zealand, with a ground resolution of 0.3 m and an image size of 512 × 512 pixels. The dataset contains a total of 8189 images, including 4736, 1036, and 2416 images in the training, validation, and test sets.

4.2. Implementation Details

The models were trained on the Google Collab platform equipped with A100 (40G memory) GPUs, Python 3.7, and Pytorch. The total epoch of the model training was 100.

Due to the small amount of data having potential to cause poor approximation and underfitting of the network, data enhancement was performed on the images of the HRI Building dataset, which were mainly vertically mirrored and rotated by 90°, 180°, and 270°. A total of 3268 images was obtained, which were randomly partitioned into training, validation, and test sets in the ratio of 8:1:1. The first 50 epochs of training were the freezing phase, which freezes the backbone feature network and can reduce the training time. The last 50 epochs were the unfreezing phase. The optimizer used was the Adam optimizer, the batch size was set to 2, the initial learning rate was 1×10^{-4}, and the cosine annealing strategy was combined with the model to attenuate the learning rate to skip the local optimum.

4.3. Loss Function Selection

In the DL model optimization strategy, the loss function is an important decision function. For classification problems, dice loss [46], focal loss [47], and cross-entropy [48]

are generally chosen as the loss functions. The dice loss function can evaluate the degree of similarity between real and predicted images in binary classification tasks based on the mathematical relationship between segmented image pixels predicted to be buildings and labeled image pixels that are actually buildings. The smaller the value of the dice, the smaller the value of the Loss, the higher the degree of similarity, and the better the model effect, which is defined as shown in Equation (13). The focal loss function is based on the binary classification cross-entropy CE, which can dynamically adjust the training weight of easily segmented samples using a dynamic scaling factor, so that the training center of gravity tends to make it difficult to segment samples. The smaller the value of the focal loss, the better the training balance between the positive and negative samples and the better the training balance of difficult and easy samples, which is defined as shown in Equation (14).

$$L_{Dice} = 1 - \frac{2|X \cap Y| + smooth}{|X| + |Y| + smooth} \tag{13}$$

$$L_{Focal} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{14}$$

where X is the set of pixels predicted to be buildings; Y is the set of pixels that are actually buildings; the *smooth* term avoids the denominator of the splitter being 0 when both X and Y are 0, which is generally set according to the experimental accuracy, and 1×10^{-5} is taken in this study; α_t is the weighting factor, which is mainly used to regulate the ratio between the negative and positive samples' losses, with the value of the interval being [0, 1], and with 0.5 being taken in this study; $(1 - p_t)^{\gamma}$ is the modulation factor, which is used to reduce the loss contribution of easy-to-split samples, where $\gamma \in [0, 5]$, and 2 is taken in this study; p is the probability that the model predicts a pixel as a building, and if the pixel is a building then $p_t = p$, otherwise $p_t = 1 - p$.

Since the dice loss function mainly alleviates the training imbalance problem of positive and negative samples, it focuses on the mining of building samples but is prone to the loss saturation problem. The focal loss function focuses on samples that are difficult to distinguish, which can reduce the misclassification of the background features that are similar to the building features to a greater extent, and it can alleviate the loss saturation problem of the dice loss function to a certain extent. Therefore, this study adds the dice loss and focal loss functions to form the DF loss composite loss function, so that the model has an improved accuracy in building segmentation while the misclassification impact of similar features is minimized.

4.4. Evaluation Metrics

To evaluate the segmentation performance of each CNN and transformer model on the building datasets reasonably and objectively, this study uses six metrics, namely precision, recall, F1 value, kappa coefficient, interaction ratio (IoU), and overall average accuracy (OA), to evaluate the six metrics based on the confusion matrix. These six metrics are calculated by the following formulas:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{17}$$

$$\text{Kappa} = \frac{po - pe}{1 - pe} \tag{18}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{19}$$

$$\text{OA} = \frac{TP + TN}{TP + FP + TN + FN} \tag{20}$$

where *po* is the number of correctly predicted pixels for each category divided by the total number of pixels, and *pe* is the sum of the products of the actual and predicted pixel counts corresponding to all the categories, respectively, divided by the square of the total number of pixels.

5. Results and Discussion

5.1. Ablation Study

To effectively evaluate the proposed three modules of up-sampling, down-sampling, and skip connection, ablation experiments are conducted on the HRI Building dataset and WHU Building dataset. A simplified U-Net network with five-layer depth is used as the baseline.

In particular, to better represent the building segmentation effect of different modules and network models for multiple working scenarios, six scenarios are set up based on the HRI Building dataset: A. Images with strong water specular reflection interference; B. Images with weak water specular reflection interference; C. Neighboring water but without water specular reflection; D. Images of complex distribution of buildings with good lighting conditions; E. Images of simple distribution of buildings with dim conditions; F. Images of complex distribution of buildings with dim conditions.

Based on the WHU Building dataset, sparse building distribution scenario (A), dense large target building distribution scenario (B, C), a scenario with both dense large target buildings and small target buildings (D), and a dense small target building distribution scenario (E) were set up.

5.1.1. Effect of RDSC

As shown in Table 1, after the introduction of the RDSC module, all the evaluation indexes of the two datasets improved to different degrees, although we focus here focusing on the OA, IoU, and F1 score indexes. In the HRI Building dataset, the F1 score improved by 1.33%, the IoU index improved from 83.68% to 85.94%, a significant improvement of 2.26%, and the OA index improved by 1.29%. In the WHU Building dataset, the F1 score improved by 0.62% from 96.62% to 97.02%, the IoU improved by 0.62% from 93.69% to 94.31%, and the OA metric improved by 0.17% from 98.65% to 98.82%.

Table 1. The results of ablation experiments.

Dataset	Modules			Evaluation Index (%)					
	RDSC	MCCT	DCF	Precision	Recall	F1	Kappa	IoU	OA
HRI Building	-	-	-	91.14	91.05	91.09	82.19	83.68	91.31
	√	-	-	92.38	92.48	92.42	84.85	85.94	92.60
	√	√	-	92.57	92.51	92.53	85.07	86.14	92.72
	√	√	√	**93.54**	**93.62**	**93.58**	**87.15**	**87.94**	**93.72**
WHU Building	-	-	-	96.41	96.85	96.62	93.25	93.69	98.65
	√	-	-	97.24	96.79	97.02	94.04	94.31	98.82
	√	√	-	**97.38**	96.80	97.09	94.17	94.43	98.85
	√	√	√	97.28	**97.06**	**97.17**	**94.34**	**94.59**	**98.88**

The bold font represents the optimal values.

To more intuitively show the improvement in the model segmentation ability allowed by the RDSC modules, the visualization results shown in Figures 8 and 9 are analyzed next. In the HRI Building dataset, the model with the introduction of the RDSC module reduces the incorrect segmentation of more buildings while slightly increasing the incorrect segmentation of the "strong water specular reflection" part of scenario A. In scenario B, the "weak water specular reflection" does not negatively affect the model's segmentation results, and the model is shown to be able segment the edges of the buildings better than in the baseline. In scenario C, the model significantly reduces ship misrecognition and allows for the finer segmentation of buildings. In scenario D, the baseline makes it difficult to segment complexly shaped buildings and cannot maintain the integrity of building segmentation, while, after adding RDSC, the complex buildings in the three red boxes on the right side are obviously better segmented. For scenarios E and F, which have poor lighting conditions, the model increases the completeness of the segmentation results for the interiors of the buildings, and at the same time can effectively separate each building. The results show that RDSC can effectively improve the model's segmentation of building shape features by reducing the loss of feature information in the down-sampling process and, at the same time, it can effectively eliminate the influence of some boats on the segmentation results, but it is not much help in improving the stronger specular reflections of the water body.

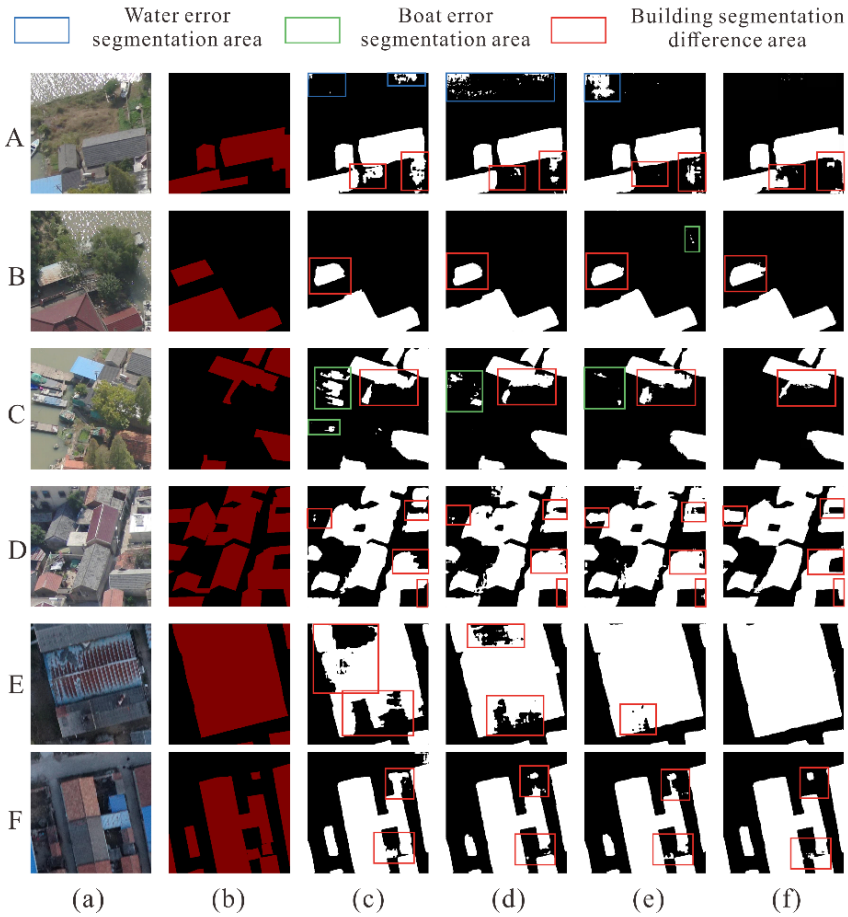


Figure 8. Results of ablation experiments on the HRI Building dataset. (**a**) Image. (**b**) Ground truth. (**c**) Baseline. (**d**) Baseline + RDSC. (**e**) Baseline + RDSC + MCCT. (**f**) Baseline + RDSC + MCCT + DCF.

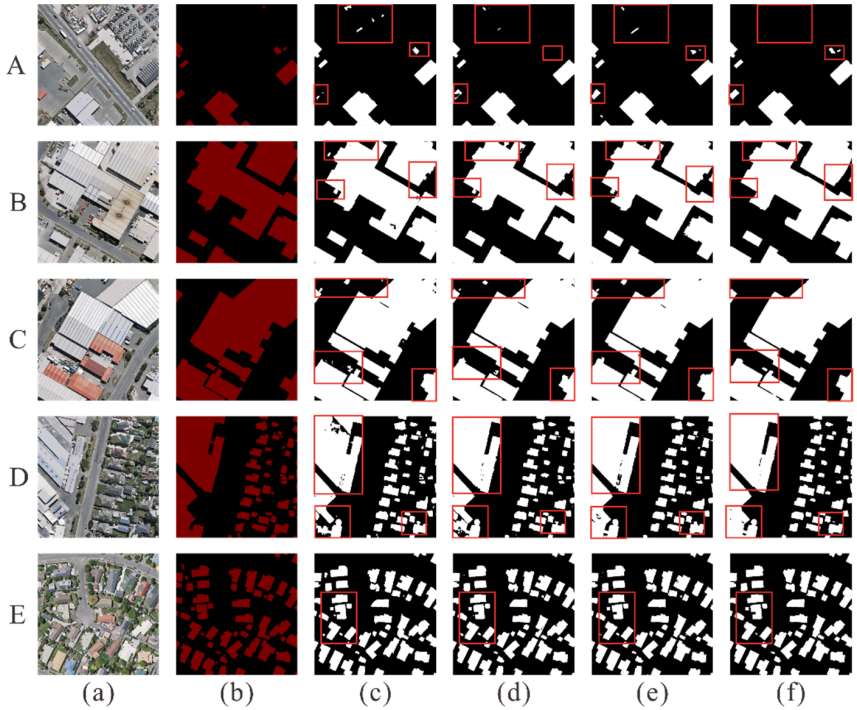


Figure 9. Results of ablation experiments on the WHU Building dataset. (**a**) Image. (**b**) Ground truth. (**c**) Baseline. (**d**) Baseline + RDSC. (**e**) Baseline + RDSC + MCCT. (**f**) Baseline + RDSC + MCCT + DCF.

In the WHU Building dataset, this model significantly reduces the mis-segmentation in the red box above scenario A and improves the accuracy of building segmentation in the left box. In scenario B, the model is shown to be able to improve the completeness of the internal segmentation of the building and has finer detailing of the edges of the building in the red box. In scenario C, the accuracy of the segmentation of the edges of the buildings in the lower two red boxes is also improved. In scenario D, the model improves the internal integrity of the segmentation results for the two large target buildings on the left, while also improving the segmentation details of the small target buildings in the left box. In scenario E, the model improves the accuracy of detection of the small target buildings by a small margin. The results confirm the effectiveness of the RDSC module. Overall, the RDSC module not only improves the segmentation integrity of the two datasets with complex building distributions, but also improves the segmentation performance in scenarios with different scales of building distribution simultaneously. The results not only confirm the effectiveness of the RDSC module, but also show that the building segmentation accuracy can be effectively improved by introducing residual structures and optimizing the convolution method to reduce the feature loss in the down-sampling process.

5.1.2. Effect of MCCT

The MCCT module is further introduced based on the RDSC module, and the experimental results are shown in Table 1. In the HRI Building dataset, the F1 score improved by 0.11% from 92.42% to 92.53%. The IoU index improved by 0.2% from 85.94% to 86.14%. The OA index improved by 0.12% from 92.60% to 92.72%. In the WHU Building dataset, the F1 score improved by 0.07% from 97.02% to 97.09%. The OA metrics improved by 0.03% from 98.82% to 98.85%.

The results of the visualization of the HRI Building dataset are shown in Figure 7. In scenario A, there is a significant reduction in building mis-segmentation in the red box. In scenario B, the accuracy of the building segmentation in the red box is further improved. In scenario C, the incorrect segmentation portion of the ship is significantly reduced and

the buildings in the red box are segmented more accurately. In scenario D, the model eliminates the omission of buildings in the upper left corner, while the internal details of the buildings in the right box are depicted in greater detail. In scenarios E and F, the model greatly improves the segmentation completeness of the building interiors and mitigates the leakage in scenario F caused by the introduction of the RDSC module.

The enhancement of the model's segmentation results for the WHU Building dataset mainly focuses on refining the portrayal of the internal details of the buildings and reducing the omission of separate building segments at different scales, as shown in the visualization of the results in Figure 8. In scenario A, the omission of buildings in the right box disappears, and the segmentation accuracy of the buildings in the left box is improved. In scenario B, the interiors and edge details of the buildings in the upper red box are more finely delineated. In scenario C, the segmentation of the buildings in the lower-left red box is more complete. In scenario D, the internal segmentation of the buildings in the lower-left box is more complete, while the precision of the segmentation of the small target building in the lower-left corner has been improved. In scenario E, the amount of wrong segmentation results becomes less in the lower-left corner of the red box.

Overall, the introduction of the MCCT module resulted in a more focused range of errors due to specular reflections in the water body, fewer incorrect segmentations for boats, and further improvements in the segmentation accuracy for buildings. The omission of separate buildings was improved in all scenarios. The fusion of features at different scales from the channel perspective to reduce the semantic differences between the encoder and decoder features can effectively improve the accuracy of building segmentation, alleviate the influence of similar features on building segmentation, and reduce the omission of independent building segments, which also reflects the effectiveness of the MCCT module.

5.1.3. Effect of DCF

As shown in Table 1, the inclusion of DCF further improves the network. In the HRI Building dataset, the F1 score improved from 92.53% to 93.58%, an improvement of 1.05%. The IoU index improved from 86.14% to 87.94%, an improvement of 1.8%. The OA index improved from 92.72 to 93.72%, an improvement of 1%. In the WHU Building dataset, the F1 score improved by 0.08% from 97.09% to 97.17%.

As shown in Figure 7, the HRI Building dataset showed a large improvement in the segmentation results for each scenario. In scenario A, the network has eliminated the effect of strong water specular reflection on the building segmentation, and also improved the segmentation accuracy of the buildings below. In scenario B, the incorrect segmentation of the ship has been eliminated, and more correct segmentation of the buildings in the red box has been obtained. In scenario C, the incorrect segmentation of the ship has been eliminated, and the segmentation integrity of the buildings was improved and has better detail. In scenario D, the buildings in the red box on the left side are completely and correctly segmented, and the interior details of the buildings on the right side are efficiently portrayed, almost being the same as the ground truth. In scenario E, the integrity of the building interiors is further improved. In scenario F, the separate buildings in the top half are better segmented, and the accuracy of the buildings in the red box below is also improved. The results show that replacing the traditional decoded feature splicing and up-sampling with the DCF module can effectively eliminate the detrimental effect of strong water body specular reflections on building segmentation and improve the interior and edge details as well as the overall accuracy of the buildings.

As also illustrated in Figure 8, in the WHU Building dataset, the final model achieves good results in the segmentation of different building distribution scenarios. In scenario A, this model eliminates the erroneous segmentation in the upper red box while improving

the accuracy of the small target buildings in the remaining two red boxes. In scenario B, the model portrays the edges and interior details of the buildings more finely. In scenario C, the model eliminates the wrong segmentation in the upper part, while there is no missed detection of buildings in the lower-left box. In scenarios D and E, the model strongly improves the internal integrity of the large target building, while the accuracy of the small target buildings is further improved. The DCF module's effectiveness has been validated.

5.2. Comparison with State-of-the-Art Methods

To better evaluate the performance of our proposed model, we conducted experiments comparing the model with the state-of-the-art models, including FCNs [9], U-Net [10], U-Net++ [24], Swin-UNet [49], ACC-UNet [50], CSC-UNet [51], UCTransNet [45], and DTA-UNet [52]. Among them, the first four models are the most classical models for semantic segmentation, and the last four models are the most recent models that have been used for semantic segmentation in the last few years.

5.2.1. Results on the HRI Building Dataset

The model proposed in this study clearly achieves the best performance in each evaluation metric in the HRI Building dataset, as shown in Table 2. In the main evaluation metrics, compared to UCTransNet, which has the second-best performance, our model performed better by 1.07% in the F1 score, 1.85% in the IoU index, and 1.03% in the OA metrics. The model significantly outperforms the current state-of-the-art models on the HRI Building dataset.

Table 2. Results of the HRI Building dataset comparison experiment.

Method	Precision/%	Recall/%	F1/%	Kappa/%	IoU/%	OA/%
FCN	92.56	92.07	92.29	84.59	85.72	92.52
U-Net	89.74	89.77	89.75	79.5	76.86	89.99
U-Net++	87.71	88.14	87.89	75.78	78.43	88.1
Swin-UNet	80.54	79.92	80.17	60.37	67.06	80.82
ACC-UNet	85.39	84.48	84.83	69.69	73.76	85.36
CSC-UNet	89.97	89.42	89.66	79.32	81.3	89.97
UCTransNet	92.49	92.53	92.51	85.02	86.09	92.69
DTA-UNet	92.16	92.46	92.29	84.58	85.70	92.44
baseline	91.14	91.05	91.09	82.19	83.68	91.31
RDAU-Net (Ours)	**93.54**	**93.62**	**93.58**	**87.15**	**87.94**	**93.72**

The bold font represents the optimal values.

As shown in Figure 10, in scenario A, only FCNs, CSC-Unet, and our proposed RDAU-Net can eliminate the effect of strong water body specular reflection on building segmentation, but RDAU-Net significantly outperforms each of the other models in the completeness and accuracy of its building segmentation results. In scenario B, all models except U-Net and U-Net++ are unaffected by the weak water body specular reflection, but, for the building segmentation results, among the remaining models only UCTransNet and our model can maintain high building segmentation completeness and high accuracy. In scenario C, only UCTransNet and our model can eliminate the negative effect of the boat on the building segmentation, and the building segmentation accuracy of our model is better than that of UCTransNet. In scenario D, the model proposed herein is the only one that can completely segment the building in the red box on the left side, and the segmentation result in the red box on the lower right side has significantly better internal details and integrity than that of the other models. In scenario E, RDAU-Net outperforms the other models in terms of segmentation completeness and edge detail for the interiors of simple buildings. In scenario F, U-Net, UCTransNet, and our model significantly outperform the other models, but our model outperforms the first two models in terms of the segmentation

completeness for stand-alone buildings. The results show that our model can effectively eliminate the effects of strong water specular reflections and boats and can effectively improve the integrity of building segmentation as well as the segmentation of edge and interior details.

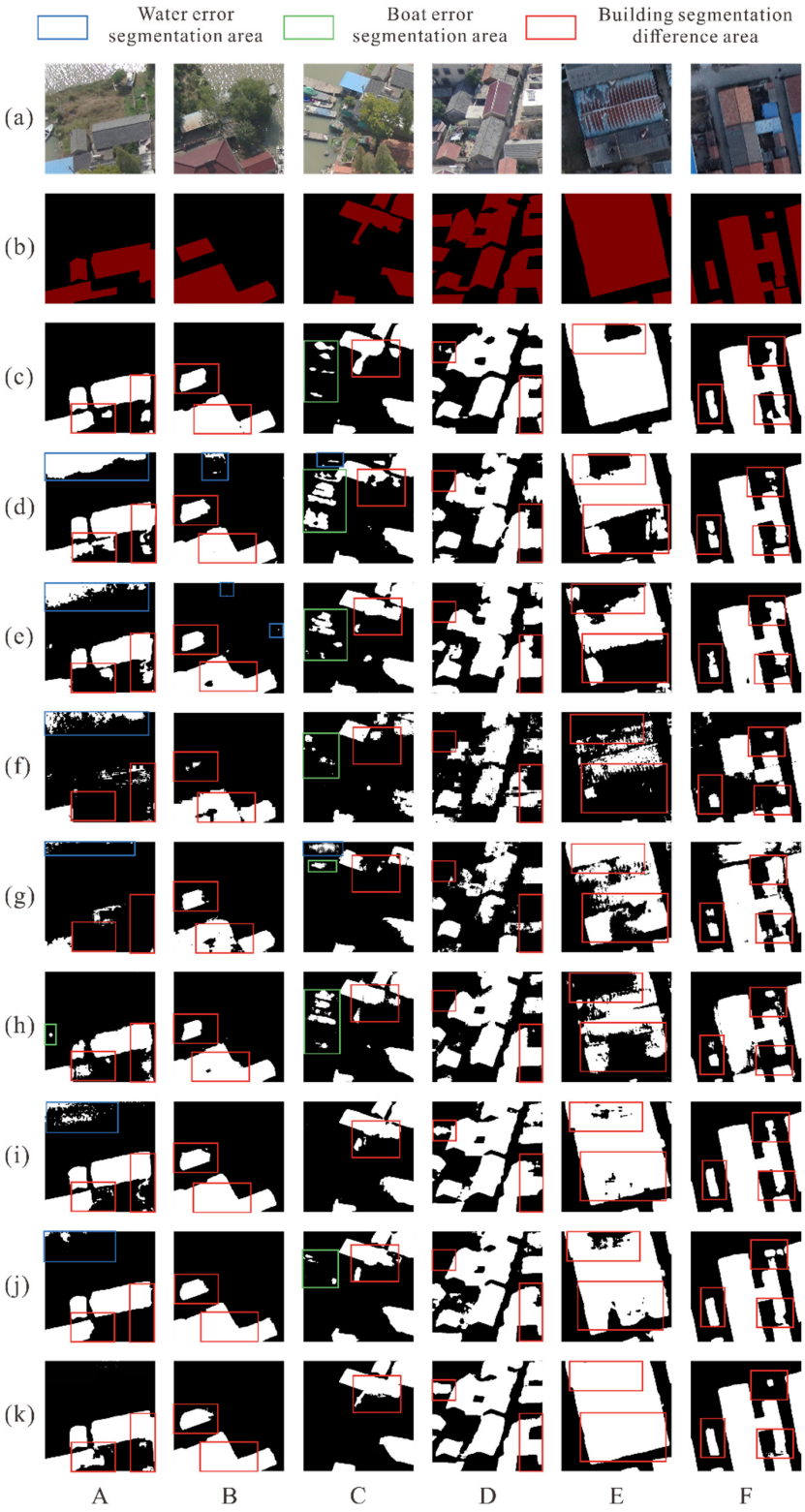


Figure 10. Visualization of the results of comparative experiments on the HRI Building dataset. (**a**) Image. (**b**) Ground truth. (**c**) FCN. (**d**) U-Net. (**e**) U-Net++. (**f**) Swin-UNet. (**g**) ACC-UNet. (**h**) CSC-UNet. (**i**) UCTransNet. (**j**) DTA-UNet. (**k**) RDAU-Net.

Overall, RDAU-Net outperforms the other models on the HRI Building dataset. It is worth noting that the FCN, as the first CNN model applied to semantic segmentation, is able to achieve good results in all five scenarios, and even slightly outperforms the latest models such as ACC-Unet, UCTransNet, and DTA-UNet in terms of eliminating the effect of specular reflection on water bodies as well as the accuracy of building segmentation. This means that, in river and lake littoral environments, some of the models that use a good combination of a CNN and transformer may not be able to be simpler and more effective than the FCN model. RDAU-Net, as the best-performing model, focuses on reducing the information loss in the feature transfer process and the semantic difference in the feature fusion, and under this design idea, the negative impacts of specular reflection of water bodies and ships are basically eliminated, and the segmentation accuracy and detail portrayal of buildings are improved. How to further reduce the information loss and semantic differences will be the focus of future research on the application of DL models in river and lake coastal scenes.

5.2.2. Results on the WHU Building Dataset

The comparison experiments on the WHU Building dataset are summarized in Table 3. RDAU-Net outperforms all other models in every evaluation metric except for recall. In comparison to the second-best model, our model shows improvements of 0.01% in the F1 score, 0.2% in the kappa coefficient, 0.19% in the IoU index, and 0.04% in the OA index.

Table 3. Results of the WHU Building dataset comparison experiment.

Method	Precision	Recall	F1	Kappa	IoU	OA
FCN	96.29	96.44	96.36	92.73	93.13	98.56
U-Net	95.92	94.74	95.32	90.63	91.29	98.17
U-Net++	96.65	96.91	96.78	93.56	93.87	98.72
Swin-UNet	95.76	95.22	95.49	90.98	91.58	98.23
ACC-UNet	96.26	96.13	96.19	92.39	92.82	98.50
CSC-UNet	96.59	97.06	96.82	93.64	93.95	98.74
UCTransNet	96.90	97.24	97.16	94.14	94.40	98.84
DTA-UNet	96.61	**97.47**	97.03	94.07	94.33	98.82
baseline	96.41	96.85	96.62	93.25	93.69	98.65
RDAU-Net (Ours)	**97.28**	97.06	**97.17**	**94.34**	**94.59**	**98.88**

The bold font represents the optimal values.

The visualization results of each model for the WHU Building dataset are shown in Figure 11. In scenario A, the only models that can eliminate the noise in the topmost red box and do not miss detecting the small target buildings in the remaining two red boxes are the FCN and our model, but our model has higher segmentation accuracy. In scenario B, only ACC-UNet and our model can effectively segment the interior details of the building in the topmost red box, the edge details of the building in the left red box, and the edge details of the building in the bottom left red box. In scenario C, only our model does not miss detecting the small target building in the bottom left red box and has the highest overall segmentation accuracy among all models. In scenarios D and E, our model outperforms the other models in terms of segmentation completeness and edge details for large target buildings, and slightly improves the segmentation accuracy for small target buildings.

Overall, RDAU-Net also outperforms the other models on the WHU Building dataset, which has more than 4000 training samples, which results in good segmentation performance for all the compared models, meaning that the difference in performance between different models is also narrowed by the rich training samples. Nevertheless, RDAU-Net still achieves the best performance in five of the six metrics. It is worth mentioning that, in Scenarios A and C, only RDAU-Net does not omit the segmentation of small-scale buildings. In addition, for the large-scale building objective, it was able to maintain the best building

interior integrity and edge detail. Therefore, this result again reflects that reducing the loss during the feature transfer and reducing the semantic gap can bring about an improvement in building segmentation accuracy, as the segmentation results of our model more closely reflect the real buildings, with internal integrity and well-defined edges.

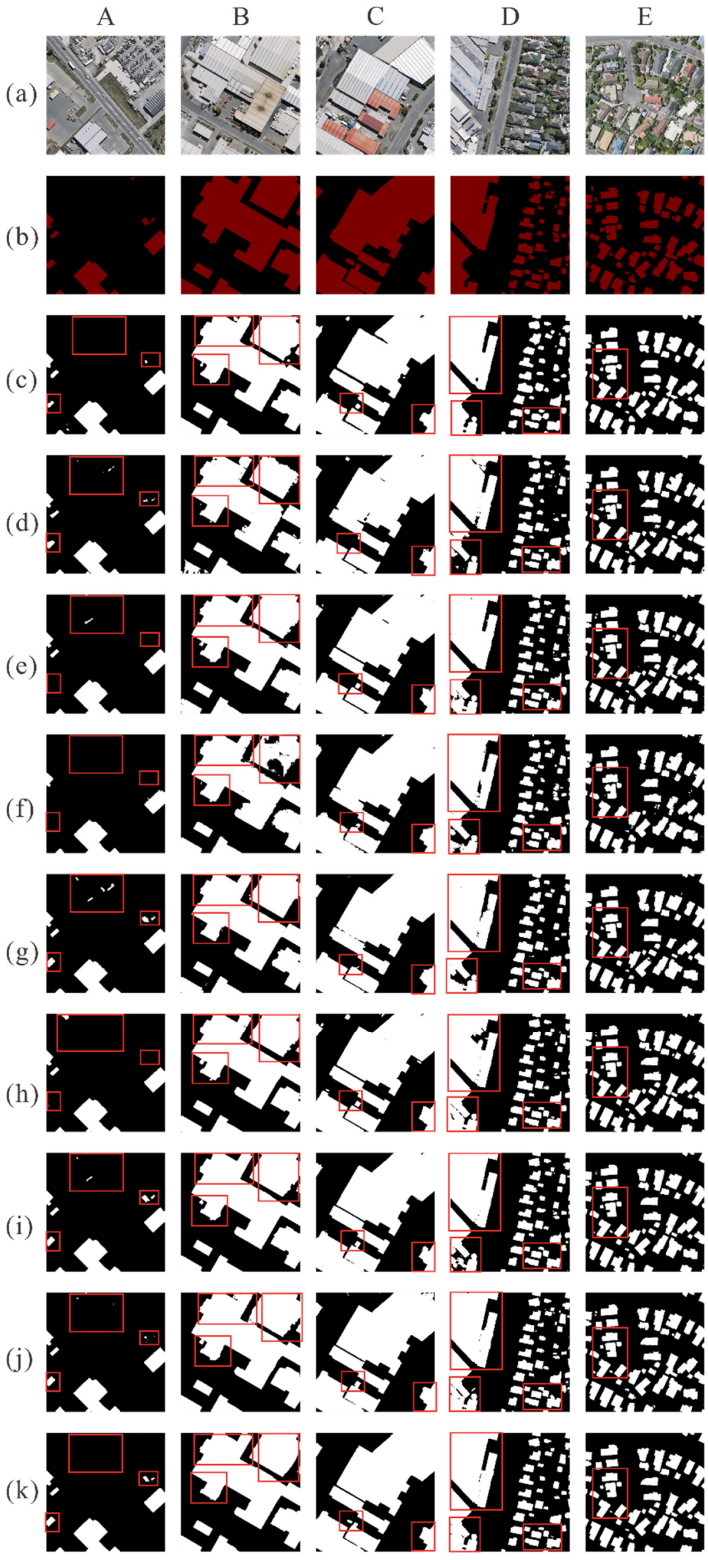


Figure 11. Visualization of the results of comparative experiments on the WHU Building dataset. (**a**) Image. (**b**) Ground truth. (**c**) FCN. (**d**) U-Net. (**e**) U-Net++. (**f**) Swin-UNet. (**g**) ACC-UNet. (**h**) CSC-UNet. (**i**) UCTransNet. (**j**) DTA-UNet. (**k**) RDAU-Net.

6. Conclusions

In the present study the RDAU-Net state-of-the-art approach for semantic segmentation of buildings in remote-sensing images is proposed. The key advantages of this approach include that it eliminates the negative effects of water body specular reflection and boat characteristics on the segmentation of buildings in river and lake management work, mitigates the omission of buildings at different scales, and improves the segmentation accuracy of buildings. In addition, we have constructed the first building dataset, the HRI Building dataset, for river and lake monitoring work scenarios in China.

Compared to current state-of-the-art models, the model proposed herein can more effectively eliminate the effects of specular reflections from the water and the effects of boats with similar characteristics to the buildings in the image. Meanwhile, our model obtains better performance in each evaluation metric for both the HRI Building dataset and WHU Building dataset. More specifically, the current building semantic segmentation models cannot segment the specular reflection of water and boats well, and they are not good for the edge detail segmentation of buildings. However, our experimental results show that RDAU-Net can solve the above issues. We designed our model to include RDSC to reduce the loss of feature information in the down-sampling process in the decoding stage, which can reduce the impact of complex building shapes and building scale differences on the segmentation results; by designing the MCCT module to enhance the relevance of multi-scale features and reduce the semantic gap between the features, the model can further enhance the internal integrity of the segmented buildings and exhibits a better ability to portray the internal details; by designing the DCF module to fuse the semantically inconsistent same-scale features in the decoding stage, it can mitigate the effect of complex backgrounds on the semantic segmentation of buildings and obtain an improved ability of interior and edge detail portrayal.

Our research is dedicated to better-applying building semantic segmentation to the supervision of rivers and lakes to reduce safety hazards. For now, the HRI Building dataset is not as large as large public datasets such as the WHU building dataset, and there may be some redundancy in the overall structure of our model. In addition, the structure of our proposed model is more complex than traditional building segmentation algorithms, which may affect the efficiency of the large number of building segmentation requirements in application scenarios. Therefore, in the next step, we plan to enrich the HRI Building dataset to provide more experimental data for scholars conducting related research, and at the same time lighten the model to improve its processing efficiency. How to better apply semantic segmentation models to practical work such as river and lake regulation is an important research direction that requires continuous research now and in the future.

Author Contributions: Conceptualization, Y.W. (Yipeng Wang) and Y.S.; methodology, Y.W. (Yipeng Wang), D.W., T.X. and Y.B.; data curation, D.W., Y.S., W.L. and Y.W. (Yihong Wang); writing—original draft preparation, Y.W. (Yipeng Wang); writing—review and editing, Y.W. (Yipeng Wang), T.X., G.P.P. and Y.B. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Major Science and Technology Program of the Ministry of Water Resources of China (SKS-2022072), the Water Science and Technology Project of Jiangsu Province (2023022), the National Natural Science Foundation of China (Grant No. 42377046), and Research Funds of Jiangsu Hydraulic Research Institute (2023z034, 2024z007).

Data Availability Statement: The author can supply the datasets related to this article upon request.

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

References

1. Cid, N.; Erős, T.; Heino, J.; Singer, G.; Jähnig, S.C.; Cañedo-Argüelles, M.; Bonada, N.; Sarremejane, R.; Mykrä, H.; Sandin, L.; et al. From meta-system theory to the sustainable management of rivers in the Anthropocene. *Front. Ecol. Environ.* **2022**, *20*, 49–57. [CrossRef] [PubMed]
2. Su, X.; Fan, Y.; Wen, C. Systematic coupling and multistage interactive response of the urban land use efficiency and ecological environment quality. *J. Environ. Manag.* **2024**, *365*, 121584. [CrossRef] [PubMed]
3. Huang, X.; Shen, J.; Li, S.; Chi, C.; Guo, P.; Hu, P. Sustainable flood control strategies under extreme rainfall: Allocation of flood drainage rights in the middle and lower reaches of the yellow river based on a new decision-making framework. *J. Environ. Manag.* **2024**, *367*, 122020. [CrossRef] [PubMed]
4. Huang, X.; Hua, W.; Dai, X. Performance Evaluation of Watershed Environment Governance—A Case Study of Taihu Basin. *Water* **2022**, *14*, 158. [CrossRef]
5. Xue, H.; Liu, K.; Wang, Y.; Chen, Y.; Huang, C.; Wang, P.; Li, L. MAD-UNet: A Multi-Region UAV Remote Sensing Network for Rural Building Extraction. *Sensors* **2024**, *24*, 2393. [CrossRef]
6. Notarangelo, N.M.; Mazzariello, A.; Albano, R.; Sole, A. Comparing Three Machine Learning Techniques for Building Extraction from a Digital Surface Model. *Appl. Sci.* **2021**, *11*, 6072. [CrossRef]
7. Chen, R.; Li, X.; Li, J. Object-Based Features for House Detection from RGB High-Resolution Images. *Remote Sens.* **2018**, *10*, 451. [CrossRef]
8. Tamilarasi, R.; Prabu, S. Automated building and road classifications from hyperspectral imagery through a fully convolutional network and support vector machine. *J. Supercomput.* **2021**, *77*, 13243–13261. [CrossRef]
9. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
10. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
11. 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]
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
13. Liu, L.W.; Huang, Z.; Lu, K.Y.; Wang, Z.X.; Liang, Y.M.; Lin, S.Y.; Ji, Y.H. UJAT-Net: A U-Net Combined Joint-Attention and Transformer for Breast Tubule Segmentation in H&E Stained Images. *IEEE Access* **2024**, *12*, 34582–34591. [CrossRef]
14. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G.; Zhang, D. DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 4005615. [CrossRef]
15. 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.
16. 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]
17. 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]
18. Liu, H.; Luo, J.; Huang, B.; Yang, H.; Hu, X.; Xu, N.; Xia, L. Building Extraction based on SE-Unet. *J. Geo-Inf. Sci.* **2019**, *21*, 1779–1789.
19. Lu, K.; Sun, Y.; Ong, S.H. Dual-Resolution U-Net: Building Extraction from Aerial Images. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 489–494.
20. Wang, S.; Hou, X.; Zhao, X. Automatic Building Extraction From High-Resolution Aerial Imagery via Fully Convolutional Encoder-Decoder Network With Non-Local Block. *IEEE Access* **2020**, *8*, 7313–7322. [CrossRef]
21. Shunping, J.I.; Shiqing, W.E.I. Building extraction via convolutional neural networks from an open remote sensing building dataset. *Acta Geod. Cartogr. Sin.* **2019**, *48*, 448.
22. Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; Liao, Y.J.N. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing* **2022**, *493*, 626–646. [CrossRef]
23. 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]
24. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. [CrossRef]

25. 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 Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 833–851.
26. Yu, M.; Chen, X.; Zhang, W.; Liu, Y. AGs-Unet: Building Extraction Model for High Resolution Remote Sensing Images Based on Attention Gates U Network. *Sensors* **2022**, *22*, 2932. [CrossRef] [PubMed]
27. Zhong, L.; Lin, Y.; Su, Y.; Fang, X. Improved U-Net Network Segmentation Method for Remote Sensing Image. In Proceedings of the 2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Beijing, China, 3–5 October 2022; pp. 1034–1039.
28. Liu, Y.; Chen, D.; Ma, A.; Zhong, Y.; Fang, F.; Xu, K. Multiscale U-Shaped CNN Building Instance Extraction Framework With Edge Constraint for High-Spatial-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 6106–6120. [CrossRef]
29. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
30. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 9992–10002.
31. Meng, X.; Yang, Y.; Wang, L.; Wang, T.; Li, R.; Zhang, C. Class-Guided Swin Transformer for Semantic Segmentation of Remote Sensing Imagery. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6517505. [CrossRef]
32. Xu, Z.; Geng, J.; Jiang, W. MMT: Mixed-Mask Transformer for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5605116. [CrossRef]
33. Chen, Y.; Dong, Q.; Wang, X.; Zhang, Q.; Kang, M.; Jiang, W.; Wang, M.; Xu, L.; Zhang, C. Hybrid Attention Fusion Embedded in Transformer for Remote Sensing Image Semantic Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 4421–4435. [CrossRef]
34. Xiao, X.; Guo, W.; Chen, R.; Hui, Y.; Wang, J.; Zhao, H. A Swin Transformer-Based Encoding Booster Integrated in U-Shaped Network for Building Extraction. *Remote Sens.* **2022**, *14*, 2611. [CrossRef]
35. Wang, T.; Xu, C.; Liu, B.; Yang, G.; Zhang, E.; Niu, D.; Zhang, H. MCAT-UNet: Convolutional and Cross-Shaped Window Attention Enhanced UNet for Efficient High-Resolution Remote Sensing Image Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 9745–9758. [CrossRef]
36. Liu, B.; Li, B.; Sreeram, V.; Li, S. MBT-UNet: Multi-Branch Transform Combined with UNet for Semantic Segmentation of Remote Sensing Images. *Remote Sens.* **2024**, *16*, 2776. [CrossRef]
37. Ding, R.X.; Xu, Y.H.; Liu, J.; Zhou, W.; Chen, C. LSENet: Local and Spatial Enhancement to Improve the Semantic Segmentation of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 7506005. [CrossRef]
38. Dimitrovski, I.; Spasev, V.; Loshkovska, S.; Kitanovski, I. U-Net Ensemble for Enhanced Semantic Segmentation in Remote Sensing Imagery. *Remote Sens.* **2024**, *16*, 2077. [CrossRef]
39. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic convolution: Attention over convolution kernels. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11030–11039.
40. Feng, M.; Sun, X.; Dong, J.; Zhao, H. Gaussian Dynamic Convolution for Semantic Segmentation in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5736. [CrossRef]
41. Hou, J.; Guo, Z.; Wu, Y.; Diao, W.; Xu, T. BSNet: Dynamic Hybrid Gradient Convolution Based Boundary-Sensitive Network for Remote Sensing Image Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5624022. [CrossRef]
42. Wang, X.; Zhang, Y.; Lei, T.; Wang, Y.; Zhai, Y.; Nandi, A.K. Dynamic Convolution Self-Attention Network for Land-Cover Classification in VHR Remote-Sensing Images. *Remote Sens.* **2022**, *14*, 4941. [CrossRef]
43. Chen, G.P.; Li, L.; Zhang, J.X.; Dai, Y. Rethinking the unpretentious U-net for medical ultrasound image segmentation. *Pattern Recognit.* **2023**, *142*, 109728. [CrossRef]
44. Ibtehaz, N.; Rahman, M.S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw.* **2020**, *121*, 74–87. [CrossRef]
45. Wang, H.N.; Cao, P.; Wang, J.Q.; Zaiane, O.R. UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer. In Proceedings of the 36th AAAI Conference on Artificial Intelligence/34th Conference on Innovative Applications of Artificial Intelligence/12th Symposium on Educational Advances in Artificial Intelligence, Electr Network, Online, 22 February–1 March 2022; pp. 2441–2449.
46. Ming, Q.; Xiao, X. Towards Accurate Medical Image Segmentation With Gradient-Optimized Dice Loss. *IEEE Signal Process. Lett.* **2024**, *31*, 191–195. [CrossRef]
47. Wang, S.y.; Qu, Z.; Gao, L.y. Multi-Spatial Pyramid Feature and Optimizing Focal Loss Function for Object Detection. *IEEE Trans. Intell. Veh.* **2024**, *9*, 1054–1065. [CrossRef]

48. Connor, R.; Dearle, A.; Claydon, B.; Vadicamo, L. Correlations of Cross-Entropy Loss in Machine Learning. *Entropy* **2024**, *26*, 491. [CrossRef]
49. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
50. Ibtehaz, N.; Kihara, D. *ACC-UNet: A Completely Convolutional UNet Model for the 2020s*; Springer Nature: Cham, Switzerland, 2023; pp. 692–702.
51. Tang, H.; He, S.; Yang, M.; Lu, X.; Yu, Q.; Liu, K.; Yan, H.; Wang, N. CSC-Unet: A Novel Convolutional Sparse Coding Strategy Based Neural Network for Semantic Segmentation. *IEEE Access* **2024**, *12*, 35844–35854. [CrossRef]
52. Li, Y.; Yan, B.; Hou, J.; Bai, B.; Huang, X.; Xu, C.; Fang, L. UNet based on dynamic convolution decomposition and triplet attention. *Sci. Rep.* **2024**, *14*, 271. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.