An Improved Mask R-CNN: Extraction of Door and Window Instances on Village Building Façade Images

Daiqi Zhong[1,2], Lin He[1,2], Yi Lin[1,2,*]

[1] College of Surveying and Geo-Informatics, Tongji University, 200092 Shanghai, China – (2231979, 2010964, linyi)@tongji.edu.cn
[2] Research Center of Remote Sensing Technology and Application, Tongji University, 200092 Shanghai, China

KEY WORDS: Object detection, Instance segmentation, Window and door extraction, Mask R-CNN, Attention mechanism.

ABSTRACT:

Rapid access to the basic structure of village buildings is conducive to the investigation of the load-bearing bodies of village houses and provides data support for disaster assessment and post-disaster rescue and reconstruction. The development of computer vision technology provides new ideas and tools for identifying and extracting basic structures of housing buildings. Considering that the original Mask R-CNN ignores the spatial association and relationship of door and window elements, an advanced deep learning model based on Mask R-CNN network is proposed in this paper to detect and segment the door and window structure from the façade images. The improved network architectures integrate the attention mechanism with the original network, containing an improved Coordinate Attention(CA) module and a relationship module-based head network. The experimental results show that the Average Precision(AP) value of the backbone combined with the improved CA module is increased by 0.7% and 0.7% on regression and segmentation tasks respectively, compared with the original Mask R-CNN network. In the head network based on the relationship module, the calculation strategy of the relational module proposed in this paper increases the AP values of detection and segmentation from 76.7% and 77.7% to 80.6% and 80.0%, respectively.

1. INTRODUCTION

Among natural disasters, earthquake damage is most closely related to housing construction. The areas with severe earthquake damages are mainly in village and town areas, therefore, the survey of village and town housing disaster-bearing body information can provide technical support and decision basis for post-earthquake emergency rescue, decision-making command and post-disaster reconstruction in the region (Xu et al., 2014).

In the building structure survey of the disaster-bearing body investigation, the size and location of openings in walls are the key factors in understanding the building structure and carrying out damage assessment of disaster-bearing bodies, in which case, window and door elements are one of the most distinctive and numerous elements of the building façade. Compared with traditional extraction methods such as contour-based methods (Haugeard et al., 2009; Lee & Nevatia, 2004; Recky & Leberl, 2010), intensity-based methods (Čech & Šára, 2009), and machine learning-based methods (Jampani et al., 2015; Reznik & Mayer, 2008; Yang et al., 2012), extracting doors and windows with UAV data and deep learning based methods would greatly reduce the difficulty of survey work and improve efficiency. First, rather than conducting large amounts of on-the-spot investigation, unmanned aerial vehicles (UAV), aircraft, satellites, and other equipment are used to scan a specific area and obtain the corresponding data. Second, the quality and accuracy of data are greatly improved by using aerial surveys and remote sensing technology. Third, deep learning has achieved state-of-the-art performance in a number of vision tasks (Dosovitskiy et al., 2020; Girshick, 2015; Long et al., 2015; Redmon et al., 2016; Ren et al., 2015). However, extracting components from building façades with UAV data and deep learning methods is barely studied in current research (Liu et al., 2020).

In this paper, a deep learning model based on Mask R-CNN (He et al., 2017) is proposed to detect and segment door and window elements from the building façade. For the detection of door and window structures of village houses, several improved architectures fused with Mask R-CNN to improve the accuracy of the extraction of doors and windows were designed. The distribution of doors and windows of houses in villages and towns is generally not as well arranged as the elements of houses in cities, especially the houses in residential areas are more significant. Therefore, it is necessary to improve the network's attention to the window and door elements in feature extraction. Also, considering that the original Mask R-CNN only extracts the appearance features of individual targets without considering the spatial relationship of door and window elements and the association between objects, the improved network architectures in this paper integrate the attention mechanism with the original network model. The specific work is as follows:
(1) An improved CA(Coordinate Attention) module is proposed to fuse coordinate attention and channel attention to optimize the features extracted by the backbone, enabling the network to better extract RoI for object detection;
(2) Relation module is embedded in the fully connected layer of the head network to integrate the appearance features and geometric features among different objects, and a novel strategy is proposed to seek a weighted sum of the attention of geometric features and appearance features to improve the degree of relationship of geometric features among door and window objects.

2. METHODOLOGY

2.1 Network Architecture

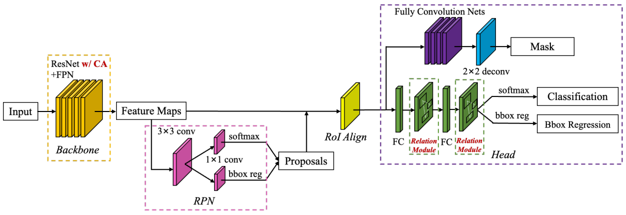


Figure 1. General framework of the proposed network.

The improved Mask R-CNN network structure is shown in Fig.1. The network model consists of three parts(from left to right): (1)

https://doi.org/10.5194/isprs-annals-X-1-W1-2023-295-2023 | © Author(s) 2023. CC BY 4.0 License.

Backbone with Improved CA Module, (2) Region Proposal Network, (3) Head with Relation Module.

Firstly, the ResNet (He et al., 2016) + FPN (Lin et al., 2017) feature extraction structure is adopted by the backbone network to extract multi-scale feature maps. Considering the balance of detection effect and speed, ResNet50 is selected. At the same time, an improved CA module is introduced, and the feature maps output by ResNet is extracted with the attention feature by the improved CA module, and then the pyramid feature structure is generated by the FPN network.

Second, the RPN network is used to predict the classification score of the anchor box generated by each pixel and the bounding box. Then, select the anchor boxes with higher scores as the proposal regions into the head network.

In head network, one branch is for classification and bounding box regression, and the other is for instance segmentation. The relation module is embedded after the two fully connected layers of the first branch and is used to learn the relationship between objects.

2.2 Backbone fused with Improved CA Module

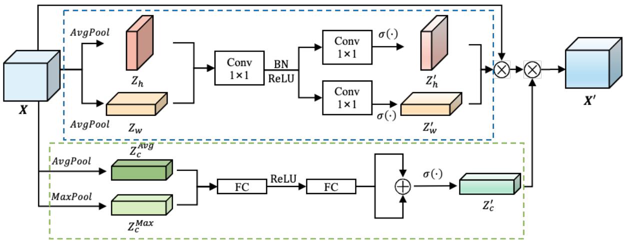


Figure 2. Improved CA Module.

Coordinate Attention Module. The CA module (Hou et al., 2021) is a way to combine spatial coordinate information to enhance the position information of the target that the model focuses on when learning image features. It adjusts the importance of each spatial location by introducing a weighting factor calculated from the coordinate information, thus making the model more focused on important spatial location information. The CA module can well handle the position relationship of objects, thereby improving the accuracy of the network in the detection task. Its structure is shown in the blue dotted box in the Fig.2.

Specifically, given an input $X \in \mathbb{R}^{C\times H\times W}$, each channel is first encoded along the horizontal and vertical coordinate directions using pooling kernels of dimensions $(H,1)$ and $(1,W)$, respectively, and the output Z_h^c of the cth channel of height h can be expressed as follows:

$$Z_h^c(h) = \frac{1}{W}\sum_{i=0}^{W-1} X^c(h,i) \tag{1}$$

Similarly, the output Z_w^c of the cth channel of width w can be expressed as:

$$Z_w^c(w) = \frac{1}{H}\sum_{i=0}^{H-1} X^c(j,w) \tag{2}$$

The above 2 transformations aggregate features along each of the two spatial directions to obtain a pair of direction-aware feature maps, allowing the attention module to capture the long-term dependencies along one spatial direction and preserve the precise location information along the other spatial direction, which helps the network to locate the region of interest more accurately. In other words, the coordinate information embedding operation is to average pool the coordinates in the X and Y directions where each pixel is located, and obtain X AvgPool and Y AvgPool.

By embedding the position information, concatenate the two position features and then perform the linear transformation operation as follows:

$$f = \delta(F_1([Z_h, Z_w])) \tag{3}$$

where Z_h and Z_w are the coordinate information embedding, $F_1(\cdot)$ is the 1×1 convolution, $\delta(\cdot)$ is the activation function (typically ReLU), $f \in \mathbb{R}^{C/r\times(H+W)}$ is the generated intermediate feature map for spatial information in the horizontal and vertical directions, and r denotes the downsampling ratio to control the size of the module.

Next, $f \in \mathbb{R}^{C/r\times(H+W)}$ is split into two parts $f_h \in \mathbb{R}^{C/r\times H}$ and $ff_w \in \mathbb{R}^{C/r\times W}$, and then two linear transformations are used to make the two features have the same number of channels, as follows:

$$Z'_h = \sigma(F_h(f_h)) \tag{4}$$

$$Z'_w = \sigma(F_w(f_w)) \tag{5}$$

where $F_h(\cdot)$ and $F_w(\cdot)$ are two 1×1 convolutions and $\sigma(\cdot)$ is the Sigmoid function.

Finally, the output $X' \in \mathbb{R}^{C\times H\times W}$ with coordinate attention features is obtained by multiplying the original features with the coordinate features using the broadcast mechanism:

$$X' = X \times Z'_h \times Z'_w \tag{6}$$

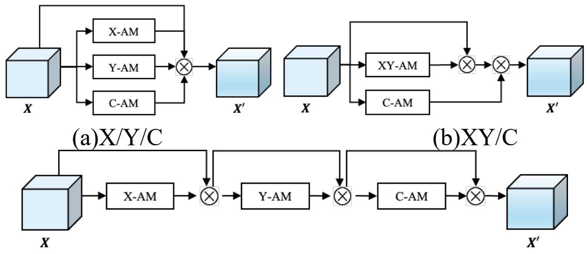
Improved CA Module. The CA Module only considers the importance of spatial location information, while it is not reflected on which channel is more worthy of attention. Therefore, An improved CA module was designed. On the basis of CA Module, the channel attention branch is added to enable the network to better detect objects by combining both their position information and channel information, and its structure is shown in the green dashed box in Fig.2.

When considering the channel attention, we are inspired by the SE (J. Hu et al., 2018) and CBAM (Woo et al., 2018) modules to perform global average pooling and maximum pooling on the channel dimension of the original feature image to generate two different global features $Z_C^{Avg} \in \mathbb{R}^{1\times 1\times C}$ and $Z_C^{Max} \in \mathbb{R}^{1\times 1\times C}$. Then, Z_C^{Avg} and Z_C^{Max} are simultaneously fed into a shared multilayer perceptron containing two fully connected layers FC_1 and FC_2. Finally, the channel attention feature $Z'_C \in \mathbb{R}^{1\times 1\times C}$ is obtained by element-wise sum as follows:

$$Z'_C = \sigma\left(FC_2\left(ReLU\left(FC_1(Z_C^{Avg})\right)\right) + FC_2\left(ReLU\left(FC_1(Z_C^{Max})\right)\right)\right) \tag{7}$$

where $\sigma(\cdot)$ refers to the sigmoid function, $ReLU(\cdot)$ refers to the ReLU function, and FC_1 and FC_2 are two fully connected layers with shared weights.

Three different fusion methods. Considering that the CA module is essentially composed of X-directional attentional feature module (X-AM) and Y-directional attentional feature module (Y-AM), plus channel attentional feature module (C-AM), different fusion methods may produce different effects, so three fusion methods are tried in this paper as shown in Fig.3.



(a)X/Y/C (b)XY/C

(c)X+Y+C

Figure 3. Three different fusion methods for improved CA module.

The first fusion method is to input the original feature map into the three branches of X-AM, Y-AM, and C-AM in parallel and obtain the output attention feature map by multiplying it with the original features in the broadcast mechanism as shown in Fig.3(a).

The second fusion method is to input the original feature map in parallel to the XY-AM and C-AM branches, where the XY-AM branch is the attention feature that extracts the features through 1×1 convolution by concatenating the coordinate encoding in the X and Y directions and then splits after linear transformation. Then multiply the XY-AM branch's feature with the original feature map and then multiply it with the output of the C-AM branch to obtain the attention feature image as shown in Fig.3(b). The third fusion method is to connect the three branches in series and multiply the input and output of each branch with skip connections to obtain the attentional feature image as shown in Fig.3(c).

Backbone with improved CA module. The ResNet was used in Backbone networkt, and then the outputs of ResNet, C1 to C5, are put into the improved CA module. The final outputs, P2 to P6, at different scales are extracted by the CA module in combination with the FPN network as shown in Fig.4.

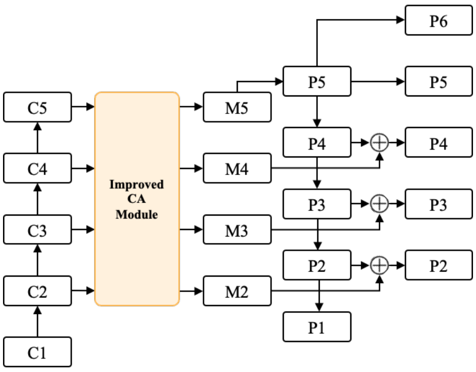
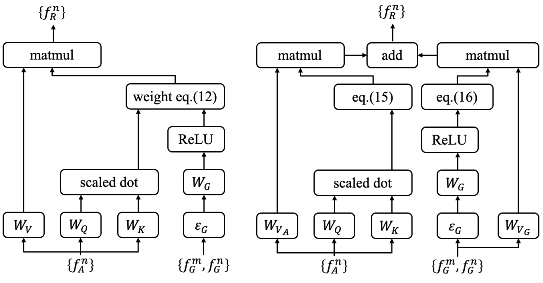


Figure 4. Backbone with improved CA module.

2.3 Head with Relation Module

Relation Network. The relation network (H. Hu et al., 2018) models the relationship between objects. For an image with N RoIs, there are appearance feature f_A and geometric feature f_G. Appearance feature f_A refers to the feature vector corresponding to each RoI, and geometric feature f_G denotes the bounding box coordinates and scale of each RoI object, whose feature set is denoted as $\{f_A^n, f_G^n\}_{n=1}^N$.



(a) (b)

Figure 5. Two strategies for calculating relation features.

Given the nth RoI object, the relationship between it and all other objects is characterized by $f_R(n)$ as follows:

$$f_R(n) = \sum_m \omega^{mn} \cdot (W_V \cdot f_A^m) \tag{8}$$

where f_A^m denotes the appearance feature of the mth object, W_V is the linear transformation matrix, and ω^{mn} denotes the relational feature weights between the mth object and the nth object. The relation feature f_R is added with the original image features f_A and used as the input of the next layer.

The calculation of the relational feature weight ω^{mn} consists of three parts:
(1) Calculating the appearance correlation ω_A^{mn} between two objects.

$$\omega_A^{mn} = \frac{dot\left(W_K f_A^m, W_Q f_A^n\right)}{\sqrt{d_k}} \tag{9}$$

where W_K and W_Q are linear transformation matrices and d_k is the number of feature dimensions.

(2) Calculating the geometric correlation ω_G^{mn} between two objects.

$$\omega_G^{mn} = max\left\{0, W_G \cdot \varepsilon_G\left({f'}_G^{mn}\right)\right\} \tag{10}$$

where ${f'}_G^{mn}$ denotes the geometric feature after coordinate transformation, W_G is the linear transformation matrix of the geometric feature, and ε_G is an operator to embed the 4-dimensional coordinate information into a 64-dimensional feature with sinusoidal position encoding.

The coordinate transformation is to ensure translation non-deformation and scale invariance between different target features:

$$f_G'^{mn} = \left(\log\left(\frac{|x_m - x_n|}{w_m}\right), \log\left(\frac{|x_m - x_n|}{w_m}\right), \log\left(\frac{w_n}{w_m}\right), \log\left(\frac{h_n}{h_m}\right)\right)^T \tag{11}$$

(3) Calculate the correlation ω^{mn} between the two targets.

$$\omega^{mn} = \frac{\omega_G^{mn} \cdot \exp(\omega_A^{mn})}{\sum_k \omega_G^{kn} \cdot \exp\left(\omega_A^{kn}\right)} \tag{12}$$

In calculating the relation module, the appearance features are divided into N_r equal parts so that they pass through different relation modules, and then sum up with the appearance features f_A as the input of the next layer:

$$f_A^n = f_A^n + \left[f_R^1(n); f_R^2(n); \ldots; f_R^{N_r}(n)\right] \tag{13}$$

The relational features W_Q, W_K and W_V correspond to the Query and the Key-Value pair in the attention mechanism, respectively. For each pair of relational features, W_Q, W_K, W_V and W_G are learnable weights, so the relation module can be integrated with the network for end-to-end network training as shown in Fig.5(a).

The geometric similarity between objects in the above relation module is involved in the calculation as the weight coefficients of object similarity, and we propose an alternative strategy that uses both appearance features and geometric features for the calculation of relation features to enable the network to learn the association between the location information of different objects. In this strategy, the relation features are computed as follows:

$$f_R(n) = \sum_m \omega_1^{mn} \cdot \left(W_{V_A} \cdot f_A^m\right) + \omega_2^{mn} \cdot \left(W_{V_G} \cdot f_G^m\right) \tag{14}$$

where ω_1^{mn} and ω_2^{mn} are the appearance similarity and geometric similarity weights between the mth and nth RoI, obtained from ω_A^{mn} and ω_G^{mn} by softmax transformation:

$$\omega_1^{mn} = \frac{\exp(\omega_A^{mn})}{\sum_k \exp\left(\omega_A^{kn}\right)} \tag{15}$$

$$\omega_2^{mn} = \frac{\exp(\omega_G^{mn})}{\sum_k \exp\left(\omega_G^{kn}\right)} \tag{16}$$

In this way, the obtained relational features integrate the weighted sum of the appearance and geometric features between different targets after a linear transformation as shown in Fig.5(b). As a result, the network is able to fully consider the geometric relationships between different objects.

Head with Relation Module. The computational flow of the head network in the original Mask R-CNN is shown in Fig.6(a). Given the nth RoI object feature, it is input to two fully connected layers of size 1024 dimensions, and then the corresponding classification scores and bounding box regression values are obtained by linear transformations. Each object is classified and regressed using only the features extracted by the preorder network. The computational flow of the head network with the relationship module is shown in Fig.6(b), where all RoI object features are used as inputs and the dimensions of the inputs and outputs are kept constant, and each target achieves correlation learning between object features by learning associations with others, including appearance features and geometric features.

$$\begin{aligned} &RoI\_feat_n \xrightarrow{FC} 1024 \\ &\xrightarrow{FC} 1024 \\ &\xrightarrow{Linear} (score_n, bbox_n) \end{aligned}$$

(a) Original head

$$\begin{aligned} &\{RoI\_feat_n\}_{n=1}^{N} \xrightarrow{FC} 1024 \cdot N \xrightarrow{RM_1} 1024 \cdot N \\ &\xrightarrow{FC} 1024 \cdot N \xrightarrow{RM_2} 1024 \cdot N \\ &\xrightarrow{Linear} \{(score_n, bbox_n)\}_{n=1}^{N} \end{aligned}$$

(b) head with relation module

Figure 6. Figure placement and numbering.

3. EXPERIMENTS

In this section, AP(Average Precision) is used as the evaluation criteria for each category. The AP value is the area below the PR curve for a certain class of targets. The recognition accuracy of a model can be measured by analyzing the AP values of the network training results. The PR curve is plotted with Recall as the horizontal axis and Precision as the vertical axis for different values. Usually, the larger the area contained below the PR curve, the better the model is. In the COCO(Microsoft Common Objects in COntext) evaluation criteria (Lin et al., 2014), AP@[0.5:0.95] is the average of the 10 AP values obtained by dividing the $IoU_{threshold}$ from 0.5 to 0.95 in steps of 0.05, which allows for a more refined evaluation of the detection accuracy of the model. AP^{50} and AP^{75} represent the AP values when $IoU_{threshold}$ is 0.50 and 0.75 respectively, while AP^S, AP^M and AP^L are the AP values for different object sizes with S, M and L representing object sizes of area < 32^2, 32^2 < area < 96^2 and area > 96^2 respectively.

3.1 Datasets

The datasets for this experiment were obtained from 3D models of building facades reconstructed from UAV aerial imagery in selected villages and towns, with four wall facades for each building as our goal is to make quick and large-batch window and door extractions for village houses.

Most of the buildings photographed are two- or three-story bungalows in height. The image is size of 2000 × 3000 pixels, which consists of Red, Green and Blue channels. To prepare the input data for training, window and door objects are extracted from the annotated images and encoded in the COCO instance segmentation format.

To maintain the consistency of the results, every tested model ran 300 epochs on a single GPU. Also, each model was trained and finetuned on the basis of the pre-trained ResNet model on ImageNet-1k dataset.

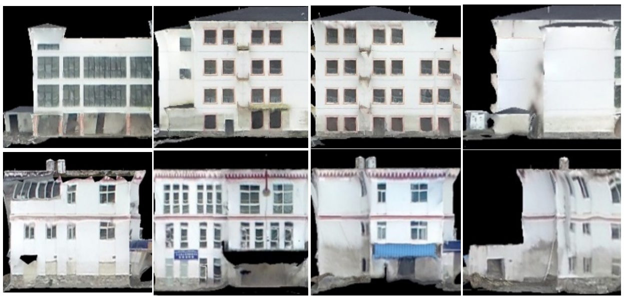


Figure 7. Dataset images.

3.2 Three Fusion Methods in Improved CA Module

Three different improved CA module fusion methods are as follows: (1)X/Y/C: The original feature map X is input in parallel to the X-AM, Y-AM, and C-AM branches, and the output attentional feature image is obtained by multiplying it with the original feature by the broadcast mechanism. (2)XY/C: The original feature map X is input in parallel to the XY-AM and C-AM branches, and then multiplied with the original image to obtain the attentional feature image. (3)X+Y+C: The three branches are connected in series and the input and output of each branch are multiplied by skip connections to obtain the attentional feature.

	Method	mAP		Method	mAP
AP$^{\text{bbox}}$	Baseline	76.0	AP$^{\text{segm}}$	Baseline	77.0
	X/Y/C	76.2		X/Y/C	77.2
	XY/C	**76.7**		XY/C	**77.7**
	X+Y+C	75.5		X+Y+C	76.5

Table 1. The mAP (%) of three different fusion methods of improved CA module.

Table.1 shows that the mAP values of both X/Y/C and XY/C fusion methods improved compared to the baseline. The mAP value of the XY/C fusion method is the highest in the regression and segmentation tasks, with 0.7% and 0.7% improvement over the baseline model, respectively. The results show that the addition of the CA module attentional feature extraction in the backbone enables the network to better extract and deliver the door and window objects as RoIs to the downstream tasks in the subsequent RPN network.

Method	**AP^{50}**	**AP^{75}**	**AP^{S}**	**AP^{M}**	**AP^{L}**
Baseline	91.4	**88.6**	77.9	72.7	79.0
X/Y/C	91.6	88.0	78.0	73.0	79.2
XY/C	**91.9**	88.4	**78.4**	**73.5**	79.2
X+Y+C	91.4	87.3	77.1	72.7	**79.6**

Table 2. The AP values(%) of three different fusion methods of improved CA module on regression task.

Method	**AP^{50}**	**AP^{75}**	**AP^{S}**	**AP^{M}**	**AP^{L}**
Baseline	77.0	90.8	88.2	65.8	71.0
X/Y/C	77.2	91	88.6	65.9	71.4
XY/C	**77.7**	**91.4**	**88.8**	**66.1**	**71.6**
X+Y+C	76.5	91	86.8	65.0	71.0

Table 3. The AP values(%) of three different fusion methods of improved CA module on segmentation task.

Although the X/Y/C fusion method is slightly better than the original Mask R-CNN, it performs not well as XY/C. This is probably because the XY/C attention extraction method integrates the X- and Y-direction features when extracting coordinate attention, which is richer in information association and long-term feature dependency than the X- and Y-direction input branches separately. As shown in Table.2 and Table.3, the X+Y+C method performs worse than the original Mask R-CNN, presumably because some of the detailed features are lost after the original features are extracted by modular cascading.

3.3 Two Strategies in Relation Module

In a relational module-based head network, two crucial hyper-parameters exist in the model: the number of relations N_r and the number of modules $\{r_1, r_2\}$. We set $r_1 = 1$ and $r_2 = 1$ on the basis on Hu's experiment(Hu et al., 2018), and compare the results for different numbers of N_r.

The AP is highest when the number of N_r is 16 according to Hu's research. However, we came to a different conclusion in our trials. Experiments on the header network were carried out on the improved CA module based on XY/C in the previous section, and the experimental results are shown in the table. According to Table.4, the mAP values for the regression and segmentation tasks showed an increasing trend as they increased overall. Specifically, when N_r is less than 8, both AP$^{\text{bbox}}$ and AP$^{\text{segm}}$ tend to increase by increasing the number of N_r. When N_r is equal to 16, the mAP values decreased slightly. When N_r is equal to 32, the mAP value is the highest, and the accuracy increases by 2.6% and 3.9% in the regression task and 2.3% and 2.2% in the segmentation task, respectively. Considering the computational efficiency and computational complexity, $N_r = 32$ is chosen.

The two different relation module computation strategies perform well on our dataset both. Considering that the output of the segmentation branch of Mask R-CNN will resize to the size of RoI, the accuracy of the regression branch will affect the accuracy of the post-segmentation processing. For the above reasons, the relational module computation strategy of Head+RM$^{\text{Eq.(15)+Eq.(18)}}$ is proposed.

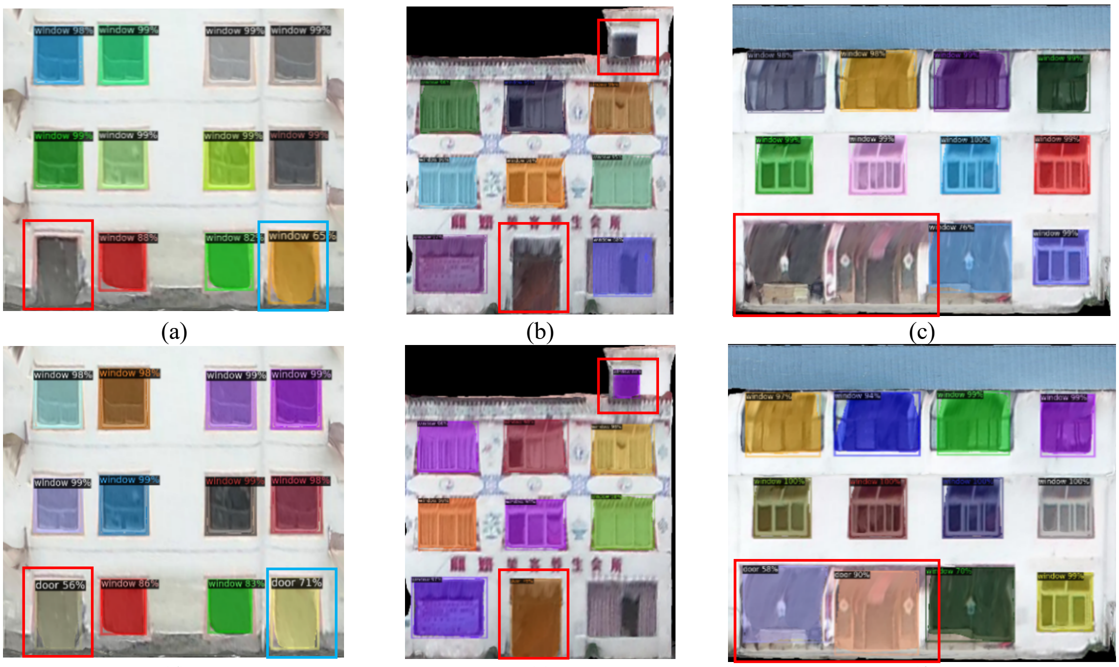
3.4 Precision Analysis

The comparison between the detection results of the original Mask R-CNN and the improved Mask R-CNN reveals that the original Mask R-CNN performs less effectively for door elements, and is prone to false and missed detections, as shown in Fig.8. After the introduction of the attention mechanism, the number of false and missed detections is significantly reduced, which is presumably due to the fact that the introduction of the attention mechanism enables the network to learn the connection and difference between the appearance and location of the target object and other objects, thus enabling the network to identify the objects more accurately.

Specifically, some of the severely missed objects were mainly door elements. On one hand, door elements were less distributed and regularly arranged in a building façade than windows. On the other hand, the appearance, style and scale of door elements varied, making them easy to be missed in the original network. In the improved network, combined with the coordinate information, the network is able to detect door elements based on the geometric relationships and the relative positions of the window elements in the façade.

	Method	Baseline	N_r					
			1	2	4	8	16	32
AP$^{\text{bbox}}$	Head+RM$^{\text{Eq. (12)}}$	76.7	77.9	78.1	78.2	78.6	78.4	**79.3**
	Head+RM$^{\text{Eq. (15)+Eq. (18)}}$		77.1	78.0	78.7	79.4	79.6	**80.6**
AP$^{\text{segm}}$	Head+RM$^{\text{Eq. (12)}}$	77.7	78.6	78.9	79.0	79.6	79.2	**80.0**
	Head+RM$^{\text{Eq. (15)+Eq. (18)}}$		78.3	78.5	79.1	79.8	79.5	**79.9**

Table 4. The mAP (%) for Evaluating the Effect of N_r on two strategies.



(a) (b) (c)

(d) (e) (f)

Figure 8. Detection and segmentation results.

(a)(b)(c) are the results of origin Mask R-CNN and (d)(e)(f) are the results of improved Mask R-CNN. Red boxes are for under-detection and blue boxes are for misdetection.

4. CONCLUSION

In this paper, a Mask R-CNN network was proposed for extracting instances of window objects improved from UAV building façade images. Two improvements have been made. First, with the combination of the attention features, the improved CA module enables the network to learn the attention features based on coordinate positions and channel attention features through coordinate encoding. Therefore, the RoIs containing the door and window objects can be extracted more accurately, which helps further tasks in the head network. Second, by combining the relation module with the head network, a set of objects can be processed simultaneously through the interaction of appearance features and geometric features, so that relationships between doors and windows can be learned during the learning process, enhancing the network's ability to represent image features and geometric relationships between objects. Through these attention mechanisms, the network can learn the appearance, space, and object relationships between objects on a deeper level, resulting in greater accuracy.

Quantitatively, the average precisions of the improved CA module are higher than the baseline by 0.7% and 0.7% on both regression and segmentation tasks. The average precisions of the head with relation module increased from 76.7% and 77.7% to 80.6% and 79.9% compared to those without relation module. Experiments show that the proposed method can fully utilize the spatial relationship features between doors and windows, and therefore can achieve better detection results.

The attention mechanism can be utilized to enhance the network's attention to the feature extraction, which reduces the occurrence of misdetections and under-detection in extracting the RoIs of the door and window elements. Besides, combined with the relation module, the network can model the appearance and geometric relationships between each object, learning the similarities between the same class of objects and the differences between different classes.

However, due to the complexity and irregularity of window and door elements in village buildings, the improvement of the network in attention modeling for window and door elements is still limited. Future work will continue to consider modeling the relative relationship between window and door elements and house facades to further investigate instance segmentation of facade elements.

ACKNOWLEDGEMENTS

This study was funded by the Key Project of Shanghai Science and Technology Innovation Action (No. 20dz1201202).

REFERENCES

Čech, J., & Šára, R. (2009). Languages for constrained binary segmentation based on maximum a posteriori probability labeling. *International Journal of Imaging Systems and Technology*, *19*(2), 69–79.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv Preprint arXiv:2010.11929*.

Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.

Haugeard, J.-E., Philipp-Foliguet, S., Precioso, F., & Lebrun, J. (2009). Extraction of windows in facade using kernel on graph of contours. *Image Analysis: 16th Scandinavian Conference, SCIA 2009, Oslo, Norway, June 15-18, 2009. Proceedings 16*, 646–656.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13713–13722.

Hu, H., Gu, J., Zhang, Z., Dai, J., & Wei, Y. (2018). Relation networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3588–3597.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Jampani, V., Gadde, R., & Gehler, P. V. (2015). Efficient facade segmentation using auto-context. *2015 IEEE Winter Conference on Applications of Computer Vision*, 1038–1045.

Lee, S. C., & Nevatia, R. (2004). Extraction and integration of window in a 3D building model from ground view images. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, *2*, II–II.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755.

Liu, H., Xu, Y., Zhang, J., Zhu, J., Li, Y., & Hoi, S. C. (2020). DeepFacade: A deep learning approach to facade parsing with symmetric loss. *IEEE Transactions on Multimedia*, *22*(12), 3153–3165.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Recky, M., & Leberl, F. (2010). Windows detection using k-means in cie-lab color space. *2010 20th International Conference on Pattern Recognition*, 356–359.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*.

Reznik, S., & Mayer, H. (2008). Implicit shape models, self-diagnosis, and model selection for 3D façade interpretation. *Photogrammetrie Fernerkundung Geoinformation*, *3*, 187–196.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.

Xu, Z., Yang, J., Peng, C., Wu, Y., Jiang, X., Li, R., Zheng, Y., Gao, Y., Liu, S., & Tian, B. (2014). Development of an UAS for post-earthquake disaster surveying and its application in Ms7. 0 Lushan Earthquake, Sichuan, China. *Computers & Geosciences*, *68*, 22–30.

Yang, M. Y., Förstner, W., & Chai, D. (2012). Feature evaluation for building facade images-an empirical study. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; 39-B3*, *39*(B3), 513–518.