Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 7258649, 15 pages
https://doi.org/10.1155/2021/7258649

Research Article

EAWNet: An Edge Attention-Wise Objector for Real-Time Visual Internet of Things

Zhichao Zhang,[1] **Hui Chen**,[2] **Xiaoqing Yin**,[3] **and Jinsheng Deng**[3]

[1]*College of Computer, National University of Defense Technology, Changsha 410000, China*
[2]*Science and Technology on Integrated Logistics Support Laboratory, National University of Defense Technology, Changsha 410000, China*
[3]*College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha 410000, China*

Correspondence should be addressed to Xiaoqing Yin; yinxiaoqing89@163.com

Received 22 April 2021; Accepted 9 June 2021; Published 12 July 2021

Academic Editor: Fa Zhu

With the upgrading of the high-performance image processing platform and visual internet of things sensors, VIOT is widely used in intelligent transportation, autopilot, military reconnaissance, public safety, and other fields. However, the outdoor visual internet of things system is very sensitive to the weather and unbalanced scale of latent object. The performance of supervised learning is often limited by the disturbance of abnormal data. It is difficult to collect all classes from limited historical instances. Therefore, in terms of the anomaly detection images, fast and accurate artificial intelligence-based object detection technology has become a research hot spot in the field of intelligent vision internet of things. To this end, we propose an efficient and accurate deep learning framework for real-time and dense object detection in VIOT named the Edge Attention-wise Convolutional Neural Network (EAWNet) with three main features. First, it can identify remote aerial and daily scenery objects fast and accurately in terms of an unbalanced category. Second, edge prior and rotated anchor are adopted to enhance the efficiency of detection in edge computing internet. Third, our EAWNet network uses an edge sensing object structure, makes full use of an attention mechanism to dynamically screen different kinds of objects, and performs target recognition on multiple scales. The edge recovery effect and target detection performance for long-distance aerial objects were significantly improved. We explore the efficiency of various architectures and fine tune the training process using various backbone and data enhancement strategies to increase the variety of the training data and overcome the size limitation of input images. Extensive experiments and comprehensive evaluation on COCO and large-scale DOTA datasets proved the effectiveness of this framework that achieved the most advanced performance in real-time VIOT object detection.

1. Introduction

Intelligent vision internet of things (VIOT) uses all kinds of image sensors, including surveillance cameras, mobile phones, and digital cameras, to obtain people, cars, objects, images, or video data; extract visual tags; and use intelligent analysis technology to process information, so as to provide support for follow-up applications as shown in Figure 1. The intelligent visual internet of things can directly, vividly, and efficiently reflect the monitoring data of the observed object and output the results of intelligent analysis. Therefore, VIOT is widely used in important places such as social public safety, intelligent vehicles, parking lots, community monitoring, land and sea traffic monitoring, urban security, and military reconnaissance. However, the performances of supervised learning are often limited by the disturbance of abnormal data and unbalanced scale of latent objects, which impair the automatic inference speed and recognition accuracy.

The intelligent visual internet of things can provide information assistance for public security departments, such as real-time monitoring, suspect tracking, and crime early warning. At the same time, it can also provide a large number of real-time traffic information for traffic management

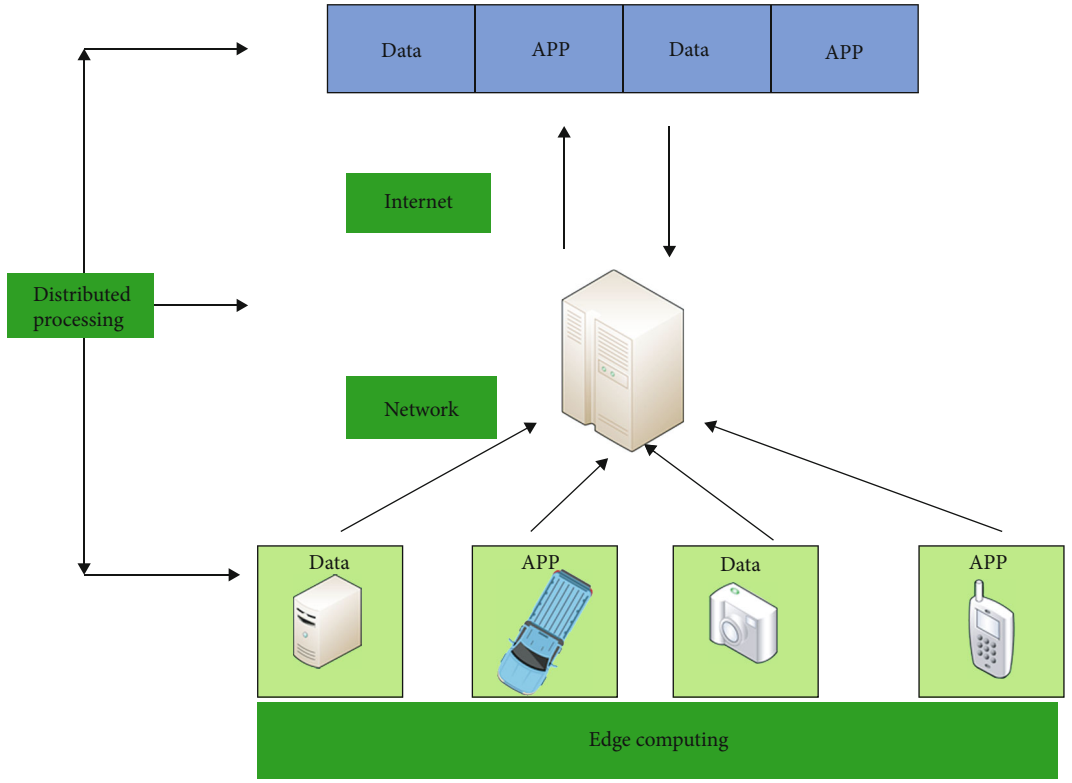


Figure 1: The whole object detection process of intelligent vision internet of things (VIOT).

departments to facilitate their traffic supervision. There is no doubt that the emergence of the internet of things in intelligent vision has brought great convenience to people's lives. The intelligent visual internet of things system needs to extract accurate image features in the application.

However, the object detection task for an outdoor internet of things vision system is very sensitive to weather conditions, especially in frequent and widely distributed blurry scenes. In addition, when vehicles and people move quickly, the captured images may be blurry. Out-of-focus cameras can also lead to a decrease in detection accuracy. Finding specific key information from traffic surveillance, astronomical remote sensing, public security investigation, and other applications therefore remains a significant challenge. Because of the lack of information, these low-quality images seriously affect the effectiveness of the intelligent visual internet of things system.

In order to conquer these challenges, we need more advanced object detectors. Advanced object detection methods have greatly improved over the past few years, and several methods have been introduced to optimize the network structure, which can be divided into single-stage and double-stage; however, the use of an attention module to improve the efficiency of searching is not well investigated. They are divided into two mainstreams: two-stage detector and single-stage detector. Two-stage detector: the R-CNN [1] directly performs the selective search [2] and classifies objects using a CNN. Compared with the traditional methods, the use of the R-CNN significantly improved the accuracy of classification, marking the beginning of the era of target detection using deep learning. In its variants (for example, fast R-CNN [3]), the two-phase framework was updated, which helps them achieve even better performance. In addition, to further improve the accuracy, some highly effective extensions were proposed, such as R-FCN [2] and mask R-CNN [4]. Single-stage detector: The most representative single-stage detectors are YOLO [5, 6] and SSD [7]. They use feature maps to predict the confidence and location of a multitarget receptive field block (RFB) network to achieve accurate and fast object detection. Both these detectors use lightweight backbones for acceleration; however, their accuracy clearly lags behind the top two-stage approach. Recently, more advanced single-stage detectors (such as DSSD [8] and RetinaNet [9]) have updated their original lightweight backbone using a deeper ResNet-101 and by applying some techniques, such as deconvolution [8] or caustics [9]. Their results are comparable to or even better than most advanced two-stage methods. However, these detectors could not achieve the balance of speed and accuracy simultaneously.

In view of the abovementioned limitations, we propose to design a generative edge detection method to perceive the object structure, make full use of attention mechanism to dynamically screen different kinds of objects, and perform object detection on multiple scales. By doing so, not only does the edge recovery effect become better, but also the object detection performance for long-distance aerial objects is significantly improved.

We design a lightweight single-stage attention rotation intelligent object detection network for wireless internet of

things. Inspired by the previously proposed methods, our framework, using an edge attention-wise conventional neural network, EAWNet, and based on attention mechanism, has the following advantages:

(i) Edge prior depiction greatly reduces the amount of attention-aware computation. We propose an edge attention-wise network beneficial for extracting features effectively and reducing the ground truth position shrinkage. The framework performs well in terms of accuracy (state-of-the-art stability for multiclass) and speed (real-time video recognition). Edge prior reconstruction and attention-wise modules are embedded into the EAWNet, which helps in performing efficient latent search and localization

(ii) With the combination of intelligent connection and residual link, rotating bounding box, and synthesis loss function, the visual loss of intensive detection is reduced to a minimum. Pass-wise connection follows a straight way to pass the initial patch information to different last stage fusion layers to restore the recognition fusion. It propagates semantically strong features and enhances all features with a reasonable classification capability. In addition, residual connections for local convolutional layers and pass-wise connections for global feature dataflow are designed to modify the architecture for faster and lighter inference

(iii) A dual parallel attention module is used to improve the efficiency of multiscale object detection. Attention-wise modules including context attention-wise module (CAW) and position refinement-wise module (PRW) are designed to reduce computing cost and improve effectiveness. These modules can match the right object and position instead of searching the entire background. A rotating bounding box, designed for aerial image object detection, proved to be beneficial for the recognition of dense and tiny objects

2. Related Work

2.1. Object Detection. In the wave of artificial intelligence sweeping the world, the intelligent visual internet of things is expected to achieve a significant social and economic promotion of the internet of things. It has become a typical successful representative of the application of the internet of things. Object detection methods are mainly based on CNNs; one-stage object detectors play a remarkable role in object detection. Most existing VIOT object detectors are classified according to whether they have suggested steps for regions of interest (two-stage [3, 4, 10]) or not (one-stage [6, 7, 11]). Although two-stage detectors are more flexible and accurate, single-stage detectors are generally considered faster and more efficient by using pretrained anchors [12]. Single-stage detectors have attracted wide attention because of their high efficiency and simplicity. Here, we mainly follow the design of single-stage detectors and prove that higher efficiency and higher accuracy can be achieved by optimizing the network structure.

A recent single-stage detector [7, 13] was designed to match the accuracy of more complex two-stage detection methods. Although these detectors show impressive results on large- and medium-sized objects, their performance on small objects is lower than expected [14]. (The size of an object is related to the pixels it occupies in the picture.) When using the most advanced single-stage RetinaNet [13], it achieves unbalanced results with a COCO AP-large of 47 but only 14 for AP-small objects (as defined in [15]). Small object detection is a challenging problem, which requires not only low intermediate information to accurately describe it but also high-level semantics to distinguish the target from the others or background.

There are five types of YOLO from YOLOv1 to YOLOv5 [16, 17]. Based on YOLOv1 to YOLOv3 [18], researchers propose an efficient and powerful object detection model called YOLOv4 [19]. A variety of modules are mixed in the YOLOv4, such as Weighted Residual Connections (WRC), Cross Stage Partial (CSP) connections, Cross Minibatch Normalization (CmBN), Self-Adversarial Training (SAT), Mish-activation, Mosaic data augmentation, CmBN, DropBlock regularization, and CIoU loss. The introduction of these modules increased the calculation time yet greatly improved the accuracy. While YOLOv5 shows the fastest speed among the series of YOLO algorithms and is comparable to YOLOv4 in terms of accuracy, YOLOv5 is remarkably lightweight.

The following methods are the most classic deep learning object detection methods from different schools in the past two years. RFBNet [20] simply employs dilated convolutions to enlarge the receptive field and achieves good vision of the extracted features; although it could be learnable for image recognition, its performance is not satisfactory. LRF [15] is a lightweight scratch network (LSN) that is trained from scratch taking a downsampled image as input and passing it through a few convolutional layers to efficiently construct low- and middle-level features. However, learning from both scratch and pretrained sacrifices too much time efficiency, and the network is so complex that the inference speed is slower than in one-stage methods. CenterMask [21] combines instance segmentation and object detection into one task, and it goes even further by using the instance segmentation to achieve the recognition of class and position that object detection demands. EfficientDet [22] balances the network depth for speed and feature flow connection strategy for accuracy, ignoring the attention mechanism to enhance the performance without occupying large resources.

Considering the mentioned advantages and weaknesses, we propose an attention-wise YOLO, which handles the feature extraction flow in a reasonable way by designing a multi-path refinement framework. In addition, pass-wise connection meets the demands in terms of balancing time efficiency and prediction accuracy. To learn semantic information wisely, attention-wise modules are introduced between the backbone and the neck.

2.2. Attention Module. The main focus of attention models in computer vision is to focus on interesting things and ignore

irrelevant information. Recently, attention models have been classified into three groups: hard attention [23] and soft attention [16], global attention [16, 17] and local attention [24, 25], and self-attention [16]. Hard attention models have been widely used for a long model without preprocessing. The computational cost of local attention is lower than that of global attention because it does not need to consider the hidden layer state of all encoders. The self-attention mechanism improves the attention model, which reduces the dependence on external information and is capable of capturing internal correlation of data features. As self-attention shows good performance, it is widely used on computer vision tasks.

The attention model mechanism is important in deep learning methods. The first to propose the self-attention mechanism were Vaswani et al. [25]. It relies on global dependencies between inputs and outputs and was applied in machine translation. In computer vision area, attention modules have been also adopted. Zhang et al. [26] created an image generator that leverages adjacent regions to object shapes rather than local regions of fixed shape for image generation by the self-attention mechanism. An adapted attention module for object detection that uses the relative geometry between objects was proposed [27]. There is a successful application in space-time dimension for videos with nonlocal operation [28]. Fu et al. designed DANet [29] based on the newly Fully Convolutional Networks (FCNs) [30] with position attention mechanism (PAM) and channel attention mechanism (CAM). DANet settles the problems of object detection in some confusing categories and objects with different appearance. In addition, Fu et al. [31] proposed a DRANet which makes an improvement in self-attention modules based on DANet. DRANet adopts the compact PAM (CPAM) and the compact CAM (CCAM), reducing computational complexity.

2.3. Detection Bounding Boxes. Current object detection algorithms may not perform good results on detecting oriented targets [32]. State-of-the-art object detection methods rely on rectangular-shaped, horizontal/vertical bounding boxes drawn over an object to accurately localize its position. Such orthogonal bounding boxes ignore object pose, resulting in reduced object localization, and limiting downstream tasks such as object understanding and tracking. Rotated faster R-CNN [33] based on the faster R-CNN [34] adds a regression branch to predict the oriented bounding boxes for aerial images. It could improve the performance on tiny things in high resolution by introducing balanced FPN. R4Det [35] is an end-to-end detector which could address the problems of images with large aspect ratios, dense distributions, and extremely imbalanced categories. Moreover, from experimental results, we could see that the detector shows strong robustness against adversarial attacks.

3. Implementation

3.1. Network Architecture. The whole network architecture is shown in Figure 2. We propose a heavily reconstructed Edge Attention-wise Convolutional Neural Network (EAWNet). It employs the multipath refinement flow network (MRFNet) [36] as the backbone, which makes it easier to extract multilevel scale features from patches. We consider not only the efficiency between the backbone and neck through MRFNet but also the fusion effect of extracting features aided by a pass-wise connection (PWC) strategy. Furthermore, this framework is learnable for multiclass and multiscale objects because we design various attention-wise modules to make the training and validation more reasonable and extract features in a global perspective. Also, we modify the above model by adding rotating bounding boxes and design a synthesis loss function to constrain the training process and to boost the training convergence using the multipath refinement flow network (MRFNet).

The MRFNet architecture is shown in Figure 2(b). It combines each convolution layer and dataflow branches. Specifically, there are two path channels $a_0 = [a_0', a_0'']$. Every stage has a downsampled fusion layer, $[a_0', a_1, \cdots, a_k]$, which will be downsampled to lower dimensions and larger output numbers. Then, the output of this refinement transfer results in a_τ, which will be concatenated with a_0' and undergo another transition layer to finally generate the output a_U. Equations (1) and (2) are the feed-forward pass and weight update of MRFNet, respectively. w_k, w_τ, and w_U are the weights of ground truth patches g_0', g_k, and g_0'', respectively. f_k, f_T, and f_U are the transformation function of downsampled layers, transfer results, and transition layer outputs, respectively.

$$\begin{aligned} a_k &= w_k * [a_{0'}, a, \cdots, a_{k-1}], \\ a_\tau &= w_\tau * [a_{0''}, a_1, \cdots, a_k], \\ a_U &= w_U * [a_{0'}, a_\tau], \end{aligned} \tag{1}$$

$$\begin{aligned} w_K' &= f_k(w_k, \{g_{0''}, g_1, \cdots, g_{k-1}\}), \\ w_T' &= f_T(w_T, \{g_{0''}, g_1, \cdots, g_k\}), \\ w_U' &= f_U(w_U, \{g_{0'}, g_T\}). \end{aligned} \tag{2}$$

We compared our architecture with mainstream CNN architectures (ResNeXt, ResNet, and DenseNet). The results are usually a linear and nonlinear combination of the outputs of intermediate layers. Thus, the output of a k-layer CNN is

$$y = F(a_0) = a_k = H_k(a_{k-1}, H_{k-1}(a_{k-2}), H_{k-2}(a_{k-3}), \cdots, H_1(a_0), a_0), \tag{3}$$

where the whole model of the convolutional neural network is denoted by F, the mapping function from a_0 to y, and H_k is the k-th layer of CNN. Generally, a set of convolutional layers and a nonlinear activate function consist of the H_k. As for ResNet and DenseNet, we can also formulate their models into the following:

$$a_k = R_k(a_{k-1}) + a_{k-1} = R_k(a_{k-1}) + R_{k-1}(a_{k-2}) + \cdots + R_1(a_0) + a_0, \tag{4}$$

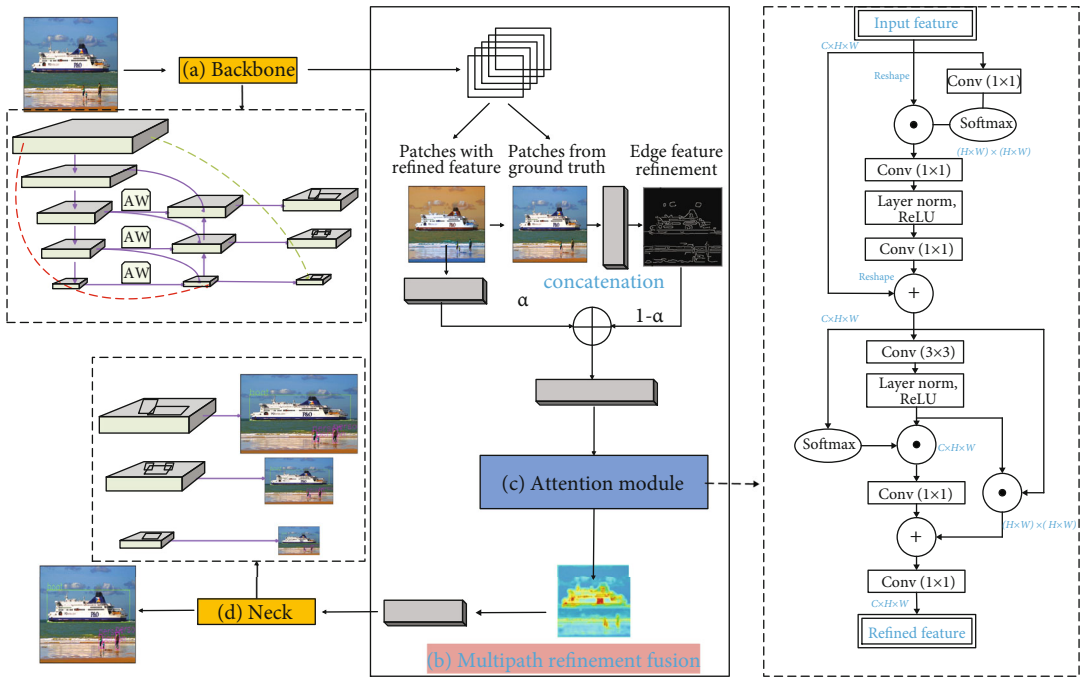


FIGURE 2: Overall network architecture of EAWNet comprising four parts: (a) backbone architecture, (b) multipath refinement fusion unit, (c) attention modules, and (d) neck consisting of the network. In addition, inside the dotted lines is the explicit implementation of the counterparts. (a) The backbone is responsible for the multiscale feature extraction and is optimized by pass-wise connection to avoid gradient disappearance. (b) The multipath refinement fusion (MRF) unit is responsible for making fusion from the edge prior which is extracted from the ground truth and refined patches. (c) The attention modules learn information of category and structure wisely quickly aided by the edge prior. The position-wise and channel-wise attention modules (in the dotted lines) consist of the attention modules in a parallel manner. (d) The neck is the decoder for object detection which is modified into rotated bounding boxes for better visual effects. All four parts are illustrated in the following sections.

$$a_k = [C_k(a_{k-1}), a_{k-1}] = [C_k(a_{k-1}), C_{k-1}(a_{k-2}), \cdots, C_1(a_0), a_0], \tag{5}$$

where R and C represent the operational computation of the residual layers and convolutional layers, respectively; both are reproduced by two or three convolutional blocks.

From the above equations, it follows that the inputs of convolutional layers originated from the previous convolutional outputs. Under this circumstance, the gradient flow could be propagated more efficiently due to the minimum path length of the gradient. However, this design would result in the reverse propagation into all layers from $k-1$ to 1, which is redundant for a repeated training process. Figure 3 illustrates the EAWNet reusing the initial features and simultaneously preventing iteratively propagating gradient information by cutting down the gradient flow. The insightful vision of the design is to separate gradient flow and refinement features and fuse the last convolutional layers, which enhances feature extraction efficiency.

The specific multipath refinement flow network exhibits the advantage of multiscale feature extraction as RefineNet [36] and CSPDarkNet-53 [19], deeply modified by introducing lightweight and residual strategy, pass-wise connection, and attention modules to enhance speed and accuracy.

3.2. Pass-Wise Connection. In this section, we design two extra paths to pass features to the next extraction stage: pass-wise and residual connections. The main purpose of pass-wise and residual connections is to learn robust features and train deeper networks. They can address gradient vanishing problems and enhance the capabilities of locating positions and propagating strong responses of low-level patterns.

It is based on the fact that high-level features responding to edges or instance parts are a strong indicator to accurately localize instances. To this end, regardless of the complicated multipath refinement dataflow, we additionally add two direct connections to pass feature maps. As depicted in Figure 3, one line in red directly passes the first layer patches of the backbone to the last layer of the neck. Another line in green directly passes the first convolutional results to the last layer of data augmentation. The data augmentation layer and the neck layer extract features in a parallel way, sacrificing memory usage to enhance the accuracy and benefit feature extraction.

3.3. Attention-Wise Module. At present, of the multimodel method can be used for semantic segmentation and object detection [37], but the multimodel method leads to too much training cost.

We keenly find that when the edge generation method is used to assist the attention module to perceive the object structure, the object category is guided and searched by dynamically adjusting the receptive field of the recognition

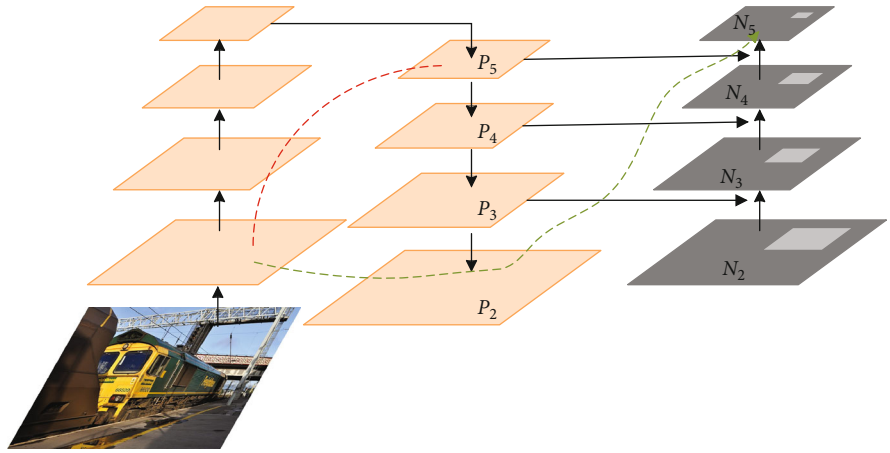


Figure 3: Pass-wise connection benefits the feature extraction and fusion.

frame, and then, the target is detected. This not only improves the performance of object detection but also overcomes the performance loss of semantic segmentation guiding the attention module. Therefore, we propose the EAWNet network structure, which perceives the object structure in a lightweight way through edge description and uses the attention module to improve the search efficiency. Finally, with the help of the multiscale network structure for object detection, it achieves the purpose of improving the training convergence detection performance.

We use a multipath refinement fusion (MRF) unit to fuse the information from the edge prior that is extracted from the ground truth and refined patches. Then, attention modules learn category and structure information and are quickly aided by the edge prior. The position-wise and channel-wise attention modules (depicted in the dotted lines of Figure 2(c)) consist of the attention modules in a parallel configuration.

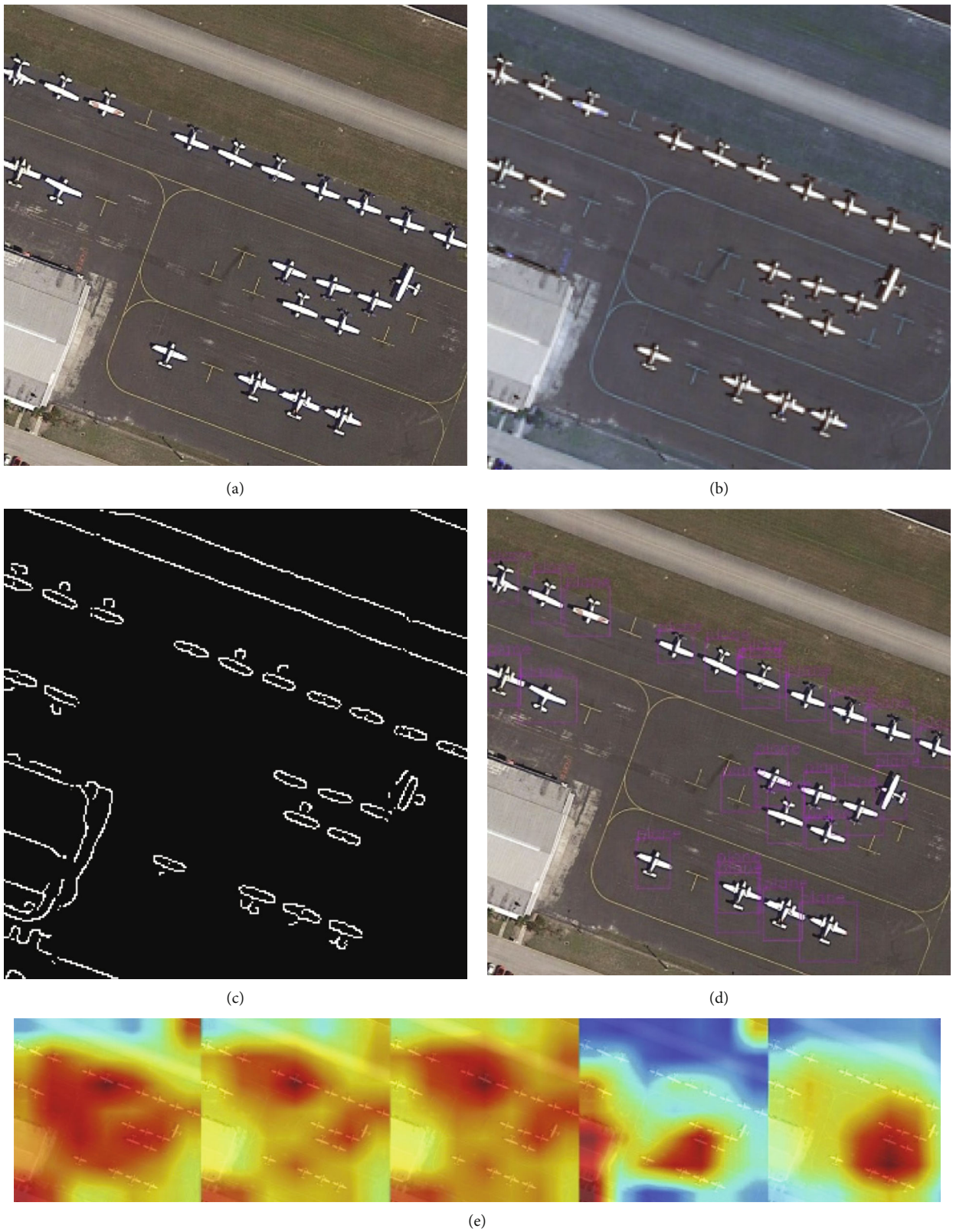
3.4. Edge Prior and Attention Mechanism. Attention plays an important role in human visual recognition [30, 38, 39]. An important feature of the human visual system is that people do not try to deal with the whole scene at once. Instead, to better capture the visual structure, humans use a series of local glimpses and selectively focus on significant parts [40]. We propose a residual attention network using encoding and decoding attention modules. By improving the feature mapping, the network not only has good performance but also has strong robustness to noise input. Instead of calculating attention scores directly, we decompose the process into a learning channel and position attention information. The individual attention score generation process of the feature map is less than that of [37]; thus, it can be regarded as a plug-and-play module for the existing basic convolutional neural networks. Hu et al. [41] introduced a compact module to take advantage of the relationship between channels. In their squash and trigger modules, they used the global average set feature to calculate channel attention. We find that these are suboptimal features, and we recommend using the maximum set feature. They also missed out on spatial attention, which plays an important role in determining the "location" of focus, as shown in [42]. Here, we utilize both channel and position attention based on an effective architecture and verify that using these two kinds of attention is better than using channel attention only [41]. Experiments show that the model is effective in detection tasks (MS-COCO and DOTA). We only need to place our module on the existing single detector [33] in the DOTA test set to achieve the most advanced performance.

Among the edge detection methods, Canny usually has the best edge restoration effect on small local objects, holistically nested edge detection (HED) [16] has the best edge restoration effect on the whole contour, and the edge of generative edge restoration training is often slightly intermittent, but the overall complex structure is the best, so we choose generative edge restoration for joint training.

On the one hand, one-stage detectors aim to handle images in a lightweight manner, making instance recognition fast and easy. On the other hand, one vital property of a visual CNN system is that it does not attempt to handle the whole scene at once. We propose to resolve this contradiction by adding the attention module between the backbone and the neck. The whole network MRF belongs in the one-stage method, while also using the attention-wise (AW) modules to preprocess the feature patches and assessing the contextual and position information several times in a multiscale manner. The design of contextual attention-wise unit and position refinement-wise unit is depicted in Figure 2(c).

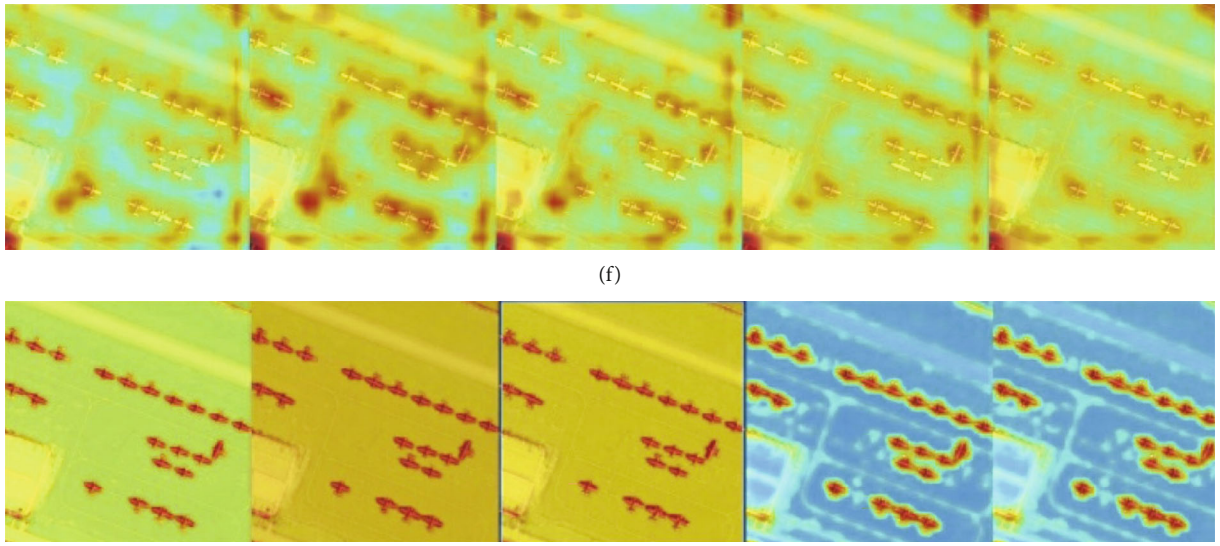
Local features generated by traditional convolutional layers would result in misrecognition of specific things. It could also lead to a high computation cost searching for specific objects from the background. To model rich contextual relationships over local features, we continue to analyze the refinement features from the context attention-wise unit and pass the patches to the position refinement-wise unit to figure out the coherency transformation to the latent position. The position refinement-wise unit enhances their representation capability by encoding a wider range of channel and spatial information into local features,

A target associated with background scenarios could be grasped by the contextual attention module. In the meantime, object positions are located by a position refinement-wise module. Specifically, the attention results could be seen



(a) (b) (c) (d) (e)

Figure 4: Continued.



(f)

(g)

Figure 4: Multifeature extraction for edge and sharpness.

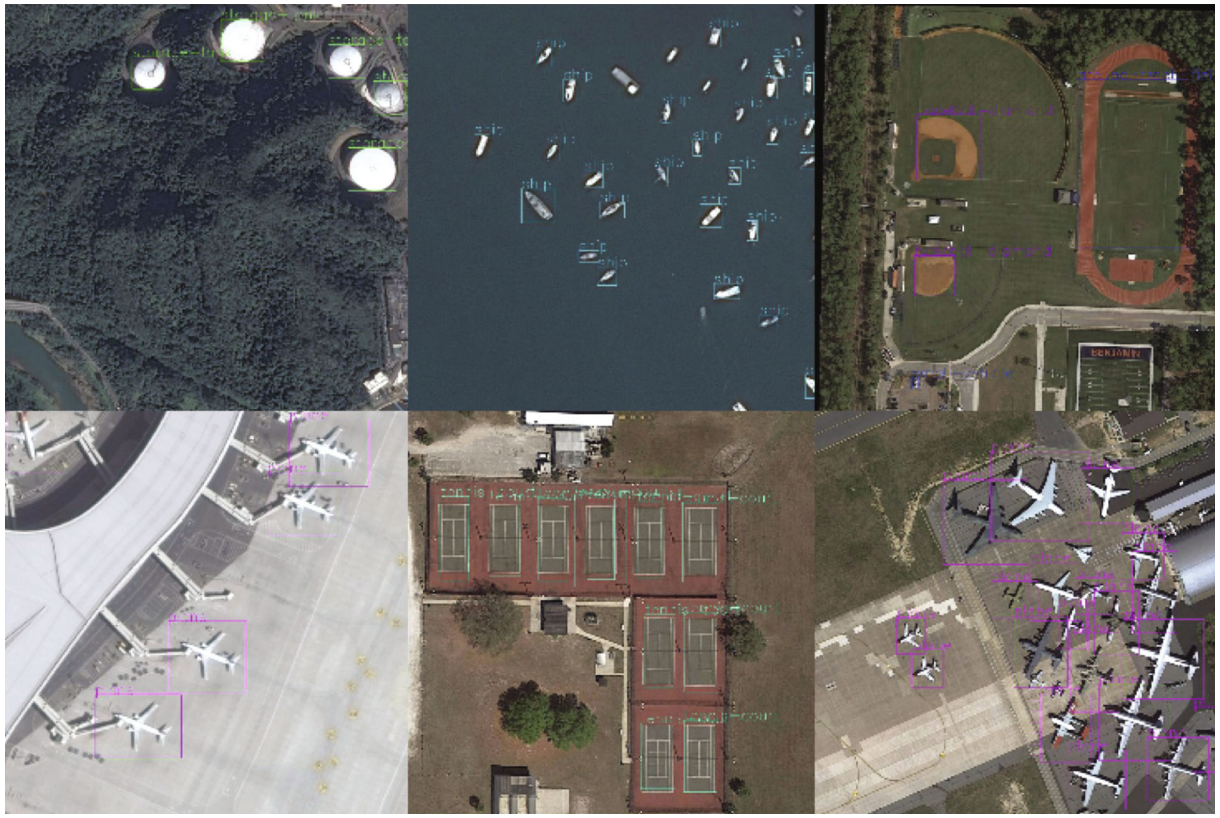


Figure 5: Our validation visual results on DOTA using EAWNet backbones.

in Figures 4(e), 4(f), and 4(g). First, the multipath refinement patches flow into the context attention-wise module, which is aimed at calculating the coherency between the bounding box and the background. This unit simplifies the channel, height, and width ($C \times H \times W$) patches into a softmax function and then combines the copies with matrix multiplication. We apply a softmax layer to obtain the context attention map m_{ij}:

$$m_{ij} = \frac{\exp\left(P_i \cdot P_j\right)}{\sum_{i=1}^{C} \exp\left(P_i \cdot P_j\right)}. \tag{6}$$

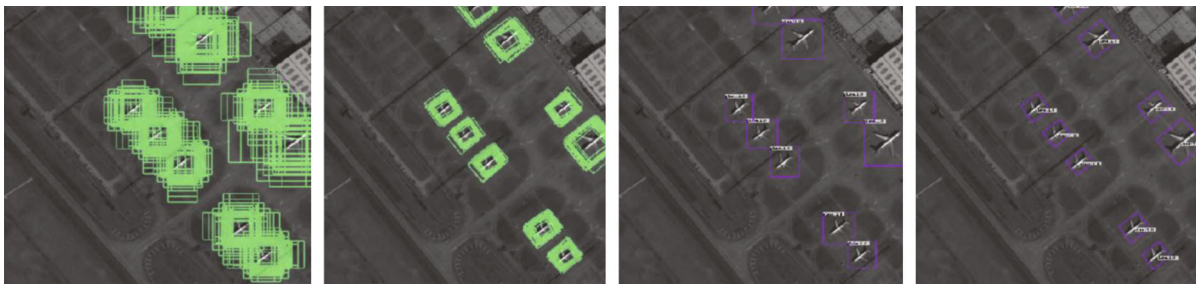


Figure 6: Visual comparison of general and rotated training processes.

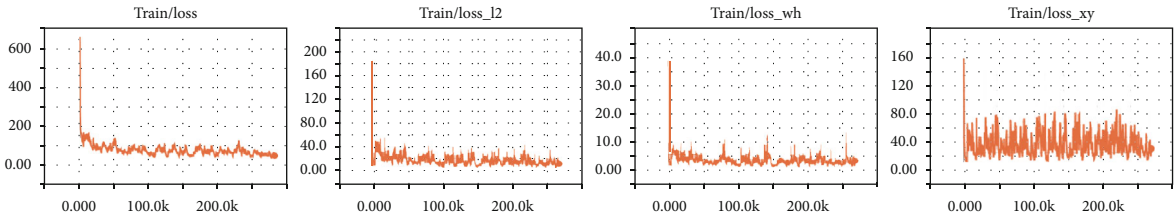


Figure 7: Different training loss function strategies are adopted in the training process which is beneficial for convergence speed.

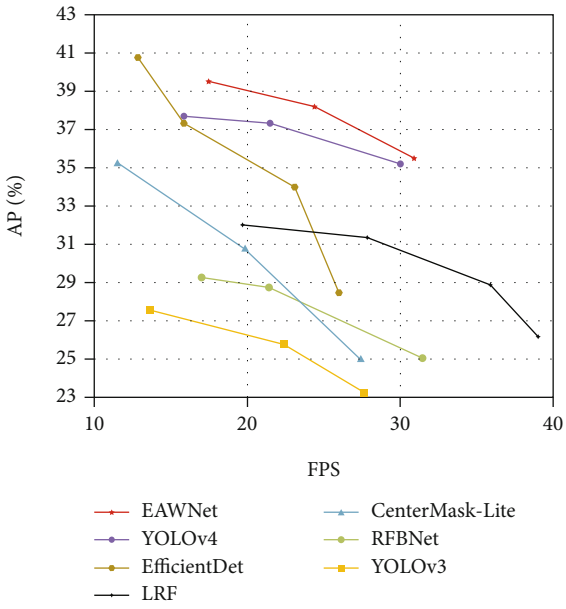


Figure 8: Our model EAWNet shows excellent performance in balancing average precision accuracy and frames per second speed compared to the SOTA methods.

The original patches P_i and reshaped branches P_j are aggregated into an average. In addition, we multiply the result by a scalable item α and add m_{ij} to obtain the output result R_{ij}:

$$R_{ij} = \alpha \sum_{i=1}^{C} \left(x_{ij} P_i\right) + m_{ij}. \tag{7}$$

The output result R is separated into the width, abscissa, and ordinate: W_i, X_j, Y_i; then, W_i and x_{ij} are reshaped and sent to the softmax function and combined as R_{ij}:

$$R_{ij} = \frac{\exp\left(W_i \cdot X_j\right)}{\sum_{i=1}^{N} \exp\left(W_i \cdot X_j\right)}. \tag{8}$$

Meanwhile, the Y_i is also combined with the reshaped S_{ij} using the sum function. We multiply it by a scalar item α and do an element-wise sum operation with the features Y_i to obtain the result $S_{ij} \epsilon R_{ij}^{C \times H \times W}$ as follows, and A_j is a fixed constant:

$$S_{ij} = \alpha \sum_{i=1}^{N} \left(R_{ij} Y_i\right) + A_j. \tag{9}$$

According to recent studies of the single object detectors, there are three ways to obtain the features that concern us. Firstly, the network uses a softmax function to weight the importance of the latent meaningful objects obtained from the background.

Then, the algorithm uses the location and class coherency to get the attention scores. This is not a promising method because the weak supervised methods cannot achieve high enough detection precision. Secondly, the object detection and the instance segmentation are combined; however, the abundant feature extraction and training computational cost are too high. Thirdly, we combine these two branches and follow a trade-off strategy: we adopt the lightweight spatial and class feature extraction channels to recognize the latent object classes, and then, the attention features are refined by the edge information to reinforce the boundary features for further inference. In this way, we use the edge information instead of instance segmentation to avoid the iterative training process and its computational cost. Thus, the

Table 1: Our model is attention-wise. The LRF also uses learnable strategies and acts more lightweight and fast; however, the accuracy is much lower than that of EAWNet. CenterMask learns in an efficient way by searching from the object center and achieves balanced performance. However, EAWNet showed a more significant improvement in learning strategy (adding attention module and rotated tight bounding boxes makes a significant progress on reconstructed deep learning models) and does achieve comparable results to similar algorithms such as LRF, RFBNet, CenterMask, EfficientDet, and YOLOv3. We can conclude that EAWNet outperforms most existing methods in terms of both accuracy and speed. The percentage of average precision on category on DOTA shows our model performs well on unbalanced and anomaly data categories.

Method	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA
RFBNet [20]	40.57	10.21	1.68	14.12	1.32	1.43	2.19	17.22	28.57	10.34	28.26	10.11	4.12
LRF [15]	40.59	21.29	37.74	24.20	9.93	2.19	5.86	45.44	39.45	35.72	17.22	38.73	48.34
CenterMask [21]	90.60	81.97	6.57	67.08	71.12	79.66	79.16	91.81	86.26	85.42	62.91	64.77	69.12
EfficientDet [22]	90.02	82.31	47.11	72.86	72.96	78.34	80.54	91.96	85.14	85.62	57.69	62.13	65.25
YOLOv4 [19]	91.13	82.13	50.28	72.64	72.78	80.43	80.47	91.89	85.76	85.73	60.12	62.64	68.09
EAWNet	90.08	86.56	54.01	74.94	76.75	82.52	81.32	91.83	87.96	86.34	65.14	61.85	70.17

Table 2: Ablation studies of network architecture (size 512 × 512).

Model	AP (%)	AP50 (%)	AP75 (%)
MRFNet [36]	37.1	58.2	38.2
MRFNet+PWC	37.3	58.1	39.8
MRFNet+RC	37.6	58.6	41.5
MRFNet+PWC+RC	36.9	59.1	44.7
EAWNet	37.9	59.7	45.2

Table 3: Average precision for ablation experiments of attention modules (size 512 × 512).

Model (with optimal setting)	AP (%)	AP50 (%)	AP75 (%)
MRFNet [36]	37.6	59.8	41.3
MRFNet+CAW	37.5	59.6	41.0
MRFNet+PRW	37.5	59.3	41.2
MRFNet+CAW+PRW	37.6	60.2	41.5

attention-wise modules also attain the location and coherency information with lightweight and refinement.

As shown in Figure 2(b), we can consider the training process in the feature extraction view instead of the network dataflow. Patches with convolutional refined features are combined with the edge feature patches which are extracted from the ground truth counterparts. We control the proportion by α in Equation (7) of the edge information and the background recognition feature extraction.

Then, the fusion middle results are sent to the attention-wise modules and finally make the inferences for object detection average precision.

Figures 4(b) and 4(c) show the input images and the edge feature maps and the attention heat maps as middle outputs which could benefit the EAWNet for efficient recognition. When we change the perceptive field radius for different object scales, heat map visualization shows the dense small objects. Owing to the smart design of network feature extraction and efficient searching latent attention strategy, this rotated attention-wise network EAWNet achieves fast real-time recognition speed and significant precision enhancement among the state-of-the-art methods.

3.5. Rotated Bounding Box and Loss Design. Horizontal and vertical bounding boxes are drawn over an object for accurate localization. However, for dense VIOT object detection, the anchors are close and boundaries are overlapped. Therefore, we designed rotating bounding boxes to obtain tighter and more precise detections.

We use five parameters x, y, w, h, θ to represent the location of the rotating bounding boxes. If (x, y) are the coordinates of the center of the latent object, wh are its width and height, and θ is the angle of rotation in polar coordinate, then t is the angle of each coordinate:

$$t_x = \frac{x - x_a}{\omega_a}, t_y = \frac{y - y_a}{h_a},$$
$$t_\omega = \log\left(\frac{\omega}{\omega_a}\right), t_h = \log\left(\frac{h}{h_a}\right), t_\theta = \theta - \theta_a, \tag{10}$$

$$t'_x = \frac{x' - x_a}{\omega_a}, t'_y = \frac{y' - y_a}{h_a}, t'_\omega = \log\left(\frac{\omega'}{\omega_a}\right), t' = \log\left(\frac{h'}{h_a}\right), t'_\theta = \theta' - \theta_a. \tag{11}$$

x is the anchor boxes, and $x^{'}$ is a prediction of bounding boxes. Thus, the loss function is expressed as

$$L = \frac{\lambda_1}{N} L_{\text{attention}} + \frac{\lambda_2}{N} L_2 + \frac{\lambda_3}{N} \mathrm{L}_{\mathrm{AW}} + \frac{\lambda_4}{N} L_{\mathrm{CLS}} + \frac{\lambda_5}{N} L_{\mathrm{Obj}} + \frac{\lambda_6}{N} L_{X,Y,W,H}, \tag{12}$$

where N denotes the number of anchors and the hyperparameters λ_k control the trade-off setting to one by default [1, 43]. The classification loss L_{CLS} is implemented by focal loss and smooth L_2 loss. L_{Obj} is the object detection loss. L_{AW} is the attention-wise loss. We also add the xy loss and wh loss $L_{X,Y,W,H}$ for the bounding box position precision and the object loss to analyze how many objects are missing. Figure 5 shows that the model is trained to detect dense and small objects in real-time VIOT fast and accurately. Thus, the rotated bounding boxes and hybrid loss function design are

Table 4: Item FPS and AP of different object detectors.

Method	Backbone	Size	FPS	AP (%)	AP50 (%)	AP75 (%)	APs (%)	APm (%)	APl (%)
YOLOv4: optimal speed and accuracy of object detection [19]									
YOLOv4	CSPDarknet-53	416	30	35.6	57.8	38.2	17.3	39.2	52.1
YOLOv4	CSPDarknet-53	512	22	37.9	60.0	41.9	19.8	41.5	49.8
YOLOv4	CSPDarknet-53	608	16	38.2	60.9	42.5	21.7	42.1	47.4
Learning rich features at high speed for single-shot object detection [15]									
LRF	VGG-16	300	39.0	26.8	46.7	29.4	8.3	30.3	42.6
LRF	ResNet-101	300	36.2	29.2	50.0	32.2	8.6	33.2	45.9
LRF	VGG-16	512	27.9	31.9	51.7	33.8	14.7	35.4	44.3
LRF	ResNet-101	512	19.7	32.6	53.2	35.1	15.2	38.0	45.4
Receptive field block net for accurate and fast object detection [20]									
RFBNet	VGG-16	300	32.0	25.3	44.5	27.1	6.9	27.2	41.3
RFBNet	VGG-16	512	22.5	29.2	50.1	31.2	11.7	32.4	42.5
RFBNet-E	VGG-16	512	17.3	29.6	50.7	31.5	12.9	31.8	42.7
YOLOv3: an incremental improvement [18]									
YOLOv3	Darknet-53	320	27	23.3	46.8	25.1	7.2	25.8	38.4
YOLOv3	Darknet-53	416	23	26.4	50.6	27.8	10.3	28.2	38.2
YOLOv3	Darknet-53	608	14	28.0	53.0	29.6	13.7	30.7	36.9
YOLOv3-SPP	Darknet-53	608	16	31.2	56.2	33.5	15.6	33.4	41.4
CenterMask: real-time anchor-free instance segmentation [21]									
CenterMask-Lite	MobileNetV2-FPN	600x	27	25.5	—	—	9.2	27.1	36.3
CenterMask-Lite	VoVNetV-19-FPN	600x	20	31.2	—	—	14.9	33.0	41.2
CenterMask-Lite	VoVNetV-39-FPN	600x	12	35.7	—	—	17.8	38.6	48.5
EfficientDet: scalable and efficient object detection [22]									
EfficientDet-D0	Efficient-B0	512	26	29.0	47.2	31.2	7.3	33.3	46.2
EfficientDet-D1	Efficient-B1	640	23	34.6	53.8	37.5	13.1	39.8	51.0
EfficientDet-D2	Efficient-B2	768	16	38.2	57.3	41.2	17.9	42.4	53.6
EfficientDet-D3	Efficient-B3	896	13	41.3	60.6	44.6	21.6	45.1	55.1
EAWNet: an Edge Attention-wise Convolutional Neural Network for real-time object detection									
EAWNet	EAWNet	416	31	36.1	59.3	39.1	18.5	40.0	53.3
EAWNet	EAWNet	512	24	38.8	60.8	42.7	20.8	42.4	50.7
EAWNet	EAWNet	608	17	39.7	62.2	43.3	23.1	42.8	48.6

FPS, AP, AP50, AP75, APs, APm, and APl represent the frame per second, average precision, reach to 50% average precision, reach to 75%, average precision, and average precision for small-, medium-, and large-scale objects, respectively.

beneficial for better visual effect and training process convergence as displayed in Figures 6 and 7.

4. Experiments

We conducted comparative experiments between EAWNet, RFBNet [20], LRF [15], YOLOv3 [18], CenterMask [21], and EfficientDet [22] in FPS, AP, and visual effects.

Frames per second (FPS) are raised by 6.25%, and AP average precision (AP) is increased by 1.51%. The results obtained with other state-of-the-art object detectors are displayed in Figure 8. Our EAWNet on the red line is on the Pareto optimality curve and is the fastest and most accurate detector.

4.1. Experimental Setup. We implemented our model with PyTorch. The model was trained with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$). A batch size of 16 was used for training in four NVIDIA RTX2080Ti GPUs with 11 GB RAM. At the beginning of each epoch, the learning rate was initialized as 10^{-4} and subsequently diminished by half every 10 epochs. We trained 100 epochs on COCO and 150 epochs on DOTA.

4.2. Dataset and Augmentation. We assessed our method on two well-known benchmarks in VIOT for city and aerial scenarios: COCO [44] and DOTA [45]. The comparative experiments were performed under equivalent conditions (training on same GPU and dataset). The image size from DOTA was 1024×1024, while that from COCO was 256×256. The COCO benchmark is a large-scale object detection,

Figure 9: Some visual result tests on EAWNet show that the attention modules benefit prediction stability and training precision.

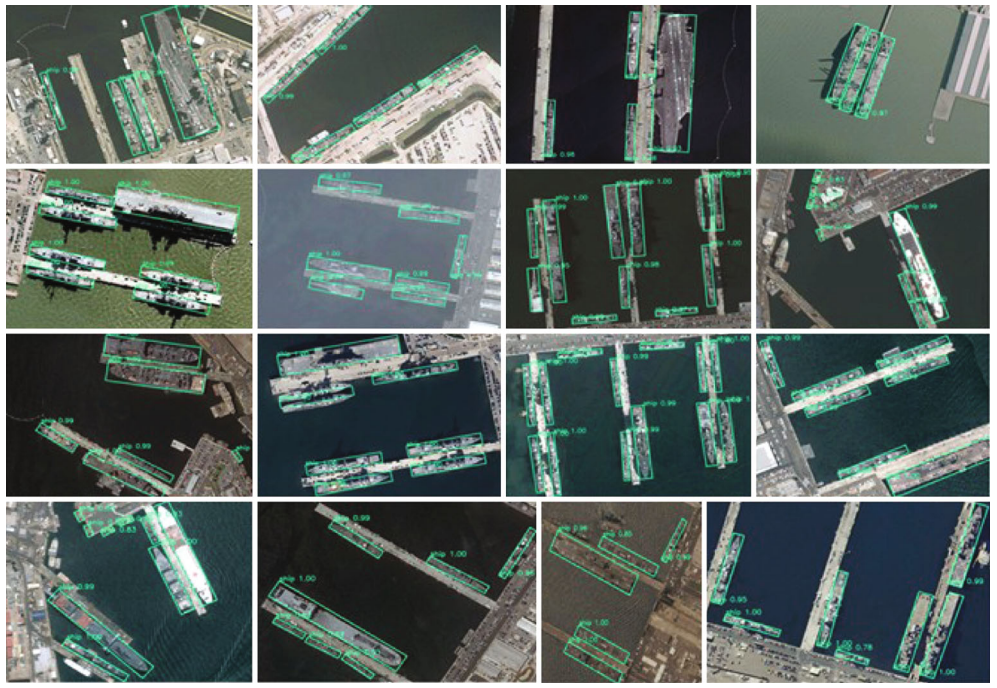


Figure 10: Visual results of rotated bounding boxes.

segmentation, and captioning dataset. It has 330k images, more than 200k labelled images, and 80 object categories, which is beneficial for object detection training.

The DOTA benchmark is the largest and most challenging dataset with oriented bounding box annotations for aerial image object detection. These images have been annotated by experts using 16 common object categories, and Table 1 shows that our approach has an excellent performance in terms of category balance and accuracy. The object categories include helicopter (HC), large vehicle (LV), small vehicle (SV), tennis court (TC), ground track field (GTF), basketball court (BC), soccer field (SBF), baseball diamond (BD), storage tank (ST), swimming pool (SP), and roundabout (RA).

In terms of data augmentation, images are flipped horizontally and vertically and rotated at random angles. For color, RGB channels are replaced randomly. For image color degradation, saturation in the HSV color space is multiplied by a random number in [0, 5].

We also conducted ablation experiments on COCO and DOTA by adopting different attention modules as shown in Table 2. Here, MRFNet is considered as the benchmark. PWC represents the pass-wise connection. RC represents the benchmark which adopts the residual connection techniques. EAWNet adopts the above strategies and modules to enhance the percentage of average precision (AP) and the inference speed of frames per second (FPS).

4.3. Experiment Analysis. We adopted different feature extraction methods and network structures as presented in Table 3. CAW represents the contextual attention-wise

modules and PRW represents the position refinement-wise modules. EAWNet adopted all the above approaches. Tables 1 and 4 illustrate the details of Figure 9 as well as the details of the experiments conducted. The experiments on COCO and DOTA validate the visual effect in dense and real-time object detection; the average precision is better than all other approaches in the comparative experiments. In addition, the results of the ablation experiments also shed light on different aspects of the connection strategies and enhancements on applying attention modules. The results further highlight the necessity of streamlining and optimizing the network structure and the effectiveness of using the attention mechanism to improve the efficiency of visual perception.

EAWNet outperforms YOLOv4 in terms of accuracy as shown in Table 4. With the same training process and dataset, we simply use the advanced MRFNet backbone as the benchmark and then add an attention-wise module.

The speed is slightly higher than YOLOv4 except for the additional module. Considering the significant accuracy improvement and fast training convergence speed, it is worthwhile to modify the model's name into an attention-wise multipath refinement flow counterpart. The average precision is shown in Figure 10. Tighter and specific detection bounding boxes benefit the training process.

As for FPS, the speed is higher than in comparative experiments. Compared to the LRF, YOLOv3, and YOLOv4, our method is slower but shows a significant improvement in terms of accuracy. Our method outperforms RFBNet, CenterMask, and EfficientDet with regard to both speed and accuracy. Therefore, our method presents a trade-off between accuracy and cost compared with YOLOv4 and outperforms most of the recent state-of-the-art methods.

5. Conclusions

Object detection has been widely used in the field of VIOT. Therefore, it is an important issue for reconstructing a smart city. However, very large images, complex image backgrounds, uneven size, and quantity distribution of training samples make detection tasks challenging, especially for small and dense objects. To solve these problems, an object detector Edge Attention-wise Convolutional Neural Network (EAWNet) is proposed in this paper. Firstly, a better training method with multiflow fusion network is designed to improve the detection accuracy. Secondly, self-attention modules are adopted to underline the meaningful information of feature maps while disregarding useless information. Finally, pass-wise connection makes key semantic features propagate effectively. Comparative experiments are conducted on the benchmark dataset COCO with state-of-the-art methods. The results indicate that our proposed object detection methods outperform the existing models. Extensive experiments and comprehensive evaluations on large-scale DOTA and daily COCO datasets demonstrate the effectiveness of the proposed framework on real-time and dense object detection inference.

In this work, we proposed a framework called EAWNet with edge attention-wise modules for real-time visual internet of things. The patches flow in the multipath refinement flow network, and features are extracted by a pass-wise connection that contributes to a considerable training efficiency. The model was evaluated on two public datasets and compared to state-of-the-art approaches. It performed quite satisfactorily in terms of both accuracy and speed under the same conditions.

In the future, we will redesign the attention modules for lower computation cost. Then, continuous improvements on object detection detectors could be conducted by applying different data augmentation skills and various feature extraction methods and network enhancement approaches as well. We are also interested in establishing whether rotated boundaries could be replaced by the instance segmentation to achieve better results on specific tasks such as an adversarial training process.

Data Availability

The DOTA dataset used to support the findings of this study have been deposited in the DOTA repository (https://captain-whu.github.io/DOTA/dataset.html). The COCO dataset used to support the findings of this study have been deposited in the COCO repository (https://cocodataset.org/).

Conflicts of Interest

There are no conflicts of interest with any affiliation and person.

Acknowledgments

We thank the National University of Defense Technology for providing the GPU clusters. This research is supported by the College of Advanced Interdisciplinary Studies. This work is supported by the National Natural Science Foundation of China (grant number: 62001493), It as also funded by the Postgraduate Scientific Research Innovation Project of Hunan Province (grant number: CX20200043).

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE conference on computer vision and pattern recognition*, pp. 580–587, Columbus, OH, 2014.

[2] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: object detection via region-based fully convolutional networks," *NIPS*, 2016.

[3] R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision (ICCV)*, Las CondesAraucano Park, Chile, 2015.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *ICCV*, 2017.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *2016 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 779–788, Las Vegas, NV, 2016.

[6] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 6517–6525, Honolulu, HI, 2017.

[7] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Cham, 2016.

[8] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: deconvolutional single shot detector," 2017, https://arxiv.org/abs/1701.06659.

[9] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, Vienice, Italy, 2017.

[10] Z. Cai and N. Vasconcelos, "Cascader-cnn: delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162, Salt Lake City, Utah, U.S., 2018.

[11] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Le Cun, "Overfeat: integrated recognition, localization and detection using convolutional networks," 2014, https://arxiv.org/abs/1312.6229.

[12] J. Huang, V. Rathod, C. Sun et al., "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 2, Honolulu, Hawaii, USA, 2017.

[13] H. Law and J. Deng, "Cornernet: detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, vol. 1, pp. 734–750, Munich, Germany, 2018.

[14] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Sod-mtgan: small object detection via multi-task generative adversarial network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 2, Munich, Germany, 2018.

[15] T. Wang, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Learning rich features at high-speed for single-shot object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1971–1980, Seoul, Korea, 2019.

[16] S. Zheng, Z. Zhu, J. Cheng, Y. Guo, and Y. Zhao, "Edge heuristic GAN for non-uniform blind deblurring," *IEEE Signal Processing Letters*, vol. 26, no. 10, pp. 1546–1550, 2019.

[17] Y. Yang and H. Deng, "Gc-yolov3: you only look once with global context block," *Electronics*, vol. 9, no. 8, pp. 1–14, 2020.

[18] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, U.S, 2018.

[19] A. Bochkovskiy, C. Wang, and H. Mark Liao, "YOLOv4: optimal speed and accuracy of object detection," in *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.

[20] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11215 LNCS*, pp. 404–419, 2018.

[21] Y. Lee and J. Park, "CenterMask: real-time anchor-free instance segmentation," in *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 13903–13912, Seattle, WA, USA, 2020.

[22] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: scalable and efficient object detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 10778–10787, Seattle, WA, USA, 2020.

[23] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," *arXiv preprint arXiv:1406.6247*, 2014.

[24] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 842–850, Boston, MA, USA, 2015.

[25] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 6000–6010, 2017.

[26] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, http://arxiv.org/abs/1805.08318.

[27] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3588–3597, Salt Lake City, Utah, U.S., 2018.

[28] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, Salt Lake City, Utah, U.S., 2018.

[29] J. Fu, J. Liu, H. Tian et al., "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3146–3154, Long Beach, California, USA, 2019.

[30] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: fully convolutional one-stage object detection," in *2019 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9626–9635, Seoul, Korea (South), 2019.

[31] J. Fu, J. Liu, J. Jiang, Y. Li, Y. Bao, and H. Lu, "Scene segmentation with dual relation-aware attention network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2146–2154, 2019.

[32] M. Tohidian, S. A. R. Ahmadi-Mehr, and R. B. Staszewski, "A tiny quadrature oscillator using low-Q series LC tanks," *IEEE Microwave and Wireless Components Letters*, vol. 25, no. 8, pp. 520–522, 2015.

[33] S. Yang, Z. Pei, F. Zhou, and G. Wang, "Rotated faster R-CNN for oriented object detection," in *Proceedings of the 2020 3rd International Conference on Robot Systems and Applications*, pp. 35–39, Chengdu, China, 2020.

[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *in IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[35] P. Sun, Y. Zheng, Z. Zhou, W. Xu, and Q. Ren, "R^4 Det: refined single-stage detector with feature recursion and refinement for rotating object detection in aerial images," *Image and Vision Computing*, vol. 103, p. 104036, 2020.

[36] G. Lin, A. Milan, and C. Shen, "RefineNet: multi-path refinement networks for high-resolution semantic segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, 2017.

[37] M. Zhen, J. Wang, L. Zhou et al., "Joint semantic segmentation and boundary detection using iterative pyramid contexts," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.

[38] E. Yang, C. Deng, C. Li, W. Liu, J. Li, and D. Tao, "Shared predictive cross-modal deep quantization," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5292–5303, 2018.

[39] C. Deng, E. Yang, T. Liu, and D. Tao, "Two-stream deep hashing with class-specific centers for supervised image search," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 6, pp. 2189–2201, 2019.

[40] H. Larochelle and G. E. Hinton, "Learning to combine foveal glimpses with a third-order Boltzmann machine," *Advances in neural information processing systems*, vol. 23, pp. 1243–1251, 2010.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," 2017, https://arxiv.org/abs/1709.01507.

[42] L. Chen, H. Zhang, J. Xiao et al., "Sca-cnn: spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, Hawaii, USA, 2017.

[43] F. Zhu, J. Yang, C. Gao, S. Xu, N. Ye, and T. Yin, "A weighted one-class support vector machine," *Neurocomputing*, vol. 189, pp. 1–10, 2016.

[44] T. Y. Lin, M. Maire, S. Belongie et al., "Microsoft COCO: common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755, Cham, 2014.

[45] G. S. Xia, X. Bai, J. Ding et al., "DOTA: a large-scale dataset for object detection in aerial images," in *2018 IEEE/CVF 8 conference on computer vision and pattern recognition*, pp. 3974–3983, Salt Lake City, UT, 2018.