SA-FRCNN: An Improved Object Detection Method for Airport Apron Scenes

LYU Zonglei[1,2*], *CHEN Liyun*[1,2]

1. College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, P. R. China;
2. Information Technology Research Base of Civil Aviation Administration of China, Tianjin 300300, P. R. China

(Received 23 February 2021; revised 26 April 2021; accepted 8 August 2021)

Abstract: The airport apron scene contains rich contextual information about the spatial position relationship. Traditional object detectors only considered visual appearance and ignored the contextual information. In addition, the detection accuracy of some categories in the apron dataset was low. Therefore, an improved object detection method using spatial-aware features in apron scenes called SA-FRCNN is presented. The method uses graph convolutional networks to capture the relative spatial relationship between objects in the apron scene, incorporating this spatial context into feature learning. Moreover, an attention mechanism is introduced into the feature extraction process, with the goal to focus on the spatial position and key features, and distance-IoU loss is used to achieve a more accurate regression. The experimental results show that the mean average precision of the apron object detection based on SA-FRCNN can reach 95.75%, and the detection effect of some hard-to-detect categories has been significantly improved. The proposed method effectively improves the detection accuracy on the apron dataset, which has a leading advantage over other methods.

Key words: airport apron scene; object detection; graph convolutional network; spatial context; attention mechanism

CLC number: TP391.4 **Document code:** A **Article ID:** 1005-1120(2021)04-0571-16

0 Introduction

Aircraft stop at the airport apron before or after the flight, and receive a series of flight ground support services, such as embarking and disembarking passengers, loading and unloading cargoes, refueling, catering, cleaning rubbish, towing, etc. Therefore, the apron scene includes multiple categories of objects such as planes, bridge vehicles, platform vehicles, refuel vehicles, luggage vehicles, food vehicles, rubbish vehicles, and tractor vehicles, etc. The process of flight ground support services requires an ordered coordination and deployment of personnel, equipment, and information[1]. Various information of the ground support service process, such as start and end time nodes, is of great significance to the airport. However, the acquisition of the data has relied on the manual filling by ground support staff for a long time. Traditional methods are inefficient, and the accuracy and completeness of the data cannot be guaranteed.

Recently, deep learning algorithms have been increasingly used in airports, and various "smart airports" solutions have been continuously proposed. Utilizing computer vision technology to process the rich surveillance image resources in airports has also become a research hotspot in the development of smart airports. At present, many airports have begun to adopt an automatic data acquisition system of flight ground support service time nodes, which makes full use of the existing video surveillance system and network resources of airports. Various objects such as planes, bridge vehicles, and other ground support vehicles in the apron scene can be

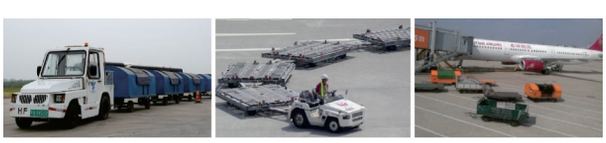
*Corresponding author, E-mail address: zllv@cauc.edu.cn.

How to cite this article: LYU Zonglei, CHEN Liyun. SA-FRCNN: An improved object detection method for airport apron scenes[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2021, 38(4): 571-586.

http://dx.doi.org/10.16356/j.1005-1120.2021.04.004

classified and located on it. Even further, time node information of each ground support service event can be automatically extracted and entered, so as to assist the airport operation and management more efficiently, accurately and intelligently. The core of the system is the apron object detection method.

Although the existing deep learning object detection methods have achieved promising performance on many benchmark datasets, there is still room for improvement of object detection in the specific scene of aprons. On one hand, the traditional object detection methods only focus on the appearance features of the region of interest that may contain the object in the image, however, ignoring the rich context information contained in the image, such as the context of the relative spatial position relationship between the objects. On the other hand, due to the numerous object categories appearing in the apron surveillance images, some different categories of ground support vehicles have similar appearances, which makes more difficult in recognizing them if the detection only relies on local visual appearance. Additionally, there may also be great differences in the appearance of the same category of ground support vehicles. As shown in Fig.1, the three pictures are all a group of luggage vehicles mounted on the trailers, but there are obvious differences between their appearances. Such a series of problems also bring challenges to the apron object detection. To solve these problems, more attention needs to be paid to the spatial context information in the special application scene of aprons.



(a) Type 1 (b) Type 2 (c) Type 3

Fig.1 Luggage vehicles in the apron

When people watch a scene, they usually do not recognize each object individually, but make a correct judgment by considering the surrounding environment. With this ability of integrating context information, human beings can still recognize objects accurately in complex situations. In the special application scene of aprons, all kinds of ground support vehicles have their own basically fixed positions relative to the aircraft according to different tasks. The relationship context between each object, such as the relative spatial relationship, can assist the accurate recognition of apron objects. As shown in Fig.2, in the manual of aircraft support facilities, the servicing diagram of flight ground support vehicles shows the spatial distribution of a series of vehicles when they provide ground services for the aircraft parked in the apron. The particularity of the apron scene restricts the spatial position relationship between objects. To further illustrate the role of spatial context information in apron object detection, we analyze an actual example detected by the faster region-based convolutional neural network (Faster R-CNN)[2] in Fig.3. Since the detector only paid attention to the visual appearance, the object in the enlarged part pointed to by the arrow was mistaken as a luggage vehicle, while it was actually a refuel vehicle. Combining the regulation of the servicing diagram with the position relationship between ob-

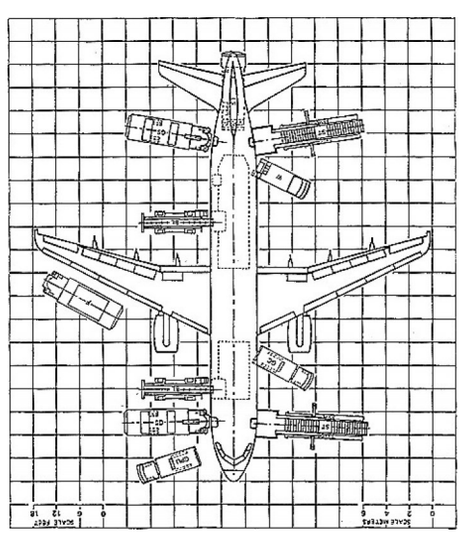


Fig.2 Servicing diagram of flight ground support vehicles

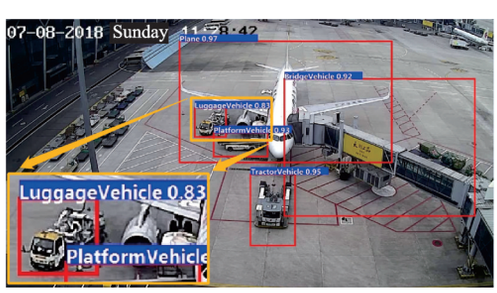


Fig.3 An actual example of Faster R-CNN in the apron

jects, it is clear that the object in this area is most likely to be a refuel vehicle rather than a luggage vehicle. If the relative spatial relationship context between objects is taken into account in the process of object detection, such a mistake could have been avoided as far as possible.

To this end, we consider making use of the spatial context information in airport apron scenes to improve the performance of apron object detection method. In this paper, we propose an improved object detection method called SA-FRCNN utilizing spatial-aware features in the apron scenes, which is a variant of Faster R-CNN. Combined with the particularity of apron application scenes, we use the graph convolutional networks[3] to incorporate the relative spatial position relationship context into the feature representations, via constructing the geometric relationship graph at the region-level, so as to make up for the limitation of only relying on visual appearance for detection. Furthermore, an attention mechanism and an advanced loss function are used to optimize the detection result.

We make the following contributions in this paper. First, we introduce a general method for mining the contextual information of relative spatial position relationships between objects in a given image to enrich the feature representations. Secondly, aiming at the particularity of the apron business scene, we advocate incorporating business rule knowledge into apron object detection, which has a practical application value and enlightening significance for the scene-oriented object detection tasks. Last, we conduct extensive experiments on the apron datasets, and the results illustrate that our proposed method significantly improves the detection accuracy and has a strong competitiveness compared with other methods.

1 Related Work

1.1 Airport apron object detection methods

Currently, object detection methods based on deep convolutional neural networks can be broadly divided into two branches: One is the two-stage methods represented by Faster R-CNN, which as well includes region-based convolutional neural networks (R-CNN)[4], Fast R-CNN[5], region-based fully convolutional networks (R-FCN)[6], etc. In the first stage, region proposals are generated, and in the second stage, the method further performs regional classification and location refinement. Two-stage methods have obvious advantages in the detection accuracy. The other is the one-stage methods represented by you only look once (YOLO) series[7-9], single shot multibox detector (SSD)[10], etc. The core idea is to perform detection by directly regressing the object location and classifying the category in grids of the image. One-stage methods have a higher detection speed, but the detection accuracy is relatively low. Anyway, these methods which only rely on visual appearance features do not consider the relationship between regions, but regard detecting objects in a given image as some isolated tasks.

The study of object detection methods for airport apron scenes has as well drawn the attention of some researchers recently. Zhan et al.[11] generated region proposals based on a streak flow method and corner pooling, and then used convolutional neural networks to identify objects in the airport. Han et al.[12] proposed an improved Faster R-CNN method for airport scenes combined with multi-scale feature fusion and online-hard-example-mining (OHEM). Huang et al.[13] proposed an object detection method based on SSD combined with feature Pyramid fusion network for the detection of small apron objects. A deep neural network compression model for airport apron object detection was proposed[1], which achieved a parameter compression on YOLOv3. The above work has made a promising progress in detecting apron objects. Differently, the application scenes of this paper need to subdivide the categories of vehicles and excavate the spatial context knowledge in the apron scenes as much as possible.

1.2 Relationship context

It is recognized that proper modeling of the relationship context between objects helps humans to understand the scene. Highly aware of this fact, there is some previous work devoted to introducing

the rich contextual knowledge into the computer vision (CV) field. Fang et al.[14] used a knowledge graph for modeling semantic consistency relationship context between objects, and further re-optimized object detection. Jiang et al.[15] mined object-verb-subject relationship and other external relationship context between categories based on label information from the visual genome (VG) dataset, guiding the object detection task. However, these methods cannot be directly applied to object detection for the apron scenes, since they are limited to specific datasets and specific contextual knowledge. In addition, there is some related work in the CV field besides object detection. Wang et al.[16] proposed relationship context-intersection region for scene graph generation. For the visual question answering (VQA) task, Li et al.[17] designed a relational encoder to capture the implicit and explicit relationship context between objects. Yao et al.[18] explored the visual relationship context for image caption task under the umbrella of attention-based encoder-decoder framework.

1.3 Graph convolutional networks

Nowadays, graph convolutional networks become quite a popular tool to deal with graph structured data. The existing graph convolutional networks can be divided into two branches: Spatial-based methods and spectral-based methods. The spatial-based methods[19-21] aggregated each central node and its neighboring nodes by defining an aggregation function. The spectral-based methods[3,22-24] used the graph spectral theory to define graph convolution. Representative methods of spectral-based graph convolutional networks are shown in Table 1.

In recent years, a few researchers have begun to try to apply graph convolutional networks in object detection tasks. Liu et al.[25] integrated graph convolutional networks into the object detection to exploit the benefit of category co-occurrence relationship among objects. Yan et al.[26] proposed a semantics-preserving graph propagation model for zero-shot object detection based on graph convolutional networks. Luo et al.[27] introduced a cascade graph model to exploit multi-scale, cross-modality information for salient object detection. Xu et al.[28] proposed a spatial-aware graph relation network (SGRN) for object detection to incorporate semantic and spatial relationships between objects, achieving improvements on VG dataset and ADE dataset in terms of detection accuracy. Kim et al.[29] proposed an object detection framework based on graph convolutional networks, utilizing spatial relationships between different parts of an object.

Table 1 Representative methods of spectral-based graph convolutional networks

Method name	Advantage	Disadvantage
Spectral CNN[22]	Generalize to non-Euclidean space	Computation-intensive, large disturbance
ChebNet[23]	Local convolution, few parameters	High computational complexity
CayleyNet[24]	Capture frequency bands of interest, flexible	Weak generalization ability
1st ChebNet[3]	Easy to train, practical	Increase in computational cost

1.4 Attention mechanisms in computer vision

Attention mechanisms have been successfully used for computer vision tasks. The attention mechanisms in CV are divided into two major types: Hard attention and soft attention. Their classification is shown in Fig.4.

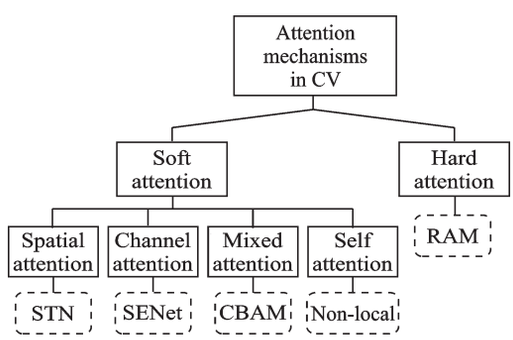


Fig.4 Classification of attention mechanisms in CV

Soft attention is a kind of deterministic attention, which can be generated directly through the network. The spatial attention represented by spatial transformer networks (STN)[30] can capture the spatial position that needs to be paid attention to in the image. Channel attention represented by squeeze-and-excitation networks (SENet)[31] allows

the network to focus on more useful channels and increase learning capabilities. As a combination of them, the convolutional block attention module (CBAM)[32] learns the importance of channels and spatial locations at the same time, which is a simple and efficient module. Non-local[33] uses the self-attention mechanism to model remote dependencies, which is computationally intensive. Differently, the hard attention like recurrent attention model (RAM)[34] is a random prediction process, and every point in the graph may extend the attention, which is achieved by reinforcement learning.

2 Methodology

Aiming at the characteristics of the apron application scene, we propose an apron object detection method called SA-FRCNN. The architecture of the proposed SA-FRCNN is shown in Fig.5.

SA-FRCNN model can be divided into four parts: Feature extraction module, region proposal generation module, spatial relationship modeling module, and regression-classification module. In the apron application scene, the spatial distribution of objects is in a structured form. In the light of the relatively fixed spatial relationship between objects in the airport apron, the graph convolutional network is used to model the spatial relationship context to obtain spatial-aware feature representations. In order to pay more attention to spatial information and key features of the apron objects, and suppress other useless information, we also introduce the convolutional block attention module[32] to redesign the feature extraction network. We apply distance-IoU (DIoU) loss[35] as the bounding box regression loss function, so as to overcome the deficiency of the SmoothL1 loss and further improve the object regression effect.

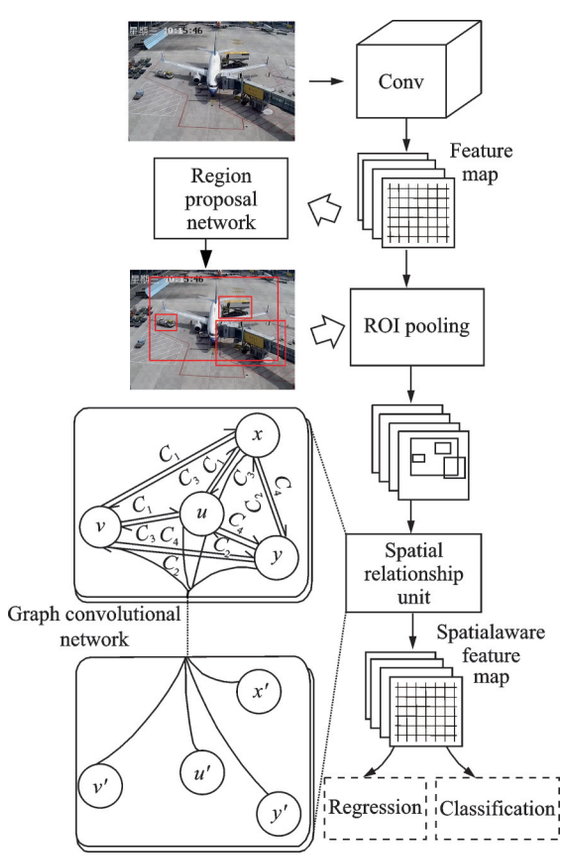


Fig.5 Architecture of the proposed SA-FRCNN

2.1 Spatial-aware feature generation

Through the problem analysis, it is found that the rich context knowledge of relative spatial position relationship between objects in the apron application scene is helpful for apron object detection. Graph convolutional network is a kind of neural network which can process graph structure data and deeply explore its characteristics and rules, deducing the embedding of each object node according to the features of neighbor nodes. Graph convolutional network can automatically learn the features of nodes and the association structure information between nodes. In recent years, it has been widely used in scene graph generation, image understanding and other emerging fields[16-18]. Inspired by these works, we attempt to introduce the graph convolutional network into apron object detection method. Specifically, the spatial relationship modeling module is embedded in the original method, and the graph convolutional network is used to encode the spatial geometric relationship of the region proposals generated by the first stage of Faster R-CNN. Then the region-level advanced feature representations are generated, which are spatial-aware.

In view of the inherent spatial relationship between objects in the apron application scene, we construct the spatial geometric relationship graph structure of the apron. This part is mainly used to obtain the relative spatial position relationship be-

tween region proposals of object detection, which can be formulated as a region-to-region directed graph

$$G=(V,\mathcal{E}) \tag{1}$$

where $V=\{v_i\}_{i=1}^{K}$ denotes the set of K region proposals, here each node $v_i\in\mathbf{R}^{D_v}$ corresponds a region proposal with the D_v-dimensional feature, $\mathcal{E}$ is the set of directed edges between region nodes, and each edge $e_{i,j}\in\mathcal{E}$ denotes the relative spatial geometric position relationship of any two region proposals region_i and region_j.

By observing the distribution regulation of various vehicle operating positions in the servicing diagram of flight ground support vehicles, this paper defines four spatial relationship categories. According to the angle between the line connecting the central points of the two regions and the positive horizontal direction, the corresponding relationship categories are assigned to the edges, which are upper right C_1, upper left C_2, lower left C_3, and lower right C_4. As shown in Fig.6, for two given regions region_i (the solid rectangular box) and region_j (the dashed rectangular box), their spatial positions are denoted by the coordinates (x_i, y_i) and (x_j, y_j) of the central point of their respective bounding boxes. A directed edge $e_{i,j}$ is established from the central point of region_i to the central point of region_j, and the angle between $e_{i,j}$ and the positive horizontal direction is marked as θ_{ij}. $e_{i,j}$ is classified in term of the size of θ_{ij}, and the index of the edge category $L_{(i,j)}$ is taken as

$$L_{(i,j)}=\left\lceil\frac{\theta_{ij}}{90^\circ}\right\rceil \tag{2}$$

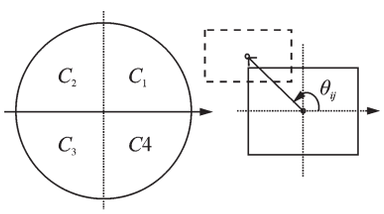


Fig.6 Definition of spatial relationship categories

In this way, the structured relative spatial position relationship information in the apron image is explicitly modeled through a way of constructing a spatial relationship graph. Based on the graph convolutional network, the spatial relationship on the spatial geometric relationship graph is obtained, and the region proposals in the image are context-encoded to generate the new regional feature representations which are optimized to be spatial-aware. The standard graph convolutional network operates on an undirected graph structure. After graph convolution operations, by encoding information about immediate neighbors, the feature representation of each node in the graph is computed as

$$v_i'=\sigma\left(\sum_{v_j\in N(v_i)}\boldsymbol{W}v_j+\boldsymbol{b}\right) \tag{3}$$

where $\boldsymbol{W}\in\mathbf{R}^{D_v\times D_v}$ and $\boldsymbol{b}$ are the weight matrix and the bias vector that need to be learnt, respectively, $N(v_i)$ denotes the set of neighbor nodes of v_i that also contains the node itself, v_i' is a newly generated feature representation, and σ an activation function, generally using the rectified linear unit (ReLU) function.

However, it does not involve any information about directionality or label of edge for encoding the context. The spatial geometric graph G constructed in this paper is a directed graph, and the edge $e_{i,j}$ of the graph structure not only has a direction, but also has different categories of labels $L_{(i,j)}$. Inspired by the work of semantic role labeling[36], in term of the structure of directed graph with labeled edges, we perform the modification on the parameters to be learnt, so as to make the network sensitive to both directionality and label of the edge. After graph convolution operations, the feature representation of each node can be computed as follows

$$v_i'=\sigma\left(\sum_{v_j\in N(v_i)}\boldsymbol{W}_{\text{dir}_{(i,j)}}v_j+\boldsymbol{b}_{L_{(i,j)}}\right) \tag{4}$$

where $L_{(i,j)}$ denotes the category of the edge $e_{i,j}$, $\text{dir}_{(i,j)}$ is used to indicate the direction of information flow, the value of which is 1 when the information flows along v_i to v_j direction, 2 when it flows along the opposite direction of the former, and 3 when it flows in self-loop. When the direction and the category of edge are different, the corresponding weight matrix and bias vector also change.

Considering that the importance of each neighbor node to the central node is "unequal", we add an attention mechanism[37] to the graph convolution

network, so as to make the network automatically focus on learning "more important" associated edge information

$$v_i' = \sigma\left(\sum_{v_j \in N(v_i)} \alpha_{ij}(\boldsymbol{W}_{\mathrm{dir}_{(i,j)}} v_j + \boldsymbol{b}_{L_{(i,j)}})\right) \tag{5}$$

$$\alpha_{ij} = \frac{\exp(\beta_{ij})}{\sum_{m=1}^{K} \exp(\beta_{im})} \tag{6}$$

$$\beta_{ij} = (\boldsymbol{Z} v_i)^{\mathrm{T}} \cdot (\boldsymbol{W}_{\mathrm{dir}_{(i,j)}} v_j + \boldsymbol{b}_{L_{(i,j)}}) \tag{7}$$

where the weight matrix $\boldsymbol{Z} \in \mathbf{R}^{D_v \times D_v}$ is the parameters to be learnt, and β_{ij} the non-normalized attention score calculated by dot product operation, which is used to measure the correlation between two features. α_{ij} denotes the attention coefficient of the whole. In this process, spatial-aware features are generated via context-encoding of the graph convolution network, and each region-level feature after optimization is endowed with the rich spatial relationship context information of apron images.

2.2 Feature extraction network based on attention mechanism

It is well known that humans always selectively concentrate on salient area instead of processing a whole scene at once, which is a significant attention mechanism in the human visual system. Similarly, in the process of feature extraction of convolutional neural networks, not all feature information needs to be equally concerned. Besides, in the process of aggregating context information, it may inevitably bring some redundant information. In order for the network to allocate spatial and channel-wise weights more reasonably when extracting features, and to suppress the influence of complex background and other unnecessary feature information, we introduce the convolutional block attention module to modify the feature extraction network in Faster R-CNN and give the network an ability to properly and selectively "pay attention" to salient visual structures.

With the depth of the deep convolutional neural networks increasing, the ability of the network to map and express image features is enhanced, however, the increase in network complexity causes the problems of gradient disappearance and network degradation at the same time. Hence the residual network came into being. In this paper, we replace the original visual geometry group network-16 (VGG16) network with residual network-50 (ResNet50)[38] as the backbone network. The attention module is a simple and effective lightweight module that can be integrated into the residual network for an end-to-end training. We add the attention module right after each residual block of ResNet50, and redistribute the weights according to the channel importance and spatial distribution characteristics to enhance the high-value feature representations while suppressing possible noise. The structure diagram of the attention module is shown in Fig.7.

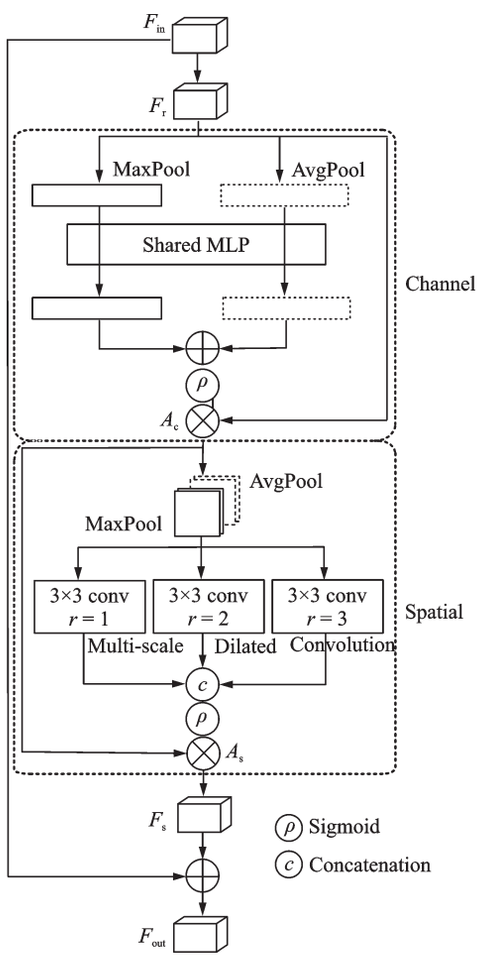


Fig.7 Structure diagram of the attention module

The input feature is represented as $F_{in} \in \mathbf{R}^{H \times W \times C}$, where H and W denote the height and width of the input image, respectively, and C denotes the number of channels. There is an intermediate feature map $F_r \in \mathbf{R}^{H \times W \times C}$ after the residual

module. The channel weight matrix $\boldsymbol{A}_c(\boldsymbol{F}_r)\in\mathbf{R}^{1\times1\times C}$ and the spatial weight matrix $\boldsymbol{A}_s(\boldsymbol{F}_c)\in\mathbf{R}^{1\times1\times C}$ are generated by the channel and spatial attention module. The overall process can be formulated as

$$\boldsymbol{F}_c=\boldsymbol{A}_c(\boldsymbol{F}_r)\otimes\boldsymbol{F}_r \tag{8}$$

$$\boldsymbol{F}_s=\boldsymbol{A}_s(\boldsymbol{F}_c)\otimes\boldsymbol{F}_c \tag{9}$$

$$\boldsymbol{F}_{out}=\boldsymbol{F}_{in}\oplus\boldsymbol{F}_s \tag{10}$$

where $\boldsymbol{F}_c\in\mathbf{R}^{H\times W\times C}$ denotes the feature map generated after $\boldsymbol{F}_r$ going through the channel attention module, and $\boldsymbol{F}_s\in\mathbf{R}^{H\times W\times C}$ the feature map generated after $\boldsymbol{F}_c$ going through the spatial attention module. Ultimately, the final result $\boldsymbol{F}_{out}\in\mathbf{R}^{H\times W\times C}$ is generated by adding $\boldsymbol{F}_s$ and $\boldsymbol{F}_{in}$, $\otimes$ the element-wise multiplication operation, and $\oplus$ the element-wise addition operation.

Channel attention focuses on what features are meaningful. Firstly, we squeeze the spatial dimension of feature maps by max-pooling and average-pooling to aggregate spatial information. Two pooled feature maps based on global information are obtained, which represent the global distribution of the response on the feature channel. Then they are fed into a shared multi-layer perceptron to model the correlation between feature channels, and two $1\times1\times C$ channel attention maps are produced. Sum them element-by-element to get the channel attention weight matrix $\boldsymbol{A}_c(\boldsymbol{F}_r)$, shown as

$$A_c(F_r)=\rho\{\mathrm{MLP}[\mathrm{MaxPool}(F_r)]+\mathrm{MLP}[\mathrm{AvgPool}(F_r)]\} \tag{11}$$

where $\mathrm{MaxPool}(\boldsymbol{F}_r)$ and $\mathrm{AvgPool}(\boldsymbol{F}_r)$ denote max-pooled feature maps and average-pooled feature maps, respectively, MLP denotes features going through a multi-layer perceptron, and ρ is the Sigmoid function. Each dimension in the channel attention weight matrix represents the importance of the corresponding dimension channel to the key information in feature maps.

Spatial attention focuses on the spatial position of salient areas, that is, in which position has the key information. Research results[39] show that performing pooling operations along the channel axis can more effectively highlight the areas containing key information. As a consequence, the max-pooling and average-pooling operations are designed along the channel dimension. Here, two $1\times H\times W$ pooled feature maps are produced. Perform a concatenation on the two output features, and then they undergo a convolution operation to learn the optimized weights of spatial. In order to capture wider-range spatial information, we leverage multi-scale dilated convolution layers with 3×3 kernel size and dilated ratio 1, 2, 3. It can obtain larger receptive field with a small amount of parameters. The process of spatial attention module is formulated as follows

$$\boldsymbol{A}_s(\boldsymbol{F}_r)=\rho(\boldsymbol{D}_1;\boldsymbol{D}_2;\boldsymbol{D}_3) \tag{12}$$

$$\boldsymbol{D}_i=E_i[\mathrm{MaxPool}(\boldsymbol{F}_c);\mathrm{AvgPool}(\boldsymbol{F}_c)]\quad i=1,2,3 \tag{13}$$

where E_i denotes the dilated convolution operation with convolution kernel size of 3×3. The convolution result is denoted as $\boldsymbol{D}_i$, and when $i=1,2,3$, the dilated ratio is 1, 2 and 3, respectively.

2.3 Loss function

The loss function of Faster R-CNN is mainly divided into two parts: RPN loss and Fast R-CNN loss. Both parts include the classification loss and the bounding box regression loss. The only difference is that the former uses two-class cross entropy loss function in the classification process, while the latter uses multiclass of cross entropy loss function. The bounding box regression loss function in Faster R-CNN is SmoothL1 loss, shown as

$$\mathrm{Loss}_{\mathrm{SmoothL1}}=\begin{cases}0.5x^2 & |x|<1\\ |x|-0.5 & |x|\geqslant1\end{cases} \tag{14}$$

$$x=t_i^*-t_i^{\mathrm{gt}} \tag{15}$$

where t_i^* represents the coordinates of the prediction bounding box, t_i^{gt} the coordinates of the ground-truth bounding box, and x the offset between them.

The SmoothL1 loss only considers the distance between the prediction bounding box and the ground-truth bounding box, while it ignores the intersection over union (IoU) of the two bounding boxes, that is

$$\mathrm{IoU}=\frac{|B^*\cap B^{\mathrm{gt}}|}{|B^*\cup B^{\mathrm{gt}}|} \tag{16}$$

where B^* represents the prediction bounding box, B^{gt} the ground-truth bounding box, and $\cap$ and $\cup$ are the intersection and union operations, respectively.

IoU is an important indicator used to evaluate the accuracy of the bounding box in the object detection method, however, the SmoothL1 loss is not a suitable choice to obtain the optimal IoU metric. To solve it, some researchers have proposed to directly calculate the gradient through IoU to perform regression. Even so, the problem is that when there is no intersection between the prediction bounding box and the ground-truth bounding box, there is no any moving gradient for further optimization. Consequently, we use the distance-IoU loss function as the bounding box regression loss function, to replace the SmoothL1 loss in Faster R-CNN, shown as

$$\mathrm{Loss}_{\mathrm{DIoU}} = 1 - \mathrm{IoU} + \frac{\gamma^2(b^*, b^{\mathrm{gt}})}{c^2} \tag{17}$$

$$d = \gamma^2(b^*, b^{\mathrm{gt}}) \tag{18}$$

where $\gamma(\cdot)$ is the Euclidean distance, c the diagonal length of the smallest enclosing box covering two bounding boxes, and $b^*$and b^{gt} are the central points of the prediction bounding box B^* and the ground-truth bounding box B^{gt}, respectively. Compared with SmoothL1 loss, DIoU loss not only directly maximizes IoU, but also minimizes the normalized distance between the central points of two boxes.

3 Experiment

3.1 Experimental environment

Our experiments are conducted on Ubuntu 16.04 operating system; the memory is 8 GB and the configured GPU is a NVIDIA Tesla V100; the deep learning framework used in experiments is the open source library PyTorch, and the language used is Python.

3.2 Dataset

Due to the lack of available public datasets in this field, we construct an airport apron dataset, which is taken from the real apron environment. Subject to the special confidentiality of airport data, current images and annotations of our dataset all come from No.209 stand of Guiyang Airport apron. Both daytime scenario and nighttime scenario are included in the apron dataset, and collected from the surveillance video data at different times of different days. The image resolution is 1 980 pixel× 1 020 pixel, and the size of the dataset is 10 075, involving eight categories in total: Planes (Plane), bridge vehicles (BridgeVehicle), luggage vehicles (LuggageVehicle), refuel vehicles (RefuelVehicle), food vehicles (FoodVehicle), rubbishvehicles (RubbishVehicle), platform vehicles (PlatformVehicle), and tractor vehicles (TractorVehicle). The distribution of the amount of labeled boxes for each category is shown in Fig.8.

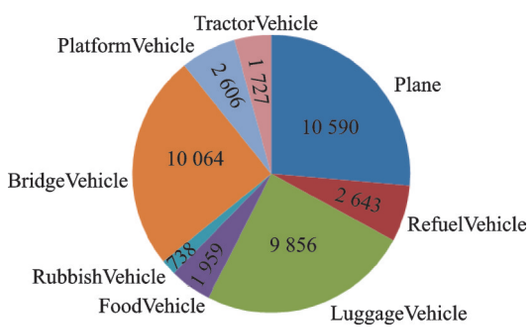


Fig.8 Distribution of the amount of labeled boxes for each category

3.3 Evaluation metrics and implementation details

The performance evaluation metrics for the object detection methods generally use the average precision (AP) and the mean average precision (mAP) of all categories. To measure the object localization accuracy, IoU is used to capture the similarity between the predicted box and the ground truth box. If the IoU of them is greater than the threshold, the object will be considered to be correctly predicted. The performance of objectiveness localization can be evaluated by IoU-based mAP metric. Better mAP indicates better localization[40]. In order to calculate mAP, introduce precision P and recall R, shown as

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{19}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{20}$$

where TP (true positive) represents the number of positive samples actually predicted, FP (false positive) the number of negative samples actually predicted to be positive, and FN (false negative) the

number of positive samples actually predicted to be negative. It is not comprehensive to evaluate the object detector only by precision or only by recall, so it is necessary to evaluate them as a whole. The appropriate performance evaluation metric is AP, shown as

$$\mathrm{AP}=\int_0^1 P(R)\,\mathrm{d}R \tag{21}$$

The average precision can be calculated separately for each category, and then we can calculate the mean of all these averages, which can obtain a mean average precision mAP to measure the performance of the object detection method for all categories, shown as

$$\mathrm{mAP}=\frac{1}{n}\sum_{i=1}^{n}\int_0^1 P(R)\,\mathrm{d}R \tag{22}$$

where n is the sum of the number of all categories contained in the dataset. As for the speed evaluation, frames per second (FPS) is adopted as the metric, representing the number of image frames processed per second.

In our experiments, the apron dataset is divided into training set and test set in the ratio of 3∶1. The network optimizer is stochastic gradient descent (SGD). The training iterations of the network are 100 000 times, and the learning rate of the first 50 000 iterations is 0.002. Then the learning rate is adjusted to 0.000 2, and it is attenuated to 0.000 02 when the iteration exceeds 75 000.

3.4 Experimental results and analysis

In order to verify the superiority of our proposed SA-FRCNN in the performance of the apron object detection task, SA-FRCNN is compared with different object detection methods. In addition to the mainstream object detection methods, we select the SGRN model[28] for comparative experiments, which also takes advantage of graph convolutional networks to process the spatial relationships between objects. Table 2 shows the experimental results of Faster R-CNN[2], SSD[10], YOLOv3[8] and SGRN, and the average precision of each category is calculated. As can be seen from Table 2, SA-FRCNN can identify each category of object in the apron object detection task more accurately.

Through the longitudinal analysis of Table 2, some characteristics and problems of the apron dataset can be found indirectly. Among the current mainstream object detection methods such as original Faster R-CNN, YOLOv3 and SSD, the average precision of rubbish vehicles is relatively low compared with other categories. It is due to the fact that in the apron dataset, the number of object samples of rubbish vehicle category is small, and the object detection network is incapable of fully learning all kinds of feature information of rubbish vehicles as other categories of objects. Owing to the imbalance of the number of categories in the dataset, the results show that the detection effect of rubbish vehicles is poor. Relatively speaking, the number of samples of planes and bridge vehicles in the apron dataset is very sufficient, so that the detection effect of various object detection methods on them is relatively satisfactory. However, for the category of luggage vehicles, despite the fact that there are a large number of object samples in the apron dataset, the detection effect is still very poor. After analysis, it is concluded that the reason for this problem may be the large differences within the luggage vehicle category: As mentioned above, there are different types of luggage vehicles in the apron, giving rise to the obvious differences in their appearance; in addition, whether and how much luggage is loaded as well has a certain impact on the appearance of the luggage vehicles. As a consequence, the category of luggage vehicles is more difficult to detect for the traditional detectors which mainly rely on visual appearance, which leads to a relatively low AP.

Through the horizontal analysis of Table 2, it can be seen that SSD and YOLOv3 are obviously worse in detection accuracy, especially for the detection of some small object categories, such as luggage vehicles, tractor vehicles and so on. The two methods obtain 77.84% and 79.85% mAP, respectively. Faster R-CNN has a significant advantage in accuracy. Compared with the above three methods, SGRN also utilizes the graph convolutional networks to integrate the spatial context and improve mAP. Especially, for individual categories such as food vehicles and rubbish vehicles, SGRN achieves better performance than SA-FRCNN. Presumably it is because that SGRN ad-

ditionally captures a certain semantic relationship. However, as a whole, SA-FRCNN has a higher detection accuracy than SGRN, which means that our SA-FRCNN can better apply apron scenes and take advantage of the spatial relationship context in the apron more effectively. The mean average precision of SA-FRCNN for eight categories of objects is improved to 95.75%, which is about 17.91%, 15.9%, 5.14% and 2.62% higher than that of the above four methods, respectively. It illustrates that the proposed SA-FRCNN improves the detection performance with better localization and classification.

Table 2 Experimental results of different object detection methods %

Parameter	Method	Faster R-CNN	SSD	YOLOv3	SGRN	SA-FRCNN
AP	Plane	94.63	91.52	91.90	97.66	99.33
	BridgeVehicle	90.87	91.84	93.31	93.74	98.61
	LuggageVehicle	82.14	58.46	61.23	86.56	93.55
	RefuelVehicle	93.85	74.96	74.14	94.53	97.60
	FoodVehicle	90.91	91.33	94.70	92.42	92.34
	RubbishVehicle	87.85	71.75	79.86	90.04	89.57
	PlatformVehicle	91.63	75.76	78.18	92.81	95.84
	TractorVehicle	93.00	67.10	65.51	97.28	99.17
mAP		90.61	77.84	79.85	93.13	95.75

In order to verify the effectiveness of the spatial relationship modeling (SRM) module proposed in this paper, we design a comparative experiment. Table 3 shows the experimental results before and after using the spatial relationship modeling module, where SRM_Faster R-CNN denotes that the spatial relation modeling module is introduced into the baseline network.

Table 3 Experimental results before and after introducing spatial relationship modeling module %

Parameter	Method	Faster R-CNN	SRM_Faster R-CNN
AP	Plane	94.63	98.45
	BridgeVehicle	90.87	96.83
	LuggageVehicle	82.14	89.22
	RefuelVehicle	93.85	95.92
	FoodVehicle	90.91	91.56
	RubbishVehicle	87.85	88.71
	PlatformVehicle	91.63	93.25
	TractorVehicle	93.00	97.21
mAP		90.61	93.89

It can be seen from Table 3 that the average precision values of eight categories of planes, bridge vehicles, luggage vehicles, refuel vehicles, food vehicles, rubbish vehicles, platform vehicles and tractor vehicles have been improved with varying degrees after the addition of spatial relationship modeling module: The average precision value of luggage vehicle category in particular has been improved by 7.08%. This is because after the SRM module encodes the spatial context, the network captures the relative spatial position relationship between the objects, so as to obtain the spatial-aware features. It makes up for the limitation of the traditional object detector before the improvement which only relies on visual appearance to identify the category of luggage vehicles. The optimized region-level features are endowed with rich spatial-aware relationship information. As a consequence, despite the large differences within the luggage vehicle category, our proposed SA-FRCNN can still achieve a satisfactory detection performance.

Nevertheless, compared with other categories, the improvement of the average precision values of food vehicles and rubbish vehicles are relatively less, increased by 0.65% and 0.86% respectively. In trying to analyze this situation combined with the apron application scene, it is found that the operation positions of the two ground support vehicles are almost the same, so that the improvement effect of two categories of vehicles after integrating spatial position relationship information is not as apparent as that of other categories. On the whole, adding SRM module makes the object detector more in line with the law of human context perception, which in-

creases the overall mAP value of the baseline network by 3.28%. It indicates that capturing the relative spatial position relationship between the objects can enhance the detection ability for the apron objects.

To thoroughly evaluate the effectiveness of each modified module, we also conduct extensive ablation experiments. As shown in Table 4, the first group uses the baseline network, and the spatial relationship modeling module is added to the baseline network in the second group. In the third group, attention module is added to further improve the feature extraction network. In the fourth group the DIoU loss function is used to replace the SmoothL1 loss function on the basis of the third group.

Table 4 Ablation experiment results

Group	SRM	Attention	DIoU loss	mAP/%
1				90.61
2	√			93.89
3	√	√		94.92
4	√	√	√	95.75

It can be seen from Table 4 that the network can achieve a better mean average precision value of 93.89% on the apron dataset after adding the spatial relationship modeling module into the baseline network. On this basis, the mean average precision value is increased by 1.03% to 94.92% after integrating the attention mechanism into the feature extraction network. This result shows the effectiveness of the attention module in the apron object detection task. The attention module makes the network pay more attention to the spatial position and key features of each object in the apron scene, and suppresses the influence of complex background and other useless feature information inevitably brought in the process of context integration. After the SmoothL1 loss function is replaced by the DIoU loss function, the mAP value is further increased to 95.75%. The introduction of DIoU loss makes the model better fit the position relationship between the predicted bounding box and the ground-truth bounding box, and makes further efforts to improve the detection ability and prediction accuracy of the network.

Fig.9 shows the comparison of AP value of each category in different groups of ablation experiments.

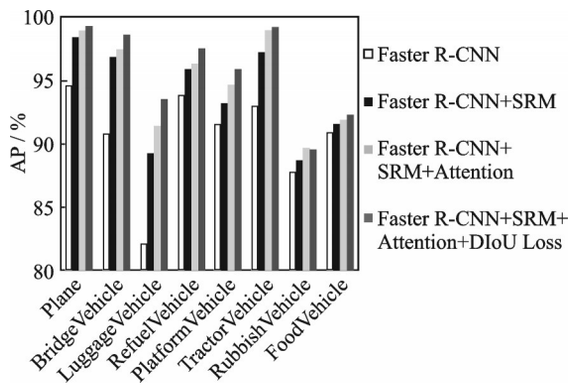


Fig.9 Comparison of AP value of each category in ablation experiments

As can be seen from Fig.9, the improvement effect of the spatial relationship modeling module is the most evident among the several modules, which also verifies the guiding significance of excavating the rich relative spatial position relationship context among the objects in the apron application scene for the object detection task. Due to the integration of the attention module into the feature extraction network, the network can learn more representative features, which makes the detection accuracy of smaller objects than plane, such as tractor vehicles, platform vehicles, luggage vehicles, and so on, has been significantly improved. The introduction of DIoU loss function has better effect on the detection of bridge vehicles, platform vehicles and refuel vehicles. For rubbish vehicles and food vehicles, the improvement effect is still relatively slight. It is speculated that the reason is the two categories of objects have the same operation position and similar appearance. This result also brings some new enlightenment: In the future, we can consider adding the temporal context information of operation time sequence of different ground support vehicles to assist the apron object detection task. From the perspective of the overall trend of the histogram chart, each module has achieved varying degrees of improve-

ment in the average precision of detection, and the proposed SA-FRCNN can significantly improve the detection effect of apron object detection task.

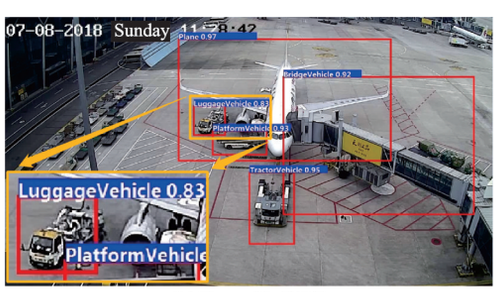
The purpose of this paper is aiming at improving the apron detection accuracy under the premise that the speed meets the actual application requirements. In order to evaluate the detection speed, we compare FPS between SA-FRCNN and the baseline Faster R-CNN. The comparison result has been tabulated in Table 5.

Table 5 FPS of two object detection methods

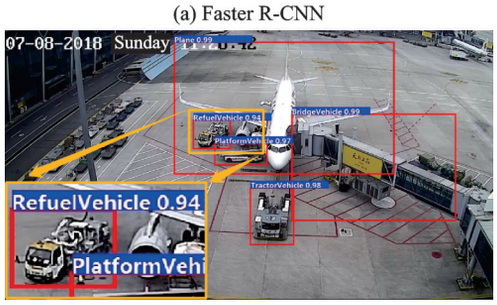
Method	Faster R-CNN	SA-FRCNN
FPS	10	7

It can be seen from Table 5 that SA-FRCNN increases the time cost due to the introduction of modified modules. Considering that the current civil aviation system usually records data in minutes, the difference in detection speed between SA-FRCNN and other methods will hardly affect airport operation. The proposed method meets the actual application requirements.

We also visualize the results of object detection on the apron dataset, as shown in Figs. 10, 11. In Fig. 10, the refuel vehicle is mistakenly identified as a luggage vehicle by Faster R-CNN. Since our SA-FRCNN considers the context of the spatial position

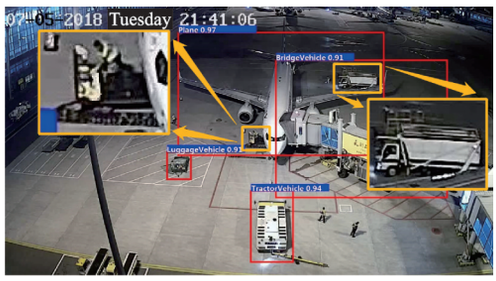


(a) Faster R-CNN

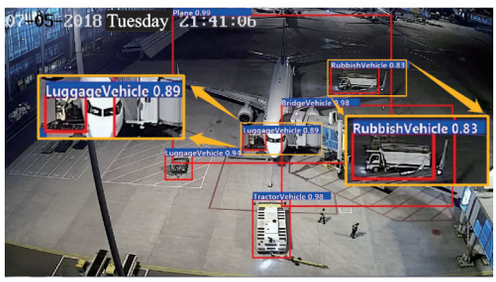


(b) SA-FRCNN

Fig.10 Visualization results in the apron at daytime



(a) Faster R-CNN



(b) SA-FRCNN

Fig.11 Visualization results in the apron at nighttime

relationship between the objects on the apron scenes, the object is correctly detected as a refuel vehicle according to the spatial distribution of the objects. In Fig.11(a), Faster R-CNN fails to detect a rubbish vehicle and one of the luggage vehicles, while SA-FRCNN in Fig.11(b) has ability to correctly recognize and locate the rubbish vehicle and the luggage vehicle. In addition, it can be found that compared with the baseline network, SA-FRCNN can improve the prediction confidence of each object in the apron and locate the objects more accurately. Fig.10 and Fig.11 are the daytime scenario and nighttime scenario of the apron, respectively. No matter in the daytime or in the nighttime with a relatively low contrast, SA-FRCNN can detect each object accurately with a better performance. In general, the visualization results prove the rationality and effectiveness of SA-FRCNN for airport apron object detection task.

4 Conclusions

This paper proposes an improved object detection method (SA-FRCNN) for airport apron scenes, which takes the particularity of the apron application scene into account. We innovatively make full use of the spatial-aware features which provide abundant spatial position relationship context information between apron objects in the detection pro-

cess. Additionally, an attention mechanism and DIoU loss are also used. Extensive experiment results show that our SA-FRCNN achieves more accurate localization and classification results on the apron dataset, and the situation of false detection and missed detection has also been rectified, which illustrates the success of introducing airport special business scenario knowledge to assist the apron object detection task. The proposed method is greatly instructive and referential for solving scene-oriented object detection tasks.

References

[1] LYU Z, PAN F, XU X. A novel deep neural network compression model for airport object detection[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2020, 37(4): 562-573.

[2] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[3] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2017-02-22) [2021-03-20]. https://arxiv. org/pdf/1609.02907.pdf.

[4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH: IEEE, 2014: 580-587.

[5] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Washington DC: IEEE, 2015: 1440-1448.

[6] DAI J, LI Y, HE K, et al. R-FCN: Object detection via region-based fully convolutional networks[EB/OL]. (2016-05-20) [2021-03-20]. https://arxiv.org/pdf/1605.06409.pdf.

[7] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 7263-7271.

[8] REDMON J, FARHADI A. Yolov3: An incremental improvement[EB/OL]. (2018-04-08) [2021-03-20]. https://arxiv.org/pdf/1804.02767.pdf.

[9] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[EB/OL]. (2020-03-23) [2021-03-20]. https://arxiv.org/pdf/2004.10934.pdf.

[10] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[11] ZHAN Zhaohuan, HAN Songchen, LI Wei, et al. A target detection method of moving objects at airport based on streak flow and deep learning[J]. Journal of Transport Information and Safety, 2019, 37(1): 49-57. (in Chinese)

[12] HAN Songchen, ZHANG Bihao, LI Wei, et al. Small target detection in airport scene via modified Faster-RCNN[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2019, 51(6): 735-741. (in Chinese)

[13] HUANG Guoxin, LI Wei, ZHANG Bihao, et al. Improved SSD-based multi-scale object detection algorithm in airport surface[EB/OL]. (2021-01-27) [2021-03-20]. http://kns. cnki. net/kcms/detail/11.2127.TP.20210127.1142.016.html.(in Chinese)

[14] FANG Y, KUAN K, LIN J, et al. Object detection meets knowledge graphs[C]//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne Australia: International Joint Conferences on Artificial Intelligence Organization, 2017: 1661-1667.

[15] JIANG C, XU H, LIANG X, et al. Hybrid knowledge routed modules for large-scale object detection[EB/OL]. (2018-10-30) [2021-03-20]. https://arxiv.org/pdf/1810.12681.pdf.

[16] WANG W, WANG R, SHAN S, et al. Exploring context and visual pattern of relationship for scene graph generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 8188-8197.

[17] LI L, GAN Z, CHENG Y, et al. Relation-aware graph attention network for visual question answering[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S. l.] : IEEE, 2019: 10313-10322.

[18] YAO T, PAN Y, LI Y, et al. Exploring visual relationship for image captioning[C]//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer, 2018: 684-699.

[19] JUSTIN G, SAMUEL S S, PATRICK F R, et al. Neural message passing for quantum chemistry[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney Australia: IMLS, 2017: 2053-2070.

[20] PETAR V, GUILLEM C, ARANTXA C, et al. Graph attention networks[J]. Lecture Notes in Computer Science, 2019, 11731(2005): 566-577.

[21] CHIANG W L, LIU X Q, SI S, et al. Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Anchorage, USA: ACM, 2019: 257-266.

[22] JOAN B, WOJCIECH Z, ARTHUR S, et al. Spectral networks and deep locally connected networks on graphs[C]//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR, 2014: 1-14.

[23] MICHAËL D, XAVIER B, PIERRE V. Convolutional neural networks on graphs with fast localized spectral filtering[C]//Proceedings of the 30th Annual Conference on Neural Information Processing Systems. Barcelona, Spain: NIPS, 2016: 3844-3852.

[24] RON L, FEDERICO M, XAVIER B, et al. Cayleynets: Graph convolutional neural networks with complex rational spectral filters[J]. IEEE Transactions on Signal Processing, 2019, 67(1): 97-109.

[25] LIU Z, JIANG Z, FENG W, et al. OD-GCN: Object detection boosted by knowledge GCN[C]//Proceedings of IEEE International Conference on Multimedia & Expo Workshops (ICMEW). [S.l.]: IEEE, 2020: 1-6.

[26] YAN C, ZHENG Q, CHANG X, et al. Semantics-preserving graph propagation for zero-shot object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 8163-8176.

[27] LUO A, LI X, YANG F, et al. Cascade graph neural networks for RGB-D salient object detection[C]// Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 346-364.

[28] XU H, JIANG C, LIANG X, et al. Spatial-aware graph relation network for large-scale object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2019: 9298-9307.

[29] KIM J U, PARK S, RO Y M. Towards human-like interpretable object detection via spatial relation encoding[C]//Proceedings of 2020 IEEE International Conference on Image Processing (ICIP). [S.l.]: IEEE, 2020: 3284-3288.

[30] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2015: 2017-2025.

[31] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[32] WOO S, PARK J, LEE J, et al. CBAM: Convolutional block attention module[C]//Proceedings of the 2018 European Conference on Computer Vision. Berlin: Springer, 2018: 3-19.

[33] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7794-7803.

[34] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention[C]//Proceedings of the 2014 Advances in Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2014: 2204-2212.

[35] ZHENG Z, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI Press, 2020: 12993-13000.

[36] MARCHEGGIANI D, TITOV I. Encoding sentences with graph convolutional networks for semantic role labeling[EB/OL]. (2017-07-30) [2021-03-20]. https://arxiv.org/pdf/1703. 04826.pdf.

[37] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-10-06) [2021-03-20]. https://arxiv.org/pdf/1706.03762.pdf.

[38] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington DC, USA: IEEE, 2016: 770-778.

[39] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[EB/OL]. (2017-02-12) [2021-03-20]. https://arxiv.org/pdf/1612.03928.pdf.

[40] LU C, LU Y, CHEN H, et al. Square localization for efficient and accurate object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2015: 2560-2568.

Acknowledgement This work was supported by the Fundamental Research Funds for Central Universities of the Civil Aviation University of China (No.3122021088).

Author Dr. **LYU Zonglei** was born in Tianjin, China, in

1981. He received the B.S. degree in computer science and technology from Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2004, and the Ph. D. degree in computer application technology from NUAA, Nanjing, China, in 2009. In 2009, he joined College of Computer Science and Technology, Civil Aviation University of China, as a Lecturer, and in 2012 became an associate professor. His current research interests include machine learning, deep learning, and object detection.

Author contributions Dr. **LYU Zonglei** designed the study, complied the models, and conducted the analysis. Ms. **CHEN Liyun** conducted the analysis, interpreted the results and wrote the manuscript. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: ZHANG Huangqun)

SA-FRCNN:一种改进的机场停机坪目标检测方法

吕宗磊[1,2], 陈丽云[1,2]

(1.中国民航大学计算机科学与技术学院,天津 300300,中国;
2.中国民航信息技术科研基地,天津 300300,中国)

摘要:机坪场景下包含丰富的空间位置关系上下文信息。传统目标检测器往往只关注单一的视觉外观而忽略上下文信息;此外机坪数据集中部分类别识别准确率较低。针对上述问题,提出一种改进的机场停机坪目标检测方法,称为SA-FRCNN。该方法利用图卷积网络来捕获机坪场景下目标间的相对空间关系,将空间位置关系上下文融入模型生成空间感知特征;在特征提取过程中引入注意力机制,聚焦机坪目标的空间位置和关键特征;使用距离交并比损失实现目标更精确地回归定位。实验结果表明,SA-FRCNN方法在机坪数据集上目标检测均值平均精准度达到95.75%,部分较难检测类别的检测效果提升显著;有效提高机坪目标检测的准确性,较其他方法具有领先优势。

关键词:机坪场景;目标检测;图卷积网络;空间上下文;注意力机制