Improving Surveillance Object Detection with Adaptive Omni-Attention over both Inter-Frame and Intra-Frame Context

Tingting Yu[1], Chen Chen[2], Yichao Zhou[1], and Xiyuan Hu[1] (✉)

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
{yutt,yczhou,huxy}@njust.edu.cn
[2] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract. Surveillance object detection is a challenging and practical sub-branch of object detection. Factors such as lighting variations, smaller objects, and motion blur in video frames affect detection results, but on the other hand, the temporal information and stable background of a surveillance video are major advantages that does not exist in generic object detection. In this paper, we propose an adaptive omni-attention model for surveillance object detection, which effectively and efficiently integrates inter-frame contextual information to improve the detection of low-quality frames and intra-frame attention to suppress false positive detections in the background regions. In addition, the training of the proposed network can converge quickly with less epochs because during multi-frame fusion stage, the pre-trained weights of the single-frame network can be used to update simultaneously in reverse in both single-frame and multi-frame feature maps. The experimental results on the UA-DETRAC and the UAVDT datasets have demonstrated the promising performance of our proposed detector in both accuracy and speed.(Code is available at https://github.com/Yubzsz/Omni-Attention-VOD.)

1 Introduction

Surveillance object detection, as an important sub-branch of generic object detection, aims at localizing objects with tight bounding boxes in each frame of a surveillance video. It has been studied for many decades from traditional background modelling and subtraction algorithms to the recent deep learning-based models. Although deep learning-based object detection models [1–10] have made significant progress on both images and videos, detecting objects in the surveillance scenario still has its own set of challenges and difficulties. First, a more accurate object detection result is needed for subsequent tracking or re-identification tasks. For example, in many benchmarks of video object detection, such as [11], [12], a threshold of 0.7 is set for calculating the average accuracy rather than a commonly used threshold of 0.5 for most image object detection. Then, the cases of occlusion, of smaller objects, and of motion blur will appear

more frequently in the surveillance videos than general images. Finally, considering the practical application of surveillance scenario, the inference speed of the detection model is another important issue, which puts high demands on the efficiency of surveillance object detection.

However, compared to the generic image object detection, surveillance object detection does have some additional prior information, such as relative stable background in the same video sequence and same object exists in consecutive frames, which can be used to improve the performance of object detection. For instance, for utilizing the intra-frame information, structural information or the layout of scenes have been used to refine the object detection results [13, 14]; and for utilizing the inter-frame information, some background modeling and optical flow-based approaches have been proposed to improve the detection of small objects [15–17]. Considering these reasons, we propose an adaptive omni-attention model for improving surveillance object detection, which effectively and efficiently combine the temporal information of the consecutive frames as inter-frame attention and the spatial context of surveillance scenarios as intra-frame attention. The experimental results on the UA-DETRAC [11] dataset have demonstrated the efficacy of our proposed detector when compared with other state-of-the-art surveillance object detection models.

The main contributions of our approach are summarized as follows.

- An inter-frame attention module, with temporal adaptive convolutions, has been added into the backbone feature extraction phase to effectively incorporate the feature maps of consecutive frames to enhance the small object detection.
- An intra-frame attention module that weights the current frame feature maps in channel and spatial dimensions has been introduced to suppress the false positive detections in the background regions.
- An efficient feature fusion module has been proposed to integrate inter-frame and intra-frame feature maps at different scales which can achieve a high detection accuracy with less training cost.

2 Related Work

2.1 Generic Object Detection

Currently deep learning-based object detection networks can be classified into two-stage and one-stage detection methods according to the detection process. The classical two-stage methods, such as R-CNN [1], SPP-Net [2], Faster RCNN [3], etc, have independent region proposal module, which can support a relative high detection accuracy, but will affect the detection speed. The one-stage methods discard the process of region proposal and directly regress the detection boxes on the images, which gains advantages in speed and gradually approaches the two-stage methods in terms of accuracy. Typical representatives of one-stage methods include YOLO [4] and its variants, SSD [5], RetinaNet [6], etc. The

recent rise of anchor-free object detection is an important improvement to one-stage detection models.

Compared with anchor-based method, anchor-free-based method refers to the detection by only dense prediction or keypoint estimation without manually designing anchors with different scale on the image. For example, CornerNet [7] and CenterNet [8], treat the object detection problem as key point detection, and respectively predict the corner and center points of the object to complete the detection. The selection of these points are determined by heatmap, and the loss function of heatmap is focal loss that proposed in [6]. What's more, FCOS [9] performs pixel-by-pixel prediction, RepPoints [10] learns a set of points to represent the object. Anchor-free-based methods solve the defects brought by the anchor, such as imbalance of positive and negative samples during training, high memory occupation, and difficulty in recognizing multi-scale objects.

Attention mechanism is another widely used technique for improving object detection, which has been developed rapidly as a training module that can significantly improve the accuracy of a model with only a small increase in model complexity or computational effort. For example, YOLOv4 [18] explores the impact of Squeeze-and-Excitation (SE) [19] and Spatial Attention Mechanisms (SAM) [20] methods for training the model, where SE is a channel attention mechanism and SAM is a spatial attention mechanism. EfficientDet [21] in backbone incorporates an SE attention module in each stage; TPH-YOLOv5 [22] integrates a convolutional attention module [23] (CBAM) to find attention regions in scenes with dense objects; in addition, attention mechanisms can also be used for feature aggregation [16].

2.2 Surveillance Object Detection

Compared with generic object detection, the biggest difference of surveillance object detection is the rich temporal and contextual information that can be utilized to improve the object detection. Existing surveillance object detection methods adopt this prior information in different ways, such as multi-frame feature aggregation, 3D convolution, dynamic foreground extraction, and so on.

Most multi-frame feature aggregation methods compute each frame detection intensively and then perform weighted average of features. FGFA [17] uses optical flow estimation to aggregate feature maps of neighboring frames. To reduce computational cost, THP [24] uses optical flow and sparse recursion to operate on sparse key frames. Because of the high computational complexity of estimating the optical flow, some non-optical flow-based feature aggregation methods have been proposed. For example, MEGA [25] integrates the precomputed features of the previous frame and stores them as global information in the remote memory module; FFAVOD [26] fuses feature matrices of the front and back frame of the current frame extracted by the backbone, and the output of the fusion module is used as the input of the current frame detection header.

3D-DETNet [27] uses 3D convolution to capture motion or temporal information encoded in multiple consecutive video frames, but the large number of 3D convolution parameters and low computational efficiency are rarely applied

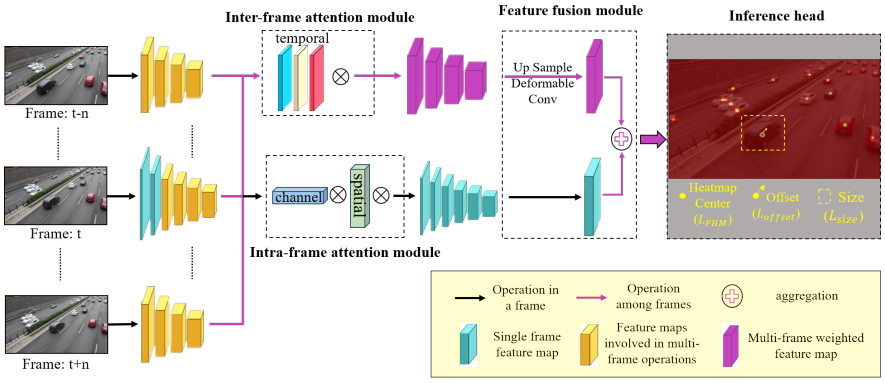

Fig. 1. Overview of our framework. The figure illustrates the detection process of 2n+1 frames video sequence. Figures 2,3 show the details of our inter-frame attention, intra-frame attention, and feature fusion modules, respectively.

directly in video detection. In the field of video understanding and object tracking, most of the temporal modeling methods usually decompose 3D convolution into the combination of 2D spatial convolution and 1D temporal convolution, such as P3D [28] and R(2+1)D [29]. In order to reduce the computation complexity of 3D convolution while giving 2D convolution the capability of temporal modeling, Huang et al. proposed Tada Conv [30], which performs temporal modeling by relaxing the temporal invariance of 2D convolution and superimposing adaptive temporal weights on it. TCTrack [31] is an example of using Tada Conv to improve the object tracking.Dynamic foreground extraction is another kind of approaches to utilize inter-frame information by updating a spatio-temporal background, which can effectively highlight moving objects, but has barely improvement for stationary objects or video jitter.

3 Proposed Method

Based on the consideration of effective and efficient use of attention in temporal, spatial and channel domain, we propose a multi-frame based omni-attention object detection network. Detailed illustration of the proposed network is provided in the following subsections.

3.1 Omni-attention based surveillance object detection architecture

The overall structure of the proposed omni-attention based surveillance object detection framework is illustrated in Fig.1, which consists of three main submodules: the inter-frame attention module, the intra-frame attention module and the feature fusion module. Since detection on each frame independently usually ignores the connection between the contextual frames, the proposed framework

utilizes multi-frame sequences instead of single-frame inputs. The input $L-$frame ($L = 2n + 1$) images are denoted as $\{I_e\}_{e=t-n,\ldots,t-1,t,t+1,\ldots t+n}$, with the target frame is I_e. The target frame I_e goes through 6 different convolution layers in the initial stage to obtain 6 different scales of feature maps. The 6-layer feature maps of the target frames contain both high-level global information and detailed information in the lower level. These feature maps are fed into the intra-frame and inter-frame attention modules, where the intra-frame module performs training of all 6 layers, while the inter-frame module trains only the last 4 layers.

For intra-frame attention, we perform operations at the target frame to exploit the channel and spatial information of the feature map at different scales. For inter-frame attention, we concatenate the last four layers of feature maps of the target frame and the involved contextual frames, assigning adaptive temporal weights to them. The reasons for selecting the last four layers of feature maps to apply temporal convolution here are: (1) the network structure of the first 2 layers of convolution is relatively simple (with only one convolution and normalization operation) and extracts low-level details; (2) the higher-level feature maps have a larger receptive field, which is more suitable for applying to multi-frame temporal contexts. Therefore, the temporal, channel and spatial information is aggregated at the same time, and then, the obtained fused feature maps are fed into 3 head branches: fusion heatmap, offset and size branches. The fusion heatmap (optimized by the loss L_{FHM}) represents the center point heatmap of each category object; offset (optimized by the loss L_{offset}) is the offset of the key point relative to the original image generated due to the image undergoing down-sampling, feature enhancement and other operations; and size (optimized by the loss L_{size}) is the distance between the object center point and the border. These three branches provide the final information to form the bounding box of the detection object.

3.2 Inter-frame attention module

Given the input image sequence, the multi-frame tandem feature map can be obtained at a certain stage of the backbone. Inspired by the temporal adaptive convolution in [30], for features of target frame enhance by the Tada Convolution with temporal modeling capability, we can obtain adaptive temporal weights specifically assigned to each frame.The input X represents the integration of all feature maps from multiple frames, and X_l represents the l-th of them. The output feature map $\tilde{X}_l$ corresponding to X_l obtained by passing through the Tada convolution is shown below:

$$\tilde{X}_l = W_l * X_l = (\alpha_l \cdot W_b) * X_l \tag{1}$$

where the $*$ indicates the convolution operation and $\cdot$ indicates the element-wise multiplication. For which the weight $W_l = \alpha_l \cdot W_b$, where W_b is the base weight shared by all frames in the network, and α_l is the calibration weight obtained from the temporal context different for each frame. Over the entire

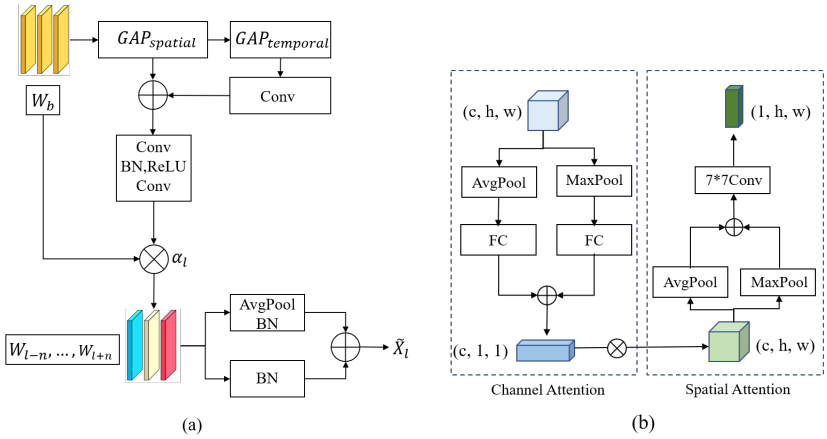

Fig. 2. (a) The inter-frame temporal attention architecture of the proposed method. GAP stands for global average pooling. (b) The channel and spatial attention architecture of the proposed method. FC stands for fully connected layer and the feature map dimensions are in the parentheses.

tandem feature map, feature map of each frame X_l is adaptively assigned with a specific calibration weight. The calibration weight α_l of the $l-$th frame not only considers the current frame, but also integrates the temporal attention among the related frame sequence. We show the process of generating the calibration weight α_l in Fig. 2(a).

Firstly, X_l is input to the Tada Convolution through a tandem operation. For obtaining the local temporal context information, a global average pooling (GAP) in the spatial dimension is first performed, such that $v_l = GAP_s(X_l)$, which is used as the frame descriptor. To obtain the global temporal context information, we perform a linear mapping of the global descriptors, superimposed on the frame descriptors, to further merge the global temporal information, i.e., $g = GAP_{st}(X)$. Here GAP_{st} represents the global average pooling in the spatial and temporal dimension. After aggregating global and local information, two convolutions to the frame descriptors as well as ReLU and normalization operations have been applied to obtain additional time-adaptive calibration weights. The whole process of spatial-temporal convolution can be described by the following equation:

$$\mathcal{F}_\alpha\left(\tilde{X}_l\right) = Conv\left(\rho\left(Conv\left(v_l + FC(g)\right)\right)\right) \tag{2}$$

where v_l and g stand for frame descriptor and global descriptor respectively, ρ stands for ReLU and batch normalization, FC stands for linear mapping.

The final weight W_l is the product of the calibration weight α_l and the base weight W_b, where the base weight can be replaced by the pre-trained weight of the single-frame network. The calibration weight is set to 0 during initialization,

which has the advantage of the training time reduction and flexibility of the convolution module embedded into the existing network.

In order to better aggregate spatial-temporal information and compensate the problem of inadequate spatial information extracted by spatial-temporal convolution, we add two branches after the adaptive convolution module as:

$$y_l = \psi_\rho \tilde{X}_l + \psi_v AvgPool\left(\tilde{X}_l\right). \tag{3}$$

Specifically, the output of the adaptive convolution as $\tilde{X}_l$ is fed into these two branches, one of which is average pooling. After performing different normalization operations on both sides, the outputs of the two branches are aggregated. In this way, according to the idea of SmallBigNet [32], the pooling branch provides a larger receptive field than the other branch, which can better integrate core and contextual semantics.

3.3 Intra-frame attention module

Besides temporal attention, attention in the spatial and channel dimensions also provides possible enhancement for feature maps of single-frame images. For a particular layer of feature maps in convolutional neural networks, the attention mechanism learns an additional weight of corresponding pixels in a particular dimension, and these weights represent the importance of a certain information that strengthens the useful features and weakens the useless ones, thus facilitate the feature screening and enhancement. Since surveillance videos often have stable background, we can use spatial and channel attention to suppress the false positive detection in the background region.

CBAM [23] is a widely used hybrid attention mechanisms combining both spatial and channel dimensions. Inspired by CBAM [23], the proposed intra-frame channel and spatial attention mechanisms are shown in Fig. 2(b), where the attention information on channels and special allocation is weighted sequentially on the 6-layer feature map generated in backbone. For channel attention, both average-pooling and max-pooling are used to compress the information on the spatial dimension, and then the pooled features are fed into a multilayer perceptron network with shared weights. For spatial attention, the feature map weighted by channel attention performs average-pooling and max-pooling on each channel and concatenates them to generate valid feature descriptors. Then a 7×7 convolutional layer is applied to obtain the weighted feature map of spatial dimension. To sum up, the proposed intra-frame attention mechanism can be described as:

$$F_c = MLP\left(AvgPool\left(X_l\right) + MLP\left(MaxPool\left(X_l\right)\right)\right. \tag{4}$$

$$F_S = f^{7\times7}\left[AvgPool\left(F_c * X_l\right); MaxPool\left(F_c * X_l\right)\right] \tag{5}$$

where F_c and F_s denote attention in the channel and spatial direction and $f^{7\times7}$ denotes the convolution operation with the filter size of 7×7.

3.4 Feature fusion module

In order to aggregate multi-frame temporal information and single-frame channel and spatial attention while applying pre-trained weights from single-frame network training, we propose an additional feature fusion module.

As shown in Fig.1, in the training phase, consecutive video sequential images are fed into the network, and each image is convolved through a 6-layer network to obtain 6 feature maps with different scales. In the intra-frame attention module, channel and spatial attention are applied to each feature map of the target frame. In the inter-frame attention module, the feature maps of the last four layers of target frame and its preceding and following frames are concatenated and then fed into the temporal adaptive convolution layer to obtain the temporal attention-weighted feature map of the target frame.

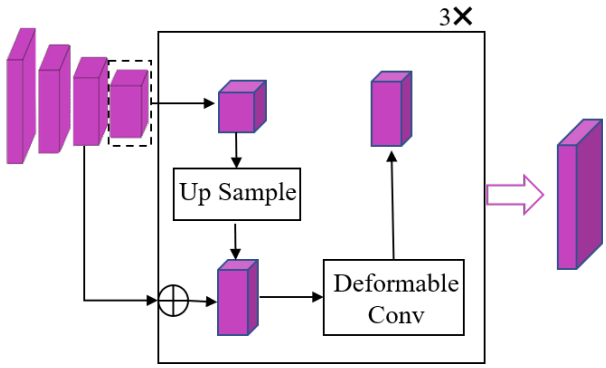

Fig. 3. Feature maps aggregation architecture after temporal adaptive convolution and intra-frame attention. 3× represents performing the fusion step for 3 times.

The four feature maps obtained after temporal weighting are shown as inputs of the feature fusion module in Fig.3. In order to combine the semantic information of the higher-level feature map and the spatial information of the lower-level feature map, the proposed feature fusion module aggregates the four feature maps. First,the feature map of the highest dimension is selected, upsampled to the dimension of the second layer feature map, and then, the upsampled feature map and the second layer feature map are pixel-wise summed and sent to the deformable convolution for better adaption of different object shapes, sizes and other geometric deformations according to the feature maps. The feature map obtained completing the above operation is used as the updated feature map of the highest dimension, and then the same operation is performed. Since the input has a total of four layers of feature maps, the above-mentioned upsampling and convolution process is performed 3 times to get the final output.

Through the above process we obtain the channel and spatial attention feature map on a single frame and the temporal attention feature map among

multiple frames, and then, we add them up to get the merged feature map. Specifically, we perform the computation of the fusion heatmap on the merged feature map, while size and offset of the bounding boxes are generated from the single-frame feature map, since the feature overlay of multi-frame has great influence on the width and height values of the bounding box.

The total loss function consists of a loss of fusion heatmap trained with focal loss, and two losses of offset and size regression trained with L1 loss. The loss function for each head block is as follows:

$$L_{FHM} = -\frac{1}{N}\sum_{xy}\begin{cases}(1-\hat{Y}_{xy})^{\alpha}\log(\hat{Y}_{xy}) & if\ Y_{xy}=1\\ (1-Y_{xy})^{\beta}(\hat{Y}_{xy})^{\alpha}\log(1-\hat{Y}_{xy}) & otherwise\end{cases} \tag{6}$$

L_{FHM} is the loss of fusion heatmap, where $\hat{Y}_{xy}$ represents the predict heatmap value for each pixel, and $\hat{Y}_{xy}=1$ denotes the pixel is the center point of an object, α and β are hyper-parameters of the focal loss. Offset and size are calculated by the loss only on the predicted object centroid.

$$L_{\text{offset}} = \frac{1}{N}\sum_{p}\left|\hat{O}_{\tilde{p}} - \left(\frac{p}{R}-\tilde{p}\right)\right| \tag{7}$$

L_{offset} is the loss of heatmap offset to the center point, where $\hat{O}_{\tilde{p}}$ represents the predict offset, $\frac{p}{R}$ is the position after downsampling the original image, and $\tilde{p}$ is the true coordinate of center point.

$$L_{size} = \sum_{k=1}^{N}\left|\hat{S}_k - S_k\right| \tag{8}$$

L_{size} is the loss of the size of bounding box, where $\hat{S}_k$ represents the predicted size, while S_k is the size of ground truth bounding box. The overall training objective is:

$$L_{inter,intra} = L_{FHM} + \lambda_{\text{size}}\, L_{\text{size}} + \lambda_{off} L_{offset} \tag{9}$$

The hyper-parameters are set as $\lambda_{\text{size}} = 0.1$ and $\lambda_{off} = 1$ in experiments , which represent the weights of the loss for regression size and offset of the bounding box, respectively.

4 Experiments

4.1 Experimental details

Datasets. As the proposed method utilizes the omni-attention including temporal information to improve surveillance object detection, the UA-DETRAC [11] dataset, a widely used real-world video object detection benchmark, is adopted to evaluate the performance. The dataset has over 140,000 frames and 1.21 million

labeled object bounding boxes, with vehicle objects covering cars, buses, trucks, etc.We also perform experimental validation on the UAVDT [12] dataset.

For training, we sample one frame out of every 10 frames for the intra-frame context extraction, and the additional frames involved in inter-frame temporal information are selected from the intervening frames.

Models. In our experiments, the proposed network architecture is built based on the CenterNet [8] implementation, with the DLA-34 selected for the backbone network and all detectors optimized with Adam. The DLA-34 is selected as backbone because it achieves reasonable balance between efficiency and accuracy among the three pre-trained backbone networks including Hourglass, DLA-34 and ResNet-101. Table 1 shows the accuracy and efficiency of the three backbone networks in [8] for object detection on COCO validation.

Table 1. Object detection result comparison on COCO validation using different backbones, according to [8], flip and multi-scale represent the use of different data enhancements.

Backbone	AP/FPS	Flip AP/FPS	Multi-scale AP/FPS
Hourglass	40.3/14	42.2/7.8	45.1/1.4
DLA-34	37.4/52	39.2/28	41.7/4
ResNet-101	34.6/45	36.2/25	39.3/4

Training details. The initialization of weights is set by the pre-training on COCO dataset, following the settings in [8]. The number of epochs is set to 50 and the learning rate is set to 2e-5, decreasing by a factor of 10 at the 30th and 40th epochs sequentially. The data augmentation operations including random scaling, random cropping, and flipping are used during training. Since the backbone network is first trained on UA-DETRAC dataset beforehand, the proposed method starts with decent feature representation capability and the losses can converge rapidly when training by the proposed method and the overall training time can be reduced.

4.2 Main results

We compare the PR curve of our method on UA-DETRAC dataset under different settings in comparison with other state-of-the-art object detection methods, and show them in figures 4 and 5, respectively. The AP scores are calculated based on the literature [33], calculating the average precisions at the fixed 11 recall values from 0 to 1: [0, 0.1, 0.2, . . . , 0.9, 1.0].

The different PR plots represent different settings in the dataset, which are shown in Fig.4 and Fig.5. The proposed method achieves a 3.33% improvement over the baseline in the overall setting. It shows that our method has provided significant average enhancement compared to the baseline in different difficulty and different lighting conditions, since the omni-attention strategy-based method

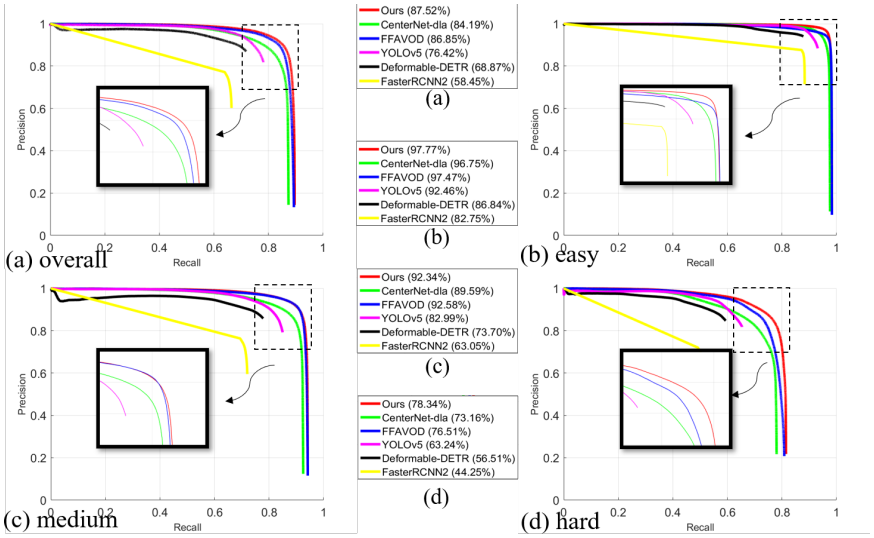

Fig. 4. PR curve comparison on UA-DETRAC dataset for object detection with the full test set and the breakdown by difficulty level: easy, medium, hard. Different colors represent different methods.

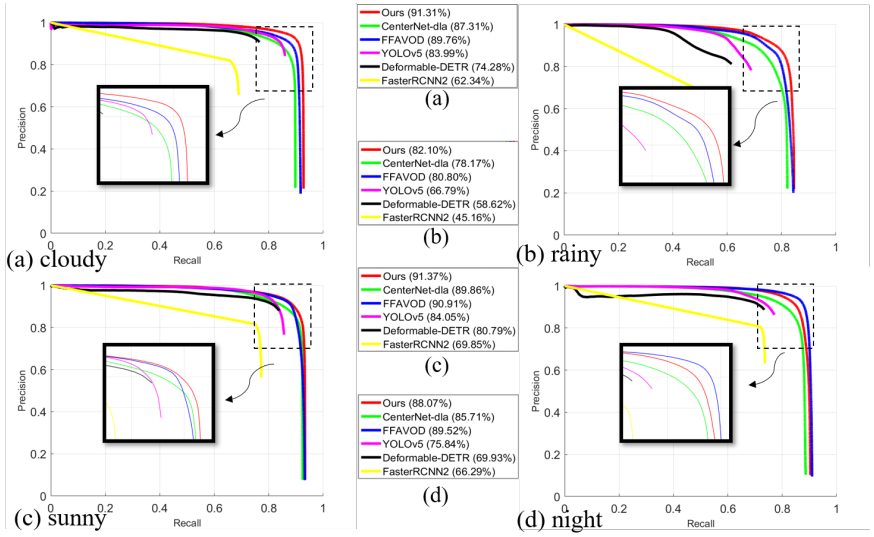

Fig. 5. PR curve comparison on UA-DETRAC dataset for object detection according to different light conditions: cloudy, rainy, sunny, night. Different colors represent different methods.

can accurately capture temporal and local contextual information to improve the effectiveness of vehicle detection per frame. The PR curve of the proposed method shows clear improvement in all settings compared with all methods.

We compare the proposed method with the state-of-the-art object detection methods and show the results in Table 2. The proposed method achieves quite competitive performance on all settings. It is worth noting that the proposed method outperforms 3D-DETnet, a method using temporal 3D convolution, with a large gap up to 34.22% in average precision. The proposed method surpasses all methods on all settings except for Medium and Night, and achieves the second-best performance on these two settings compared with FFAVOD with a slight decrease of 0.24% and 1.45% in precision. Considering the FFAVOD is a multi-frame feature fusion method using an hourglass backbone with deeper network layers, the proposed method with DLA-34 backbone verifies the effectiveness of itself. As the inference speed shown in Table 2, the proposed method achieves the optimal tradeoff between accuracy and efficiency.

The results on the UAVDT dataset are reported in Table 3.Our method achieves 2.87% gain compared with the sophisticated FFAVOD-SpotNet.

Table 2. Comparison with the state-of-the-art methods on UA-DETRAC dataset under different settings. The best result is shown in bold.

Method	Overall	Easy	Medium	Hard	Cloudy	Night	Rainy	Sunny	FPS	Environment
CenterNet(dla)	84.19	96.75	89.59	73.16	87.31	85.71	78.17	89.86	24	GPU@A30
FFAVOD(hourglass)	86.85	97.47	**92.58**	76.51	89.76	**89.52**	80.8	90.91	6	GPU@A30
FG-BR Net	79.96	93.49	83.6	70.78	87.36	78.42	70.5	89.8	10	GPU@M40
3D-DETnet	53.3	66.66	59.26	43.22	63.3	52.9	44.27	71.26	26	-
Illuminating	80.76	94.56	85.9	69.72	87.19	80.68	71.06	89.74	14	GPU@TitanX
FasterRCNN2	58.45	82.75	63.05	44.25	62.34	66.29	45.16	69.85	11.1	GPU@TitanX
Deformable DETR	68.87	86.84	73.7	56.51	74.28	69.93	58.62	80.79	2.5	GPU@A30
YOLOv5	76.42	92.46	82.99	63.24	83.99	75.84	66.79	84.05	-	-
ours	**87.52**	**97.77**	92.34	**78.34**	**91.31**	88.07	**82.1**	**91.37**	11	GPU@A30

Table 3. Comparison of the mAP of our method on the UAVDT dataset with other state-of-the-art methods.

Method	overall
FFAVOD	52.07%
FFAVOD-SpotNet	53.76%
CenterNet	51.18%
ours	56.63%

The proposed method utilizes the temporal information over multiple frames in surveillance video for better detection of the objects in motion in the mean while combines the intra-frame channel and spatial attention information for accurate feature representation. Fig. 6 shows the detection results of the proposed method compared with the baseline method. As we can see, the proposed

method shows superior performance applied to surveillance video object detection with the enhancement in missed detection and false detection, especially on vehicle objects of small size or far distance and in environment with low light conditions.

Fig. 6. Comparison of detection performance of baseline and our method, yellow arrows indicate the missed detection, red arrows indicate the false detection.

4.3 Ablation Study

Module contribution. The proposed detector adaptively utilizes omni-attention within and between frames by the designed components of both inter-frame attention module, intra-frame attention module, and we perform the ablation studies to evaluate these components. We add the inter-frame temporal attention as well as intra-frame channel and spatial attention sequentially on top of the baseline, and show the effect of each component in Table 4. We find that the method with inter-frame temporal attention or the method with intra-frame attention both detects better than the baseline method, and the proposed method with both two modules performs the best. According to the proposed omni-attention mechanism, the intra-frame channel and spatial attention suppresses the false positive detection in the background region and improves the accuracy of the detection of the target objects within the single frame, while the temporal context features compensate for the unclear object caused by insufficient illumination.
Temporal fusion configurations. We also investigate the effect of temporal fusion configurations on the detection results with different number of frames and the frame sampling schemes. We evaluate the setting the number of fused frames $n = 3$ or 5, and the fused frames are selected as consecutive frames or interval frames. The detection results for different number of frames and different interval

Table 4. Ablation study of the proposed method on UA-DETRAC dataset.

Baseline	✓	✓	✓	✓
+ Inter-frame Attention		✓		✓
+ Intra-frame Attention			✓	✓
AP(%)	84.19	87.05	86.58	87.52

Table 5. Temporal fusion configuration evaluation on the intra-frame module.

Frame sampling scheme	Numbers of frames			
	N=3		N=5	
Consecutive	✓		✓	
Interval		✓		✓
AP(%)	87.31	87.29	86.99	87.52

frames are shown in Table 5. The best results are obtained when $n = 5$ and the fused frames are set as discontinuous, that is, the $l-1, l-3, l+1, l+3$ frames of the current frame l are selected as the temporal context. The table shows that when the number of fused frames is set as 3, there is little difference between consecutive or interval frames, it might because the information on the adjacent frames in surveillance video is similar and the weighting is valid for the current frame. And when the number of fused frames is set as 5, the effect of temporal weighting on consecutive frames becomes worse, which might be caused by the redundant information in multiple consecutive frames. while the performance achieves the best when the sampling scheme is changed into the interval frames, because the model integrates more temporally global information and has greater improvement in the object detection of the current frame.

5 Conclusion

In this work, we have proposed an adaptive omni-attention model for surveillance object detection based on anchor-free object detector. Using the proposed method, we train and evaluate our network on the dataset in the field of traffic surveillance. Our experiments demonstrate that our method compares favorably against the widely-used multi-frame and single frame methods. Our method makes efficient in the 3D temporal dimension, which has positive significance for subsequent research of video object detection.To improve the detection speed of our proposed model is another issue that warrants further study.

Acknowledgements This work was supported by the National Natural Science Foundation of China (62172227) and National Key R&D Program of China (2021YFF0602101).

References

1. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
2. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence **37**(9), 1904–1916 (2015)
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
4. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
5. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
6. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
7. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV). pp. 734–750 (2018)
8. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
9. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019)
10. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: Reppoints: Point set representation for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9657–9666 (2019)
11. Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M.C., Qi, H., Lim, J., Yang, M.H., Lyu, S.: Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. Computer Vision and Image Understanding **193**, 102907 (2020)
12. Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., Tian, Q.: The unmanned aerial vehicle benchmark: Object detection and tracking. In: Proceedings of the European conference on computer vision (ECCV). pp. 370–386 (2018)
13. Loganathan, G.B., Fatah, T.H., Yasin, E.T., Hamadamen, N.I.: To develop multi-object detection and recognition using improved gp-frcnn method. In: 2022 8th International Conference on Smart Structures and Systems (ICSSS). pp. 1–7. IEEE (2022)
14. Wang, T., He, X., Cai, Y., Xiao, G.: Learning a layout transfer network for context aware object detection. IEEE Transactions on Intelligent Transportation Systems **21**(10), 4209–4224 (2019)
15. Fu, Z., Chen, Y., Yong, H., Jiang, R., Zhang, L., Hua, X.S.: Foreground gating and background refining network for surveillance object detection. IEEE Transactions on Image Processing **28**(12), 6077–6090 (2019)
16. Wang, X., Hu, X., Chen, C., Fan, Z., Peng, S.: Illuminating vehicles with motion priors for surveillance vehicle detection. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 2021–2025. IEEE (2020)

17. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 408–417 (2017)
18. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
19. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
20. Zhu, X., Cheng, D., Zhang, Z., Lin, S., Dai, J.: An empirical study of spatial attention mechanisms in deep networks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6688–6697 (2019)
21. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020)
22. Zhu, X., Lyu, S., Wang, X., Zhao, Q.: Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2778–2788 (2021)
23. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
24. Zhu, X., Dai, J., Yuan, L., Wei, Y.: Towards high performance video object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7210–7218 (2018)
25. Chen, Y., Cao, Y., Hu, H., Wang, L.: Memory enhanced global-local aggregation for video object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10337–10346 (2020)
26. Perreault, H., Bilodeau, G.A., Saunier, N., Héritier, M.: Ffavod: Feature fusion architecture for video object detection. Pattern Recognition Letters **151**, 294–301 (2021)
27. Li, S., Chen, F.: 3d-detnet: a single stage video-based vehicle detector. In: Third International Workshop on Pattern Recognition. vol. 10828, pp. 60–66. SPIE (2018)
28. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: proceedings of the IEEE International Conference on Computer Vision. pp. 5533–5541 (2017)
29. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)
30. Huang, Z., Zhang, S., Pan, L., Qing, Z., Tang, M., Liu, Z., Ang Jr, M.H.: Tada! temporally-adaptive convolutions for video understanding. arXiv preprint arXiv:2110.06178 (2021)
31. Cao, Z., Huang, Z., Pan, L., Zhang, S., Liu, Z., Fu, C.: Tctrack: Temporal contexts for aerial tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14798–14808 (2022)
32. Li, X., Wang, Y., Zhou, Z., Qiao, Y.: Smallbignet: Integrating core and contextual views for video classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1092–1101 (2020)
33. Everingham, M., Eslami, S., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)