IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 June 2024, accepted 8 August 2024, date of publication 12 August 2024, date of current version 21 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3441837

RESEARCH ARTICLE

EA U^2-Net: An Efficient Building Extraction Algorithm Based on Complex Background Information

FEIFEI XIE[1], **MINGZHE YI**[1], **ZHILING HUO**[2], **LIN SUN**[1], **JINGYU ZHAO**[1], **ZHIPENG ZHANG**[1], **JINPENG CHEN**[1], **JINRUI ZHANG**[1], **AND FANGRUI CHEN**[1]

[1]College of Geodesy and Geomatics, Shandong University of Science and Technology, Qingdao 266590, China
[2]Beijing City Interface Technology Company Ltd., Beijing 100032, China

Corresponding authors: Feifei Xie (xff@sdust.edu.cn), Mingzhe Yi (202283020118@sdust.edu.cn), and Jingyu Zhao (202182020030@sdust.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 42171439, in part by Shandong Provincial Natural Science Foundation Project under Grant ZR2021MD026, in part by Shandong Higher School Science and Technology Project under Grant J18KA214, and in part by Shandong Province Innovation Training Program under Grant 202505.

ABSTRACT Effective extraction of building edge information based on high-resolution remote sensing images is the basis for efficient urban 3D modeling. Existing building extraction methods still have some problems, such as an uncertain segmentation scale, effective feature selection, and sample selection. In this paper, we propose a practical building extraction method based on convolutional network edge-enhanced attention U^2-Net (EA U^2-Net) to accurately achieve multi-scale extraction of buildings from remote sensing imagery. First, the U^2-Net is used as the backbone network for building extraction because each stage of the network is filled by residual U-block (RSU), and the network can better aggregate multi-scale features. Second, the building edge feature map is introduced into the generation network to compensate for the problems of insufficient extracted building edge features and loss of detail. Finally, the convolutional block attention module is used to achieve effective feature extraction of buildings. We performed the experiment on the WHU building dataset, and the experimental results showed that the EA U^2-Net model has significantly improved the ability to extract buildings, with an accuracy of 96.30%, a recall rate of 94.91%, f1 of 95.26%, and iou of 91.57%. This proves that EA U ^{2} - Net can achieve better remote sensing image-building segmentation results. Finally, in view of the problem that the deep learning network relies on training samples, this study examined the influence of the number of building samples, sample purity, and sample resolution on the effect of building extraction. The results confirmed that reasonable sample parameter settings can improve the target extraction accuracy and the optimal sample parameter combination was verified in this experiment.

INDEX TERMS Building extraction, EA U^2-Net, high-resolution remote sensing imagery, sample parameter study.

I. INTRODUCTION

The spatial distribution of buildings plays a pivotal role in various human activities, including economic development [1], urban planning [2], disaster prevention and mitigation [3], [4], as well as national defense security. The rapid development of high-resolution remote sensing technology has provided abundant data resources for building extraction research [5]. Compared with medium and low-resolution remote sensing images, high-resolution remote sensing images have finer ground features, more specific geometric contours, and texture features of ground features and landscapes. However, the increase in details also introduces a large amount of noise information, making the spectral characteristics of ground

The associate editor coordinating the review of this manuscript and approving it for publication was Kan Liu.

objects more complex. Moreover, the phenomena of ''different body with same spectrum'' or ''same body with different spectrum'' have increased significantly, which greatly limits the improvement of building extraction accuracy and restricts the large-scale application of high-resolution remote sensing imagery in urban information extraction.

Traditional research on building extraction typically relies on comprehensive information such as thresholds, edges, and regions. Threshold-based segmentation methods, such as Li et al. [6], proposed an adaptive global threshold method to solve the problems of over-segmentation and under-segmentation. Wu et al. [7] measured the saliency of building targets in images of different scales through multi-feature fusion and combined it with the Otsu algorithm to automatically obtain thresholds and achieve automatic detection of buildings. Edge-based segmentation methods mainly target the texture, grayscale, shape, and other features of buildings. Qu et al. [8] used the Sobel operator and linear support vector machine to classify features for building detection. Cui et al. [9] extracted buildings based on the grayscale and geometric features of the image, and completed the extraction by detecting the spatial distribution characteristics of the basic elements in the image. Region-based segmentation methods utilize the spatial relationships between segmented regions. Izadi et al. [10] used the spatial relationship between segmented regions and their regional features to identify potential regions in multi-layer segmented images and used tree structures to describe these relationships for accurate extraction of buildings. Wegner et al. [11] proposed an irregular pattern combining optical and interferometric synthetic aperture radar (SAR) functions for building detection, using a Conditional Random Domain (CRF) framework to study the advantages of irregular graphic structures in building extraction. Tao et al. [12] integrated object-oriented thinking into building segmentation methods and effectively extracted buildings using multi-feature fusion methods and Bayesian criteria. Although these comprehensive methods have improved the mining of buildings, they still face challenges. The selection of thresholds and the design of image features often depend on professional knowledge and experience. Faced with complex urban land cover images, the comprehensive ability of these features is weak, making it difficult to meet the practical requirements of building extraction from high-resolution remote sensing images.

With the rapid development of machine learning technology, deep learning algorithms combine low-level features with multi-layer neurons to form abstract high-level features [13] (attribute categories or features), combine feature extraction with classifier modeling, reduce the need for manual feature design, and have achieved significant results in the field of image semantic segmentation [14], [15]. In 2015, Jonathan Long [16] first proposed a fully convolutional neural network (FCN), laying the foundation for the application of deep learning algorithms in image segmentation. However, FCNs are not sensitive enough to details and do not consider pixel relationships. Inspired by FCNs, Olaf Ronneberger [17] proposed U-Net, a strictly symmetric encoder-decoder structure that can achieve more accurate resolution, and introduce corresponding scale feature information into the upsampling process through skip connections to obtain finer segmentation results. Due to the progressiveness of U-Net, many scholars have made improvements based on U-Net to improve the ability to extract buildings [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [54], [55]. Although the above methods can effectively improve the accuracy of building extraction, remote sensing images have the characteristics of rich semantic information and diverse target categories compared to ordinary photos. The colors and shapes of buildings are also different, which cannot fully utilize the spatial features of remote sensing images. Therefore, there are still issues of boundary blurring and information loss in the extraction results.

Compared with U-Net, U^2-Net uses an RSU instead of a convolutional layer, which can capture more contextual information. Wei et al. [28] directly used U^2-Net to extract building, and the experimental results showed that compared with segmentation networks Segnet [29], U-Net [17], Deeplab [43], and FCN [16], and edge detection networks RCF [30], HED [31], and DexiNed [32], U^2-Net can obtain higher precision and more precise location. Zhou et al. [33] used U^2-Net to extract dense low-rise buildings to achieve large-scale mapping of urban buildings and achieved good results, verifying the effectiveness of U^2-Net in building extraction. However, the complex network structure of U^2-Net achieves a deeper architecture by sacrificing high-resolution feature maps, and it is easy to lose important information at low frequencies, especially in the face of small-target detection [34]. When extracting buildings, the model does not pay enough attention to small-area buildings, which are prone to missed detection, and ignore edge information, resulting in blurred edges of buildings.

However, deep learning methods rely heavily on large-scale annotated data, and high-quality annotated data are mostly manually operated, resulting in a large workload and low efficiency. This defect greatly limits the application of deep learning methods in practical image recognition tasks, and hence, many scholars have generated fake data through rotation, cropping, scaling, etc. [35], generative adversarial networks [36], transfer learning [37], and other methods to enhance the dataset. These methods have effectively expanded the dataset. Sufficient and rich training samples can enable the deep learning network to fully learn the characteristics of the samples; however, blindly increasing the training samples may lead to overfitting of the network model and reduce the training effect of the model, which in turn affects the segmentation accuracy in the testing phase to a certain extent [38]. In addition to quantitative parameters, compared with ordinary photos, remote sensing images have the characteristics of target category diversity, feature information variability, and complexity of interference factors,

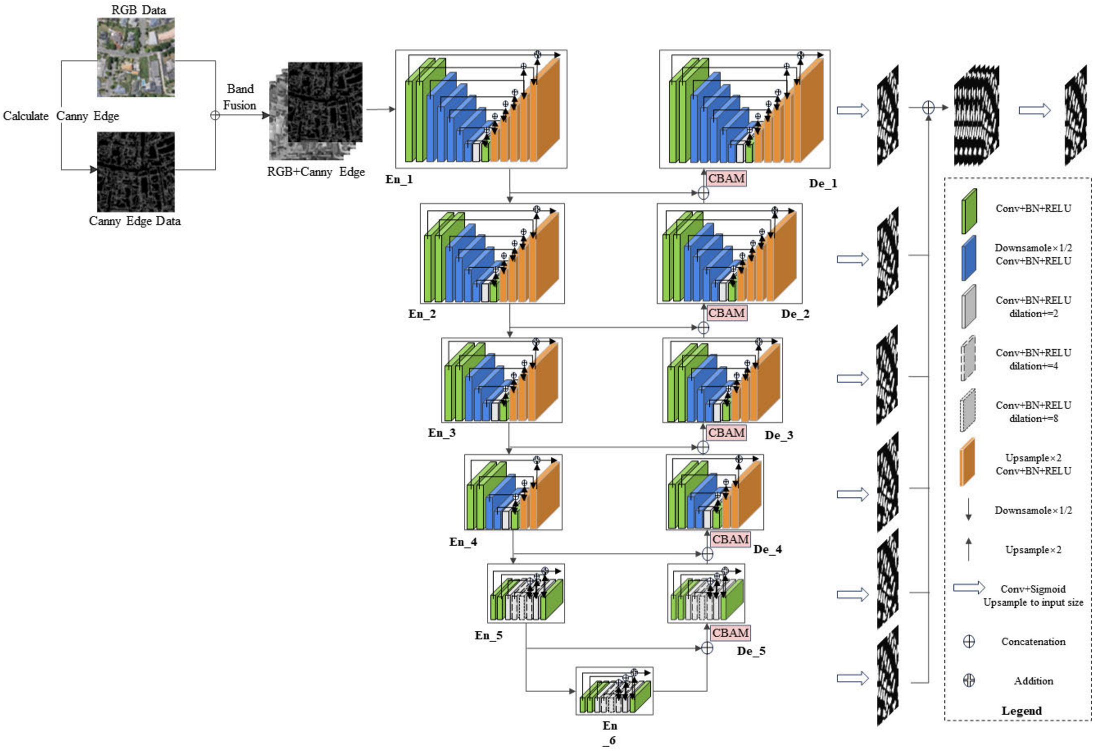

FIGURE 1. EA U^2-Net structure. The network uses U^2-Net as the main structure, introduces building edge features in the data input phase, and introduces the convolutional block attention module in the decoding phase of the network.

which will also affect the extraction of buildings. Therefore, it is necessary to study the influence parameters of building samples [39].

To address the above problems, this paper introduces the deep learning network U^2-Net for the purpose of building extraction from high-resolution remote sensing images and proposes the EA U^2-Net for the characteristics of high-resolution images and building features to achieve accurate and effective building extraction. At the same time, we study the influence of sample parameters on the deep learning model, including the number of building samples, sample purity, and sample resolution, to explore the minimum demand for remote sensing data from different sources for building extraction; it is also intended to be weakly supervised learning under sparse annotation [40] provides theoretical support. The main contributions of this study are as follows:

(1) To solve the insufficient learning of edge features in the process of building extraction by U^2-Net, enhancing the learning ability of the deep learning network for edge features according to the edge features of buildings is obtained through Canny edge feature detection, which can compensate for edge blurring in the segmentation results.

(2) To make the network more focused on building features and suppress non-building features, the convolutional block attention module is introduced into U^2-Net to enhance the connection between each feature in space and channel [45].

(3) To verify the effectiveness of this paper's method, we compare our method with other building extraction methods on the WHU building dataset, and the results show that our method is more effective than other methods.

(4) The deep learning method relies heavily on the training samples. To understand the impact of different sample parameters on the deep-learning model, this study considers building extraction as an example to study the impact of the number of samples, sample purity, and sample resolution on the segmentation results.

The rest of this paper is organized as follows. In Section II, the proposed method is described in detail, including the basic structure of the network, the dataset enhancement method, and the convolutional block attention module. Section III presents the details of the experiment, including the datasets, experimental platform, evaluation metrics, and experimental results, Section IV presents the conclusions.

II. METHODOLOGY

EA U^2-Net uses U^2-Net as the main backbone network. Before inputting the training data into the network, the edges and contours of the buildings in the image are effectively detected by using the canny edge detection algorithm [51], which increases the ability of the U^2-Net network to perceive the details and edges, and thus improves the boundary quality of the segmentation results. Moreover, combining the convolutional block attention module in the decoding stage of the network can help the network better focus on the region of interest and suppress irrelevant background information, which can help to reduce the interference of background noise on the segmentation results, and also increase the network's ability to perceive the local image regions, which can better capture the local features and textures [52] The combination of the canny edge detection algorithm and the attention module enhances the interpretability of the network and improves the robustness of the network. The network structure of EA U^2-Net is shown in FIGURE 1.

A. EXTRACTION OF EDGE INFORMATION

The training data for building extraction usually have only RGB three-band information, and the information richness is limited. Canny edge detection [44] has a strong adaptive ability and capability to remove interference, identify as many actual edges as possible, and provide edge information for building training data. Therefore, this study uses the fusion data of the original RGB band data and the Canny edge detection result data as the training datasets.

The input single-channel grey-scale image is first smoothed using a Gaussian filter, where the standard deviation of the Gaussian filter is 1.4 by default, and the filter kernel size is 5. The gradient values and gradient directions are then calculated for each pixel of the returned smoothed image. After that, it is processed by applying non-maximum suppression to the gradient values and gradient directions. Finally, the image is subjected to a double thresholding method to compute the edges, where the small threshold controls the edge connectivity and the large threshold controls the initial segmentation of the strong edges and outputs the binary image.

The calculation steps of Canny edge detection are as follows:

Step 1: Using Gaussian filtering to complete the image smoothing, the formal description of Gaussian filtering is shown in (1).

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{(\frac{x^2+y^2}{2\sigma^2})} \tag{1}$$

Gaussian filtering was used to remove noise. The Gaussian kernel used in Gaussian filtering is a Gaussian function with two dimensions, x and y, and the standard deviation in these two dimensions is the same.

Step 2: Calculating the pixel gradient using the Sobel operator. The calculations of the Sobel operator are shown in (2)-(4).

$$G_x = S_x * I \tag{2}$$

$$G_x = S_y * I \tag{3}$$

$$G_{xy} = \sqrt{G_{(x^2)} + G_{(y^2)}} \tag{4}$$

where S_x and S_y represent the pixel gradient matrices in the x and y directions, respectively, I represents the grayscale image matrix, and $*$ represents the cross-correlation operation.

Step 3: Non-maximum suppression.

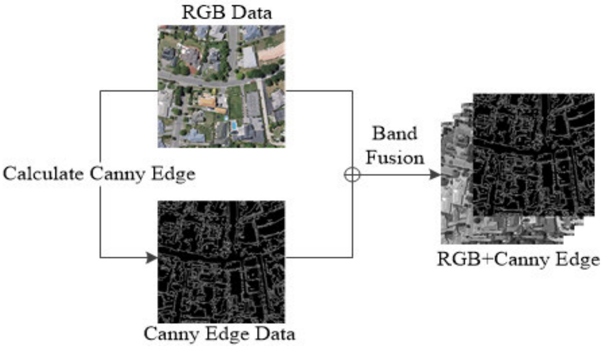

FIGURE 2. The flowchart of dataset generation.

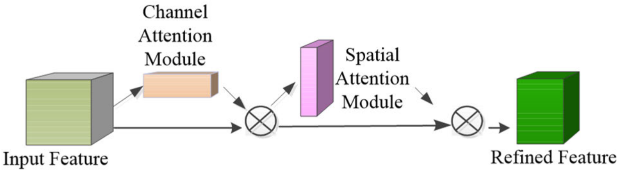

FIGURE 3. The CBAM structure.

The purpose of non-maximum pixel gradient suppression is to eliminate spurious effects caused by edge detection. The basic method is to compare the gradient strength of the current pixel with the gradient strength of adjacent pixels along the positive and negative gradient directions and maintain the maximum value as the edge point.

Step 4: Threshold hysteresis processing.

We defined high and low thresholds. Pixels whose gradient strength is lower than the low threshold are suppressed. Conversely, pixels higher than the high threshold are defined as strong edges and are reserved as edge points, and those between the high and low thresholds are defined as weak edges and are left for further processing.

Step 5: Isolate weak edge suppression.

The edge is judged according to the connection between the weak edge pixel and strong edge. As long as one of the neighboring pixels in the weak edge is a strong edge pixel, the weak edge can be retained as a strong edge, that is, a real edge point.

The flowchart of dataset generation is shown in FIGURE 2. First, the RGB data are used to generate the Canny edge detection data, and then the RGB and Canny edge detection data are fused to obtain the fusion data.

B. CONVOLUTIONAL BLOCK ATTENTION MODULE

The convolutional block attention module (CBAM) allows the network to pay more attention to the target to be detected, focus on important features, and suppress unimportant features [47] The addition of CBAM to U^2-Net can improve the connection between the channels and spaces of each feature. During the training process, the CBAM can serially generate attention feature map information in the channel and space dimensions and then multiply the two types of feature map information with the input feature map to obtain the final feature map. The CBAM structure is shown in FIGURE 3,

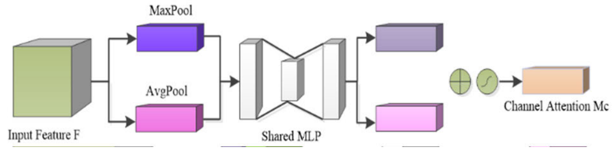

FIGURE 4. The CAM structure.

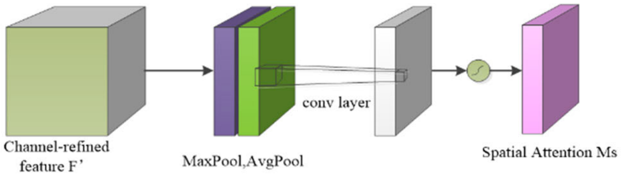

FIGURE 5. The SAM structure.

and the process is shown in (5)-(6):

$$F' = M_C(F) \otimes F \tag{5}$$

$$F'' = M_S(F') \otimes F' \tag{6}$$

where ⊗ presents element-wise multiplication, F denotes the input feature map, M_c denotes the channel attention map, M_S denotes the spatial attention map, F'' denotes the final refined output.

1) CHANNEL ATTENTION MODULE

The channel attention module [48] (CAM) firstly performs global and average pooling on the input feature map to obtain two feature vectors, then applies two fully connected operations to generate two two-dimensional vectors. The corresponding elements are added. The weight containing the channel information is multiplied by the input feature map to obtain the feature map weighted by channel attention. The CAM structure is shown in FIGURE 4, the channel attention is shown in (7):

$$M_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{7}$$

2) SPATIAL ATTENTION MODULE

The spatial attention module (SAM), similar to the channel attention module, first performs global pooling and average pooling on the input feature map containing channel information to obtain two compressed feature maps and then combines these two features. The graph is superimposed, and a convolution with a convolution kernel of 7 × 7 is performed to adjust the number of channels. The obtained spatial weight information is multiplied by the input feature map to obtain a feature map containing spatial and channel information. The SAM structure is shown in FIGURE 5, the spatial attention is computed in (8):

$$M_S(F) = \sigma(f^{7\times 7}([AvgPool(F); MaxPool(F)])) \tag{8}$$

C. U^2-NET

Qin et al. [41] proposed a U^2-Net composed of a two-level nested U-shaped structure, which includes a multi-scale RSU in the extraction stage and a U-block connected to the RSU module. The network consists of a six-level encoder and a five-level decoder, each of which is filled with a residual block RSU, so that the network can better extract intra-stage multi-scale features and aggregate inter-stage multi-level features. In the encoding stage, the first four stages pass through the RSU modules with layers 7, 6, 5, and 4, and stages 5 and 6 are RSU modules with dilated convolutions. Each decoder stage takes as input the concatenation of the upsampled feature maps from the previous stage and the feature maps from the symmetric encoder stage and finally fuses the feature maps obtained from each stage as the output. When using a small sample dataset, we can reduce the risk of overfitting by using a pre-trained model through migration learning, whereas in this paper a larger dataset is used to avoid the risk of overfitting due to a small dataset, so the network does not require any pre-training, and can be trained from scratch to achieve very competitive performance, increasing the depth of the network to obtain a high-resolution feature map without increasing the memory and computational cost.

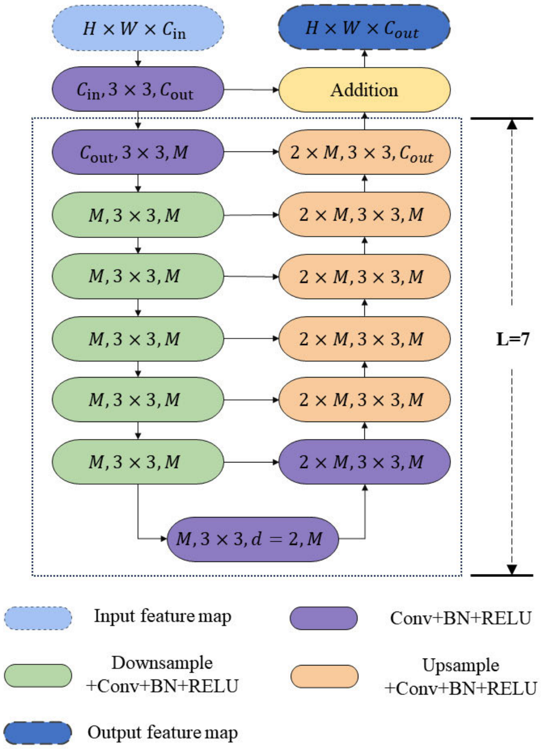
$H \times W \times C_{in}$
$H \times W \times C_{out}$
$C_{in}, 3\times 3, C_{out}$
Addition
$C_{out}, 3\times 3, M$
$2\times M, 3\times 3, C_{out}$
$M, 3\times 3, M$
$2\times M, 3\times 3, M$
$M, 3\times 3, M$
$2\times M, 3\times 3, M$
$M, 3\times 3, M$
$2\times M, 3\times 3, M$
$M, 3\times 3, M$
$2\times M, 3\times 3, M$
$M, 3\times 3, M$
$2\times M, 3\times 3, M$
$M, 3\times 3, d=2, M$
L=7
Input feature map
Conv+BN+RELU
Downsample +Conv+BN+RELU
Upsample +Conv+BN+RELU
Output feature map

FIGURE 6. Residual U-block RSU structure.

The residual U-block RSU consists of three parts: a convolutional layer that converts the input feature map into an intermediate feature map, a U-shaped structure that extracts multi-scale features, and a residual connection layer that fuses local and multi-scale features. and this design enables the network to extract multiple scales of features directly from each residual U-block RSU [42] as shown in FIGURE 6. The left side is the encoding stage, and the right side is the decoding stage. The larger the number of layers in the

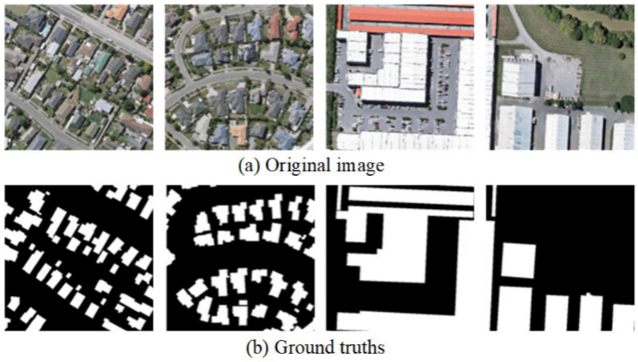

FIGURE 7. Sample image from the WHU building dataset. (a) Original image. (b) Ground truth.

encoder, the deeper the RSU structure, and more pooling operations will increase the receptive field and enrich the extraction of global features [43]. L is the number of layers in the encoder, C_{in}, C_{out} denote the input and output channels, and M denotes the number of channels inside the RSU.

D. LOSS FUNCTION

To solve the problems of gradient disappearance and low convergence speed in the training process, the backbone network was supervised by deep supervision [31], and the training loss function was calculated using (9):

$$L = \sum_{m=1}^{M} \omega_{side}^{(m)} l_{side}^{(m)} + \omega_{fuse} l_{fuse} \tag{9}$$

where $l_{side}^{(m)}$ is the loss of each stage feature map, and l_{fuse} is the loss of the final output feature map. $\omega_{side}^{(m)}$ and ω_{fuse} are the loss weights. For each term l, we used the standard binary cross-entropy to calculate the loss, and the standard binary cross-entropy was calculated using (10):

$$l = -\sum_{(r,c)}^{(H,W)} [P_{G(r,c)} \log^{P_{S(r,c)}} + (1 - P_{G(r,c)}) \log^{(1-P_{S(r,c)})}] \tag{10}$$

where (r, c) is the pixel coordinates, and (H, W) is the height and width of the image. $P_{G(r,c)}$ and $P_{S(r,c)}$ are the pixel values of the ground truths and predicted feature map. The total loss L is minimized during training. In the testing process, l_{fuse} is output as the final feature map.

III. EXPERIMENTAL RESULT

A. DATASETS

The building sample dataset used in this study was the aerial image dataset in the WHU building dataset [48] (http://study.rsgis.whu.edu.cn/pages/download/). The original data are shown in FIGURE 7. The dataset contains approximately 22,000 individual buildings, and the images have an original ground spatial resolution of 0.075 m, down sampled to 0.3 m ground resolution, and cropped to 8189 tiles with 512 × 512 pixels.

FIGURE 8. Image-processed samples.

To further enhance the generalization ability of the model, this study also expanded the original dataset through image processing methods such as Gaussian noise and salt and pepper noise to increase the diversity of experimental samples. The sample after image processing is shown in FIGURE 8.

B. EVALUATION METRICS

Precision, recall, F1 score, and intersection-over-union (IoU) were used as evaluation metrics to evaluate the accuracy of building extraction. The calculation of the different evaluation metrics are as follows.

(1) Precision is expressed as the ratio of pixels, which is the ratio of the number of pixels for which a building is correctly predicted, to the number of pixels for all predicted buildings. The precision is calculated using (11):

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

(2) Recall is expressed as the ratio of classified pixels, which is the probability of correctly predicting a pixel in a building. Recall is calculated using (12):

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

(3) The F1 score takes into account both the accuracy and recall of the segmentation model. The calculation of the F1 score is given by (13):

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{13}$$

(4) IoU is the ratio of the intersection region to the concurrent region. The IoU is calculated using (14):

$$IoU = \frac{TP}{FP + TP + FN} \tag{14}$$

where TP stands for true positive, which is the image element that predicts the positive class as a positive class; that is, the real building pixels are predicted as the number of building pixels. TN stands for true negative, which is the image element whose negative class is predicted to be negative; that is, background pixels are predicted as the number of background pixels. FP is a false positive, which is the number of pixels that predict a negative class as a positive class; that is, the

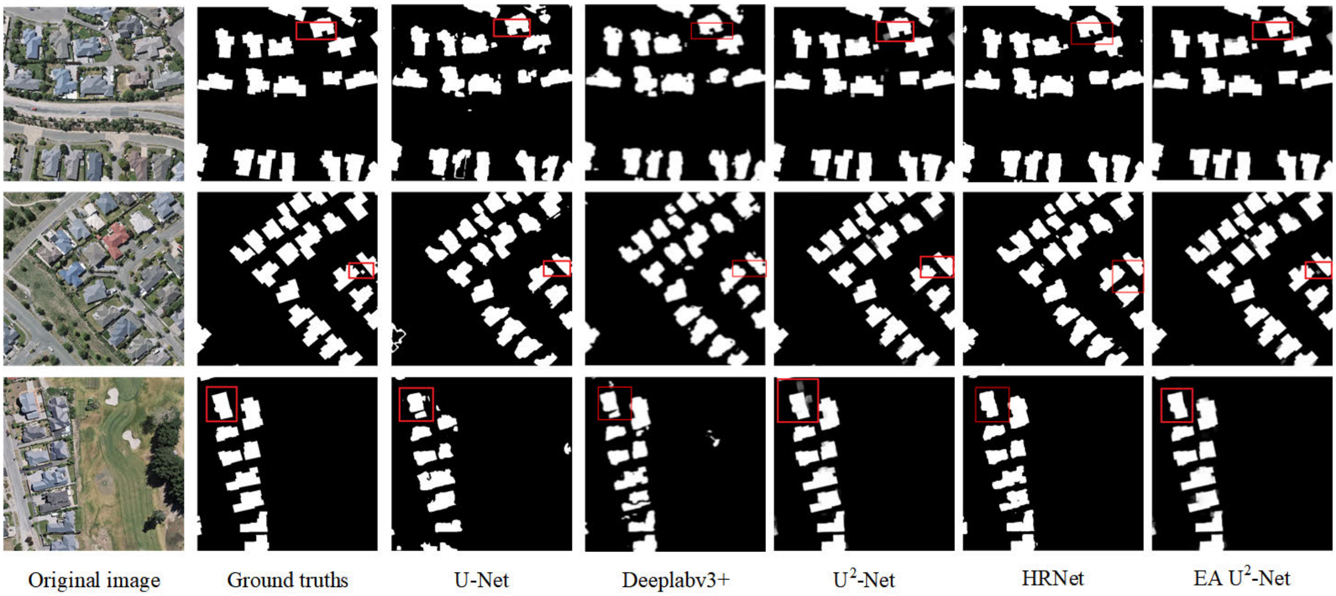

FIGURE 9. The building extraction results of different methods on the WHU building dataset.

background pixels are predicted as the number of building pixels. FN stands for false negative, which predicts a positive class as a negative class; that is, the number of real building pixels is predicted as background pixels.

C. IMPLEMENTATION DETAILS

The proposed EA U^2-Net adopts an Intel Core i7-12700H@2.30GHz 14-core processor, equipped with 16.0 GB memory, and an Nvidia GeForce RTX 3050 Ti 4 GB graphics card. In terms of software environment, we used Windows 10 Professional Edition 64-bit operating system, the programming language, Python, the CUDA11.3 version of the GPU computing platform and the cuDNN8.2.1 deep learning GPU acceleration library.

To ensure the objectivity of the experimental results, all experiments were optimized using the Adam algorithm [50], and the initial learning rate was set to 0.001. During the training process, two images were input into the model each time, the iteration round was 100, and the network training parameters were saved every 10 epochs.

D. COMPARISON TO DIFFERENT BUILDING EXTRACTION METHODS

To evaluate the algorithm performance on the WHU buildings dataset, we compared the proposed EA U^2-Net with other state-of-the-art methods, which includes the U-Net [18], U^2-Net [28], Deeplabv3+ [43], [52] and HRNet [53] with the same sample datasets (the number of samples was 2000). FIGURE 9 shows the building extraction results of different methods on the WHU buildings dataset. For large-area buildings, the prediction results of the three methods had different degrees of edge blurring, but the extraction results were generally accurate. For small-area buildings, there were certain missed detections and false detections. Among them, U-Net had the worst extraction effect, and the extraction of large-area buildings had the phenomenon of broken edges and incomplete extraction. There were also a large number of false detections and missed detections for small-area buildings. Secondly, HRNet and Deeplabv3+ extract better than U-Net, but there still exists a weakening of the edges of the building contours and the omission of small-area buildings. U^2-Net yielded better extraction results. The outline of buildings was more complete as well as accurate, and the missed and false detections of small-area buildings were also greatly improved, but blurred edges still existed. Compared with U-Net and U^2-Net, EA U^2-Net could capture more detailed information, distinguish the boundaries of buildings better, and extract small-area buildings more accurately. Overall, the outlines of the buildings were clearer, and the extraction was more complete, and accurate.

Table 1 presents the accuracy evaluation of the extraction results of different deep learning methods. It can be seen that EA U^2-Net achieved the best performance in various evaluation indicators, with a precision of 96.30%, a recall of 94.91%, an F1 of 95.62%, and an IoU of 91.57%. Compared with U-Net, Deeplabv3+, U^2-Net, and HRNet, the precision increased by 5.63%, 3.37%, 0.57%, and 0.82%, the recall increased by 14.40%, 7.00%, 0.94%, and 0.08%, F1 increased by 10.35%, 5.28%, 0.78%, and 0.47%, and IoU increased by 18.84%, 8.74%, 1.93% and 0.82%, respectively; thus the effectiveness of this method was verified.

E. ABLATION STUDY

There are two improvements to our proposed model, the introduction of building edge feature maps and the convolutional block attention module. In order to evaluate the effect of

TABLE 1. The accuracy evaluation of the extraction results of different deep learning methods.

	Precision	Recall	F1	IoU
U-Net	0.9067	0.8051	0.8527	0.7273
Deeplabv3+	0.9293	0.8791	0.9034	0.8283
U²-Net	0.9573	0.9397	0.9484	0.9018
HRNet	0.9548	0.9483	0.9515	0.9075
EA U²-Net	**0.9630**	**0.9491**	**0.9562**	**0.9157**

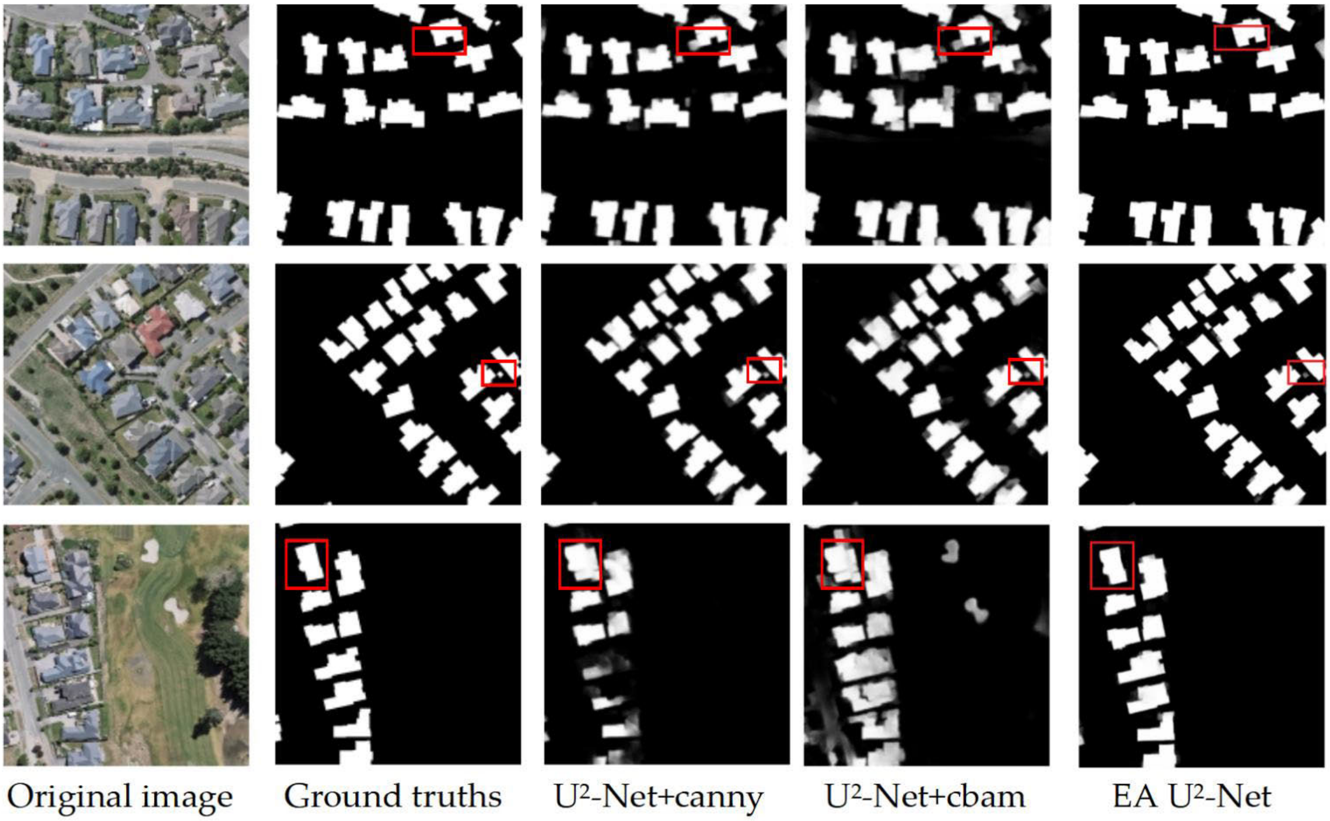

FIGURE 10. Ablation experiments on the WHU buildings dataset. Extraction results for adding different modules on buildings.

these two in modules, we conducted ablation experiments. FIGURE 10 shows the results of building extraction for the WHU building dataset by adding different modules.

As shown in Table 2, U^2-Net+canny brings considerable accuracy improvement, which indicates that the edge information of the building is critical. We also compare the experiments with U^2-Net+CBAM. As shown in the table, we show that the proposed EA U^2-Net has a significant performance gain by comparing U^2-Net+canny, U^2-Net+CBAM with EA U^2-Net. When using the method containing two modules on the WHU building dataset, the accuracy is 96.30%, F1 reaches 95.62%, and IoU is 91.57%. Thus, the validity of the method is verified.

TABLE 2. Ablation experiments on the WHU building dataset. These results show the impact of the canny edge detection and attention modules.

	Precision	Recall	F1	IoU
U² -Net	0.9573	0.9397	0.9484	0.9018
U² -Net +canny	0.9614	0.9473	0.9543	0.9125
U² -Net +CBAM	0.9593	0.9482	0.9537	0.9115
EA U² -Net	**0.9630**	**0.9491**	**0.9562**	**0.9157**

F. INFLUENCE OF TRAINING SAMPLES

Compared with ordinary photos, remote sensing imagery has the characteristics of target category diversity, feature information variability, and complexity of the interference factors. It is proposed to study the impact of the sample number, sample purity, and sample resolution on building extraction

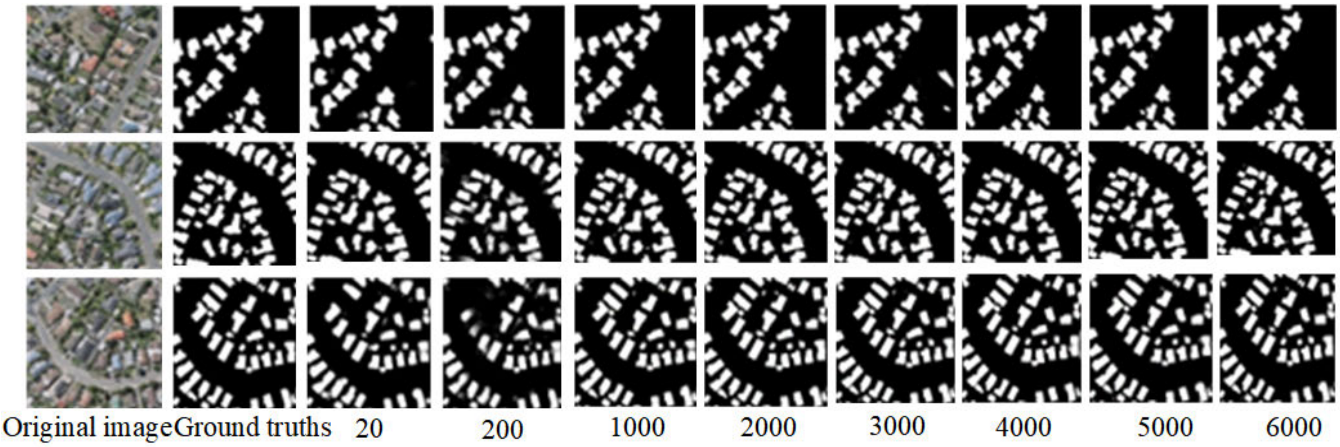

FIGURE 11. The extraction results for different numbers of building samples.

TABLE 3. The sample dataset settings.

sample number	sample purity (%)	sample resolution(m)
20	20	0.3
200	50	0.6
1000	65	1
2000	80	2
3000	90	5
4000	95	10
5000	100	
6000		

results when a sample database is constructed. The sample dataset settings are presented in Table 3.

The extraction results for different numbers of building samples are shown in FIGURE 11. When the number of building samples was small (only 20 or 200), the overall building extraction results were poor, the building boundaries were not clear, and there were a large number of false and missed detections. When the number of building samples reached 2000, the extraction effect was significantly improved, and the building boundaries were relatively clear. With a gradual increase in the number of building samples, the extraction effect also improved. The building boundaries were clearest when the number of buildings reached 4000. However, when the number of samples increased to 5000 or 6000, the extraction effect of buildings began to decline and edge blurring occurred. This shows that an appropriate increase in the number of samples is beneficial for improving the extraction results; however, blindly increasing the number of samples will lead to a decrease in the extraction effect.

Our presents the accuracy evaluation of the building extraction results for different numbers of samples. When the number of samples reached 3000, the recall and F1 reached their highest values of 95.70% and 95.14%, respectively, and the precision and IoU were reduced by only 1.23% and 0.79%, respectively, compared with the highest values.

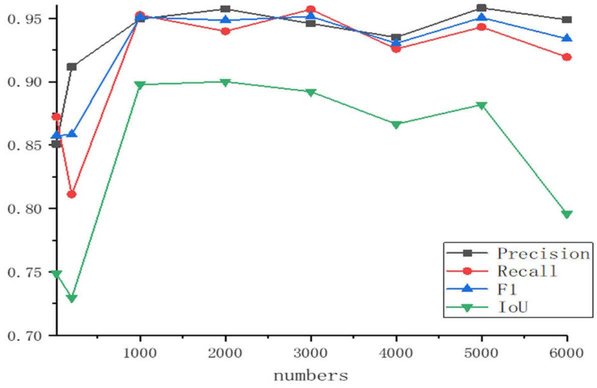

FIGURE 12. The curves of various accuracy evaluation indicators for different numbers of samples.

Compared with the case of 20 samples, the precision, recall, F1, and IoU increased by 9.49%, 8.48%, 9.41%, and 14.35%, respectively. When the number of samples was 4000, the precision, recall, F1, and IoU decreased by 1.09%, 3.13%, 2.12%, and 2.56%, respectively. When the number of samples was 5000, the precision, recall, F1, and IoU improved by 2.32%, 1.74%, 2.03%, and 1.54%, respectively, compared to 4000 samples. When the number of samples was 6000, the precision, recall, F1, and IoU decreased by 0.93 %, 2.37 %, 1.67 %, and 8.62 %, respectively, compared to the case of 5000 samples.

The curves of various accuracy evaluation indicators are shown in FIGURE 12. The precision and IoU curves exhibit the same trend. They increased at the beginning with the increase in the number of samples, and then slowly decreased until the number of samples reached 4000, and then began to rise to 5000; at this time, the precision curve reached a peak, and with the continuous increase in the number of samples, it showed a downward trend. The recall curve and F1 curve also tended to be consistent. In the early stage, as the number of samples continued to increase, both the curves first decreased and then increased. When the number of samples

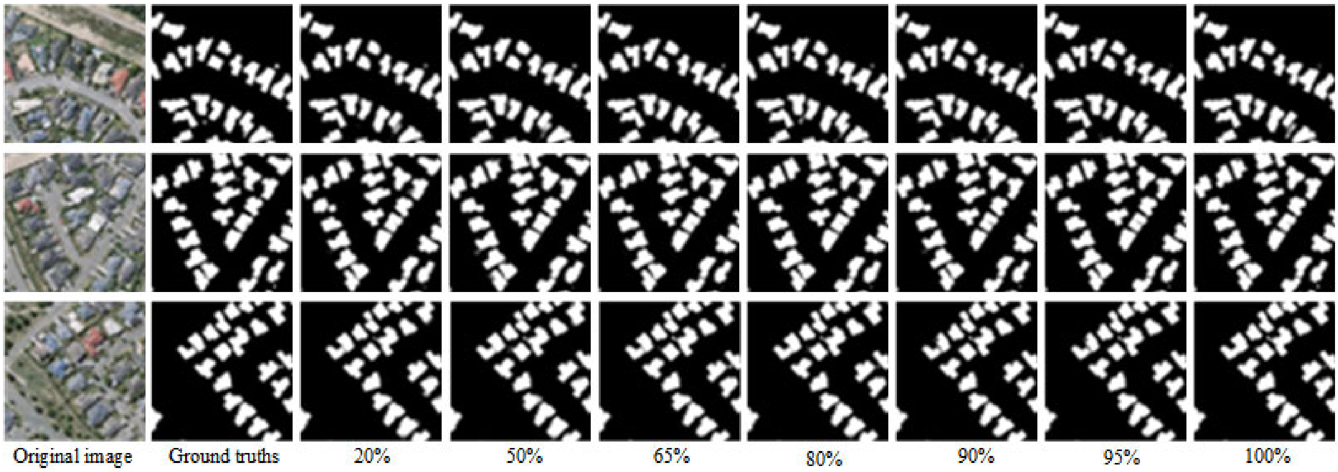

FIGURE 13. Building extraction results for different values of training sample purity.

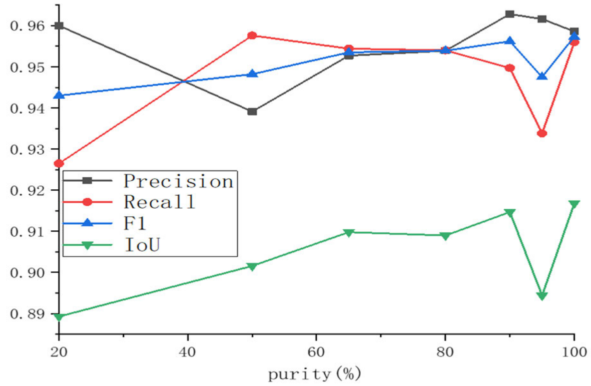

FIGURE 14. The curves of various accuracy evaluation indicators for different values of purity of samples.

reached 3000, both reached their peaks. As the number of samples continued to grow, the two curves first fell, then rose, and then continued to fall.

The buildings are densely distributed. When the number of samples reached 2000, the building extraction effect, that is, the accuracy, was significantly improved, and the time consumption was shorter. Therefore, in this study, we selected 2000 samples to study sample purity, that is, the impact of building density on the extraction results of buildings. The building extraction results are shown in FIGURE 13. When the sample purity was 20 or 50, the buildings exhibited obvious edge blurring and loss of detail. When the sample purity was 65, the blurring of the building edges was significantly improved. With the improvement of the sample purity, the building extraction effect improved, and the building boundaries became increasingly clearer. When the sample purity reached 95, the building extraction effect was the best, and the boundary was the clearest. When the sample purity reached 100, the edge definition of the building decreased, and the misdetection phenomenon became serious. It is easy to predict that non-buildings are buildings. This shows that the higher the purity of the dense samples, the better the extraction effect of buildings; however, an appropriate increase in blank samples gives the model enough information to distinguish between buildings and non-buildings, to better learn the background information, to help balance the distribution of categories, and to improve the overall extraction effect.

Our presents the accuracy evaluation of the building extraction results for different values of sample purity. When the sample purity was 20, the recall, F1, and IoU were 92.65%, 94.30%, and 88.93%, respectively, which are the lowest values, but the precision was high. When the sample purity was 50, recall, F1, and IoU increased by 3.11%, 0.52%, and 1.23%, respectively, but precision decreased by 2.09%. When the sample purity was 65, the precision, F1, and IoU were improved by 1.36%, 0.53%, and 0.82%, respectively, compared with a sample purity of 50, and the recall was slightly decreased by 0.32%. When the sample purity was 80, the precision indicators were almost identical to those when the sample purity was 65. The precision and F1 increased by 0.12% and 0.04%, respectively, and the recall and IoU decreased by 0.04% and 0.08%, respectively. When the sample purity was 90, the precision reached the highest value of 96.28%, and the recall, F1, and IoU were only 0.79%, 0.11%, and 0.21%, respectively. When the sample purity was 95, compared with a sample purity of 90, all the evaluation indicators decreased, and the precision, recall, F1, and IoU decreased by 0.12%, 1.59%, 0.87%, and 2.03%, respectively. When the sample purity was 100, F1 and IoU reached their highest values of 95.73% and 91.68%, respectively. Compared with the sample purity of 95, the recall, F1, and IoU increased by 2.22%, 0.98%, and 2.24%, respectively, but the precision dropped by 0.30%.

The curves of various accuracy evaluation indicators are shown in FIGURE 14. In the initial stage of the increase in sample purity, the various accuracy indicators gradually increased, but the precision curve gradually decreased. When the sample purity increased to approximately 50%, the precision curve decreased to its lowest value, whereas

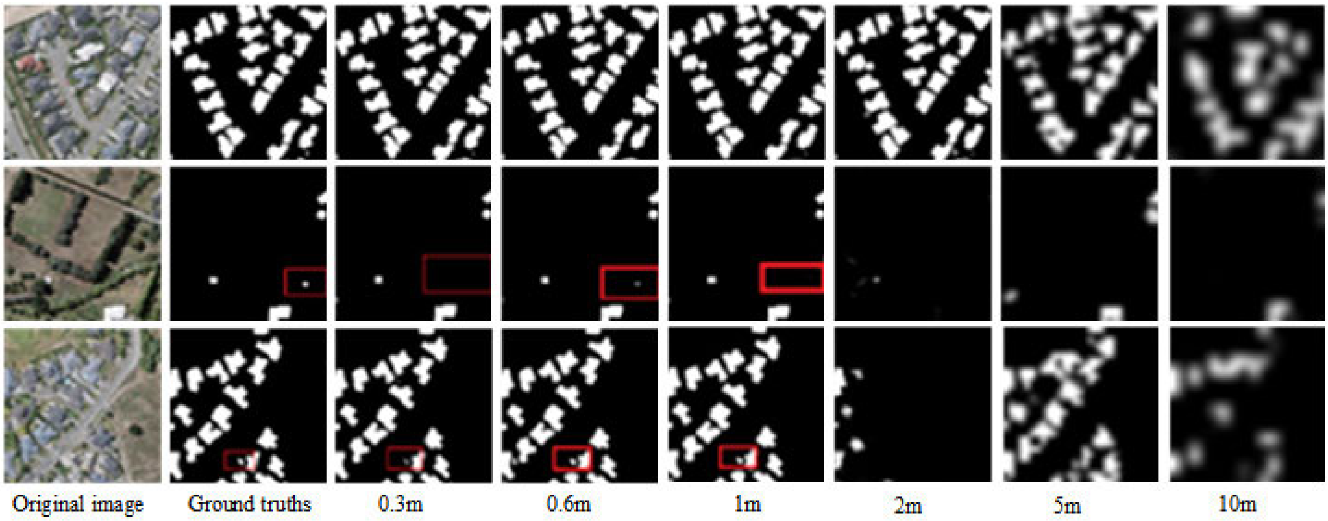

FIGURE 15. Building extraction results for different training sample resolution.

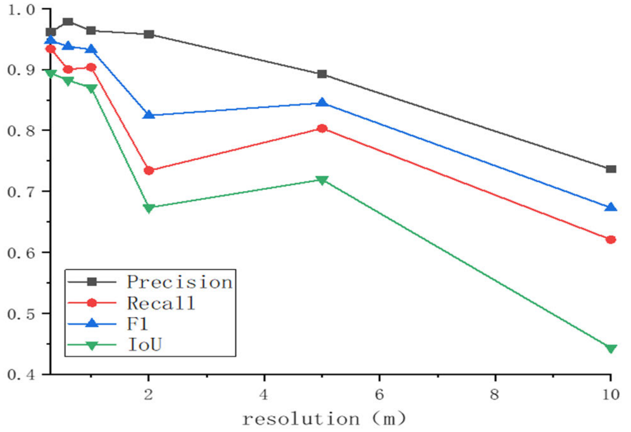

FIGURE 16. The curves of various accuracy evaluation indicators for different resolutions of samples.

the recall curve increased to the highest value. Thereafter, as the sample purity increased, the precision, F1, and IoU curves gradually increased, whereas the recall curve gradually decreased. When the sample purity was increased to 90%, each curve began to decline. When the sample purity increased to 95%, the recall, F1, and IoU curves gradually increased, whereas the precision curve continued to decline.

The diversity of remote sensing imagery sources leads to the diversity of scales of remote sensing imagery; thus, research on the sample scale is executed. The visualization results are shown in FIGURE 15. When the sample resolution was 0.3 m, the building boundary was the clearest, but it was easy to miss certain discrete small-area buildings. When the sample resolution was downsampled to 0.6 m, the building boundaries were also relatively clear, but some discrete small-area buildings that were missed when the sample resolution was 0.3 m were accurately extracted. When the sample resolution was downsampled to 1 m, the boundary of the buildings was blurred, and the extraction effect of small-area buildings also deteriorated. When the sample resolution was downsampled to 2 m, 5 m, or even 10 m, the boundary of the building was blurred, and a large number of missed and false detections occurred. This shows that with the reduction of image resolution, the extraction effect of buildings also gradually declines, but the sample resolution is not the higher the better. Appropriately reducing the sample resolution is beneficial for extracting small-area buildings.

Our presents the accuracy evaluation of the building extraction results with different sample resolutions. When the sample resolution was 0.3 m, the recall, F1, and IoU reached the highest values, which were 93.38%, 94.75%, and 89.44%, respectively, and the precision differed from the highest value of 1.70%. When the sample resolution is 0.6 m, the precision reaches the highest value of 97.86%, and the recall, F1, and IoU decreased by 3.36%, 0.97%, and 1.21%, respectively, when the sample resolution is 0.3 m. When the sample resolution was 1 m, the precision, F1, and IoU decreased by 1.50%, 0.51%, and 1.21%, respectively, and recall increased by 0.36% when the sample resolution was 0.6 m. When the sample resolution was 2 m, the precision, recall, F1, and IoU dropped by 0.57%, 16.97%, 10.81, and 19.68% when the sample resolution was 1 m, respectively. When the sample resolution was 5 m, the recall, F1, and IoU improved by 6.93%, 2.06%, and 4.59%, respectively, but the precision dropped by 6.57% when the sample resolution was 2 m. When the sample resolution was 10 m, each evaluation index had the lowest value, and the precision, recall, F1, and IoU were 73.64%, 62.08%, 67.32%, and 44.32%, respectively.

The curves of various accuracy evaluation indicators are shown in FIGURE 16. In the early stage of the decline in sample resolution, the precision curve gradually increased. When the sample resolution reached 0.6 m, the accuracy curve peaked. Subsequently, as the sample resolution decreased, the precision curve gradually decreased. The recall, F1, and IoU curves tend to be consistent. At the beginning of the decline of the sample resolution, the three curves slowly decreased, and the sample resolution was down-sampled to 0.6 m, and the three curves increased briefly and then decreased. When the sample resolution was downsampled to approximately 2 m, the three curves started to rise, and when the sample resolution was downsampled to approximately 5 m, the three curves started to fall again.

IV. CONCLUSION

This section is not mandatory but can be added to the manuscript if the discussion is unusually long or complex.

Considering the phenomenon of blurred boundaries in the process of building extraction, in this study, we propose a practical framework EA U^2-Net based on U^2-Net to extract buildings. The Canny edge detection data were added to the RGB data of the training data for training, and CBAM was added to the U^2-Net. The experimental results show that, compared with U-Net and U^2-Net, the EA U^2-Net method extracts clearer building boundaries, more complete outlines, and more accurate extraction of small buildings, which can effectively improve the building extraction accuracy. At the same time, in view of the phenomenon that deep learning relies too much on samples, this paper considers building extraction as an example to study the influence of sample parameters on the extraction results. The experimental results of the number of samples show that when the number of samples is small, increasing the number of samples can effectively improve the extraction accuracy of buildings. However, when the number of samples reaches a certain level, continuously increasing the number of samples will not only lengthen the training time of the model but also reduce the extraction effect and accuracy of buildings. The experimental results of sample purity show that the higher the purity of dense samples, the better the extraction effect of buildings, but the appropriate addition of blank samples is conducive to the network learning the characteristics of non-buildings and improves the overall extraction effect. The experimental results of sample resolution show that with a decrease in image resolution, the extraction effect of buildings also gradually declines, but the sample resolution is not the higher the better, and appropriately reducing the sample resolution is beneficial to the extraction of small-area buildings. The experimental results of the sample parameters confirm that reasonable sample parameter settings can improve the accuracy of building extraction and provide a reference for the sample settings of other target extraction experiments.

DATA AVAILABILITY STATEMENT

The WHU building dataset was provided by the Ji Shunping Team of Wuhan University (http://study.rsgis.whu.edu.cn/pages/download/).

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCES

[1] W. Xu, X. You, W. W. Zhang, and C. Deng, "Method of building scene structure extraction based on 2D map and its application in urban augmented reality," *Acta Geodaetica et Cartographica Sinica*, vol. 49, no. 12, pp. 1619–1629, 2020, doi: 10.11947/j.AGCS.2020.20190382.

[2] L. Quanhai, D. Fei, L. Lou, and R. Huiming, "A method on a rapid generation of exquisite textures of Building's roof towards planning," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 40, no. 8, pp. 1054–1060, 2015, doi: 10.13203/j.whugis20140771.

[3] G. Xiaosan, C. Xi, Z. Wenzhi, and L. Ruixiang, "Detection of damaged buildings based on generative adversarial networks," *Acta Geodaetica et Cartographica Sinica*, vol. 51, no. 2, pp. 238–247, 2022, doi: 10.11947/j.AGCS.2022.20200318.

[4] L. Dong and J. Shan, "A comprehensive review of earthquake-induced building damage detection with remote sensing techniques," *ISPRS J. Photogramm. Remote Sens.*, vol. 84, pp. 85–99, Oct. 2013, doi: 10.1016/j.isprsjprs.2013.06.011.

[5] W. Sun, G. Yang, C. Chen, M. Chang, K. Huang, X. Meng, and L. Liu, "Development status and literature analysis of China's Earth observation remote sensing satellites," *Nat. Remote Sens. Bull.*, vol. 24, no. 5, pp. 479–510, 2020, doi: 10.11834/jrs.20209464.

[6] L. Li, C. Wenting, and M. Shuli, "Segmentation of remote sensing images based on adaptive global threshold and fused markers," *Trans. Chin. Soc. Agricult. Machinery*, vol. 44, no. 7, pp. 222–228, 2013, doi: 10.6041/j.issn.1000-1298.2013.07.039.

[7] L. Wu and X. Hu, "Automatic building detection of high-resolution remote sensing images based on multi-scale and multi-feature," *Remote Sens. Land Resour.*, vol. 31, no. 1, pp. 71–78, 2019, doi: 10.6046/gtzyyg.2019.01.10.

[8] Y. Qu, "Salient building detection based on SVM," *J. Comput. Res. Develop.*, vol. 44, no. 1, p. 141, 2007.

[9] S. Cui, Q. Yan, and P. Reinartz, "Complex building description and extraction based on Hough transformation and cycle detection," *Remote Sens. Lett.*, vol. 3, no. 2, pp. 151–159, Mar. 2012.

[10] M. Izadi and P. Saeedi, "Automatic building detection in aerial images using a hierarchical feature based image segmentation," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 472–475.

[11] J. D. Wegne, U. Soergel, and B Rosenhahn, "Segment-based building detection with conditional random fields," in *Proc. Urban Remote Sensing Event*, Apr. 2011, pp. 205–208, doi: 10.1109/JURSE.2011.5764756.

[12] T. Chao, T. Yihua, C. Huajie, D. Bo, and T. Zhiwen, "A new calibration algorithm of interferometric parameters for dual-antenna airborne InSAR systems," *Acta Geodaetica et Cartographica Sinica*, vol. 39, no. 1, pp. 39–45, 2010.

[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[14] T. Xuan, W. Liang, and D. Qi, "Review of image semantic segmentation based on deep learning," *J. Softw.*, vol. 30, no. 2, pp. 440–468, 2019.

[15] Y. Meng, F. Cheng, and L. Xiong, "Semantic segmentation method of indoor obstacle images based on improved BiSeNet," *J. Huazhong Univ. Sci. Tech. (Natural Sci. Ed.)*, vol. 50, no. 6, pp. 133–138, 2022, doi: 10.13245/j.hust.220617.

[16] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany. Cham, Switzerland: Springer, 2015, pp. 234–241.

[18] K. Lu, Y. Sun, and S.-H. Ong, "Dual-resolution U-Net: Building extraction from aerial images," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 489–494, doi: 10.1109/ICPR.2018.8545190.

[19] J. Hui, M. Du, X. Ye, Q. Qin, and J. Sui, "Effective building extraction from high-resolution remote sensing images with multitask driven deep neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 786–790, May 2019, doi: 10.1109/LGRS.2018.2880986.

[20] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019, doi: 10.1109/TGRS.2018.2858817.

[21] W. Kang, Y. Xiang, F. Wang, and H. You, "EU-net: An efficient fully convolutional network for building extraction from optical remote sensing images," *Remote Sens.*, vol. 11, no. 23, p. 2813, Nov. 2019.

[22] W. Liu, J. Xu, Z. Guo, E. Li, X. Li, L. Zhang, and W. Liu, "Building footprint extraction from unmanned aerial vehicle images via PRU-Net: Application to change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2236–2248, 2021, doi: 10.1109/JSTARS.2021.3052495.

[23] C. Li, Y. Liu, H. Yin, Y. Li, P. Du, L. Zhang, and Q. Guo, "Hybrid attention cascaded U-Net for building extraction from aerial images," in *Proc. 7th Int. Conf. Big Data Comput. Commun. (BigCom)*, Aug. 2021, pp. 294–301, doi: 10.1109/BigCom53800.2021.00014.

[24] Z. Wang, Y. Zhou, S. Wang, F. Wang, and Z. Xu, "House building extraction from high-resolution remote sensing images based on IEU-Net," *Nat. Remote Sens. Bull.*, vol. 25, no. 11, pp. 2245–2254, 2021, doi: 10.11834/jrs.20210042.

[25] L. Xu, Y. Liu, P. Yang, H. Chen, H. Zhang, D. Wang, and X. Zhang, "HA U-Net: Improved model for building extraction from high resolution remote sensing imagery," *IEEE Access*, vol. 9, pp. 101972–101984, 2021, doi: 10.1109/ACCESS.2021.3097630.

[26] Z. Yuxin, Y. Qingsong, and D. Fei, "Multi-path RSU network method for high-resolution remote sensing image building extraction," *Acta Geodaetica et Cartographica Sinica*, vol. 51, no. 1, pp. 135–144, 2022, doi: 10.11947/j.AGCS.2021.20200508.

[27] H. Wang and F. Miao, "Building extraction from remote sensing images using deep residual U-Net," *Eur. J. Remote Sens.*, vol. 55, no. 1, pp. 71–85, Dec. 2022.

[28] X. Wei, X. Li, W. Liu, L. Zhang, D. Cheng, H. Ji, W. Zhang, and K. Yuan, "Building outline extraction directly using the U2-net semantic segmentation model from high-resolution aerial images and a comparison study," *Remote Sens.*, vol. 13, no. 16, p. 3187, Aug. 2021.

[29] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[30] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5872–5881.

[31] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[32] X. Soria, E. Riba, and A. Sappa, "Dense extreme inception network: Towards a robust CNN model for edge detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1912–1921.

[33] D. Zhou, G. Wang, G. He, R. Yin, T. Long, Z. Zhang, S. Chen, and B. Luo, "A large-scale mapping scheme for urban building from Gaofen-2 images using deep learning and hierarchical approach," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11530–11545, 2021.

[34] C. Qihao, S. Lin, and Z. Qian, "Scratch detection method of transparent parts based on improved U^2-Net," *Sci. Technol. Eng.*, vol. 22, no. 2, pp. 620–627, 2022.

[35] H. Pham, Q. Xie, and Z. Dai, "Meta pseudo labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 11557–11568.

[36] R. Zhang, T. Che, and Z. Ghahramani, "MetaGAN: An adversarial approach to few-shot learning," in *Proc. Neural Inf. Process. Syst.*, Red Hook, NY, USA: Curran Associates, 2018, pp. 1–11.

[37] Z. Li, S. Li, and X. Ge, "Transfer learning method for landslide extraction from GF-1 images after the wenchuan earthquake," *Nat. Remote Sens. Bull.*, vol. 27, no. 8, pp. 1866–1875, 2023, doi: 10.11834/jrs.20211020.

[38] Y. Cui, A. Dou, and S. Yang, "Deep learning sample enhancement method for 3D point cloud seismic damaged buildings," *Nat. Remote Sens. Bull.*, vol. 27, no. 8, pp. 1876–1887, 2023, doi: 10.11834/jrs.20211009.

[39] Q. Feng, B. Chen, G. Li, X. Yao, B. Gao, and L. Zhang, "A review for sample datasets of remote sensing imagery," *Nat. Remote Sens. Bull.*, vol. 26, no. 4, pp. 589–605, 2022, doi: 10.11834/jrs.20221162.

[40] A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 549–565.

[41] X. Qin, Z. Zhang, and C. Huang, "U-Net: Going deeper with nested U-structure for salient object detection," 2020, *arXiv:2005.09007*.

[42] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1741–1750.

[43] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[44] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[45] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[46] K. Xu, J. Ba, and R. Kiros, "Show, attend and tell: Neural image caption generation with visual attention," 2015, *arXiv:1502.03044*.

[47] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[48] J. Shunping and W. Shiqing, "Building extraction via convolutional neural networks from an open remote sensing building dataset," *Acta Geodaetica et Cartographica Sinica*, vol. 48, no. 4, pp. 448–459, 2019, doi: 10.11947/j.AGCS.2019.20180206.

[49] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[50] D. Zheng, S. Li, F. Fang, J. Zhang, Y. Feng, B. Wan, and Y. Liu, "Utilizing bounding box annotations for weakly supervised building extraction from remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4702517, doi: 10.1109/TGRS.2023.3271986.

[51] C. Long, S. Wenlong, S. Tao, L. Yizhu, J. Wei, L. Jun, L. Hongjie, F. Tianshi, G. Rongjie, H. Abbas, M. Lingwei, L. Shengjie, and H. Qian, "Field patch extraction based on high-resolution imaging and U2-Net++ convolutional neural networks," *Remote Sens.*, vol. 15, no. 20, p. 4900, Oct. 2023, doi: 10.3390/rs15204900.

[52] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[53] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[54] J. Chen, "Efficient seismic data denoising via deep learning with improved MCA-SCUNet," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5903614, doi: 10.1109/TGRS.2024.3355972.

[55] G. Chen, J. Chen, K. Jensen, C. Li, S. Chen, H. Wang, J. Li, Y. Qi, and X. Huang, "Joint data and model-driven simultaneous inversion of velocity and density," *Geophys. J. Int.*, vol. 237, no. 3, pp. 1674–1698, Apr. 2024, doi: 10.1093/gji/ggae128.

FEIFEI XIE received the Ph.D. degree from Wuhan University, in 2014. She is currently an Associate Professor with the College of Geodesy Geomatics, Shandong University of Science and Technology. With over nine years of teaching experience, her current research interests include drone photogrammetry and close-range photogrammetry.

MINGZHE YI received the B.S. degree from Shanxi Datong University, Shanxi, China, in 2016, and the M.S. degree from Shandong University of Science and Technology, Qingdao, China, in 2024. His research interests include category-level 6D bit-position estimation.

ZHILING HUO received the M.S. degree from Shandong University of Science and Technology, Qingdao, China, in 2022. She is currently with Beijing City Interface Technology Company Ltd.

LIN SUN received the Ph.D. degree from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, in 2006. From 2012 to 2014, he was a Research Scholar with the Department of Geographical Sciences, University of Maryland. Since 2015, he has been a Professor with Shandong University of Science and Technology. With over 17 years of teaching experience, his current research interests include quantitative analysis of remote-sensing images and machine vision.

JINGYU ZHAO is currently pursuing the bachelor's degree in remote sensing science and technology from Shandong University of Science and Technology. Her research interest includes the study of intelligent extraction methods for buildings from high-resolution remote sensing images.

ZHIPENG ZHANG received the B.S. degree from Shandong University of Science and Technology, Tai'an, China, in 2017, and the M.S. degree in photogrammetry from Shandong University of Science and Technology, Qingdao, China, in 2023. His research interests include point cloud segmentation and registration.

JINPENG CHEN received the B.S. degree from Jining University, Jining, China, in 2017, and the M.S. degree in photogrammetry from Shandong University of Science and Technology, Qingdao, China, in 2023. His research interest includes 6D object pose estimation using deep learning framework.

JINRUI ZHANG received the B.S. and M.S. degrees from Shandong University of Science and Technology, Qingdao, China, in 2018 and 2024, respectively. His research interests include bin-picking and hand-eye calibration.

FANGRUI CHEN received the bachelor's degree from Shandong University of Science and Technology, in 2018, where he is currently pursuing the master's degree. His research interest includes point cloud segmentation.