Multiframe CenterNet Heatmap ROI Aggregation for Real-Time Video Object Detection

YOU ZHOU, YONG BAI (Senior Member, IEEE), AND YONGQING CHEN

State Key Laboratory of Marine Resource Utilization in South China Sea, College of Information and Communication Engineering, Hainan University, Haikou 570228, China

Corresponding author: Yong Bai (bai@hainanu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61961014, and in part by the Hainan Provincial Natural Science Foundation of China under Grant 620RC556.

ABSTRACT

Though Two-stage video object detectors cannot perform detection in real time, the accuracy of them is normally higher than that of one-stage video object detectors. One essence is that two-stage video detectors can easily use feature information from adjacent frames to augment key frame features. How to extract and exploit temporal features in the video stream for one-stage detectors needs further exploration. CenterNet is an anchor-free one-stage object detector that regresses bounding boxes from heatmap peaks. We propose to use detected peaks and regressed boxes which encompass peak points to determine the heatmap ROIs as the extracted object heatmap features. A new relation module is designed to evaluate the similarity of heatmap ROI features and output the relation features which can effectively augment the heatmap ROI features. In the video sequence the heatmap ROIs of multiple adjacent frames are aggregated to a heatmap ROI of the key frame. Compared to CenterNet and other CenterNet-based video object detectors, our method achieves improved online real-time performance on ImageNet VID dataset with 78.8% mAP at 36 FPS.

INDEX TERMS

Video object detection, heatmap ROI, real-time object detection, one-stage detector, anchor-free.

I. INTRODUCTION

With the rapid advancing of convolutional neural networks in recent years, a number of convolutional neural network-based models have been proposed to solve image detection tasks and achieve high accuracy on still images. Video detection is needed in many scenarios such as autonomous driving, video intelligent surveillance, and industrial inspection. The video detection needs to deal with fast moving objects that cause dynamic blur, obscured objects in certain frames, and objects with rare pose. In addition, video detection is generally expected to operate in real-time with more than 25 frames per second processing speed.

Previous studies in [4], [5] based on two-stage models use temporal information to augment image features for improving accuracy, but the two-stage object detection model structure is more complex compared to the one-stage model and cannot meet the demand of real-time detection. On the other hand, the one-stage network is difficult to extract the object information to remove the background information in the feature map because it has no RPN (Region Proposal Network) as in R-CNN, and therefore it is also difficult to use the temporal information to aggregate and enhance feature map. In this paper, we propose a multi-frame CenterNet-based temporal heatmap ROI (Region of Interest) aggregation module to improve the performance of video object detection. CenterNet is an anchor-free one-stage object detector. It selects the peak on the heatmap as the object center point and regresses the aspect of the object at the peak. Then it decodes the center point, size (width and height), and bias information into the original image to detect the object. In light of the effectiveness of RPN in two-stage object detectors, We propose the heatmap ROI, a kind of region of interest on heatmap, for extracting heatmap features in CenterNet. Each selected heatmap ROI corresponds to a possible object heatmap region. Then we design a multi-frame heatmap ROI aggregation module to extract heatmap ROIs in multiple adjacent video frames and aggregate them on the heatmap of key frame to enhance the heatmap feature of candidate objects in the key frame. The rest of paper is organized as follows. Section 2 introduces the related work of object detectors and video object detectors. Section 3 presents our proposed
method. Section 4 describes the details of our implementation and demonstrates our experimental results.

II. RELATED WORK

A. OBJECT DETECTION FROM IMAGES

The current high accuracy object detectors are basically designed based on Convolutional Neural Networks (CNNs) [11], [13], [18], [19]. According to whether or not the ROI feature extraction process exists, they are divided into one-stage object detectors and two-stage object detectors. Two-stage object detectors such as Fast(er) R-CNN [8], [17] and Cascade R-CNN [3] extract ROIs via RPN module in the first stage, and classify and further refine the localization of the object in the second stage. Direct classification and detection in the feature map is generally considered a one-stage object detector, such as the YOLO series [2], [16], RetinNet [15]. In general, two-stage object detectors are more accurate but relatively slow. Some work [14], [24] has shown that the accuracy of one-stage object detectors is comparable to that of two-stage.

The difference in the process of mapping from bounding box to feature maps is divided into two types of detectors: anchor-based detectors and anchor-free detectors. Anchor-based detectors require setting a priori anchors to regress the difference between the real bounding box and the object [8]. Anchor-based detector usually generates many anchors and then performs regression or filtering on these anchors, which causes increase of computation complexity as the drawback of the real-time performance of the detectors. In anchor-free detectors, CornerNet [14] detects the two vertices of the object bounding box, and CenterNet [24] detects the center point of the object bounding box. The coordinates and attributes of the object itself are used during the encoding of the label. It can outperform the anchor-based detectors in terms of detection speed.

B. VIDEO DETECTION

Unlike object detection for still images, objects in video detection tasks can change their appearance, aspect ratio, shape, and other attributes during motion. Other issues can also appear during motion, such as motion blur, obscured by other objects, etc. Therefore, it is challenging for video object detector to maintain the consistency of the object in the detection process with respect to the temporal order to ensure that the object can also be detected in multiple frames.

Video has more information in the time domain than still pictures, and methods [1], [4], [5], [7], [9], [12], [20], [22], [23], [25], [26] of using the information in the time domain to improve the detection of the current frame have been explored to resolve the difficulty in video detection. FGFA [25] and MANet [20] uses the optical flow predicted by FlowNet [6] to propagate feature information between frames. STSN [1], STMN [22] and other methods for direct aggregation of multi-frame features. RDN [5] based on Relation Network learns the relationship between multi-frame candidate boxes and uses self-attention to pass relation features from other frames to key frames. Most of the multi-frame feature aggregation methods are based on two-stage models, which are difficult to be used in real time detection. One-stage model does not have RPN structure, and can not extract object features on the feature map. The relation features of these model ROI features calculated by the fully connected layer and the relationship module cannot be used to augment the original feature map. In this paper, we design the relation module that can calculate the relation features of two ROI features on the heatmap. The most relevant work to ours is [23], which propagated a heatmap of the previous frame to enhance the detection results of upcoming frame. In our work, we propose the concept of heatmap ROI based on CenterNet to extract the potential object features on specific heatmap regions instead of the whole heatmap of a video frame, and further aggregate heatmap ROIs on multiple frames to boost the object detection results.

III. PROPOSED METHOD

A. BACKGROUND: CenterNet

CenterNet is an anchor-free one-stage object detector with a relatively simple and intuitive network structure. Given an input RGB image of width $W$ and height $H, I \in \mathbb{R}^{W \times H \times 3}$. CenterNet use fully-convolutional encoder-decoder backbone network to predict a downsampled heatmap $\hat{Y}$ from image $I$, $\hat{Y} \in [0, 1]^{\frac{W}{R} \times \frac{H}{R} \times C}$, where $R$ is the output stride and $C$ is the number of classes. A prediction $\hat{Y}_{x,y,c}$ corresponds to the detected center of an object of class $c$ at position $(x, y)$, while $\hat{Y}_{x,y,c} = 0$ corresponds to the background. Let an object be $k$ and its corresponding class be $c_k$, the bounding box of the object is $[x_k, y_k, w_k, h_k]$. Its center point is at $p_k$. Hence CenterNet is a heatmap-based detector since it uses heatmap $\hat{Y}$ to predict the center points of objects. CenterNet network flow chart is shown in Fig. 1. The input image is generated into a high-resolution feature map (heatmap) with a fully-convolutional encoder-decoder backbone network. The backbone network can be selected from alternatives such as ResNet with DCNv2 or hourglass network to extract feature map of an image. The extracted feature map can be refined by up-convolutional network and deep layer aggregation (DLA). Then the refined feature map is processed by 3 separate heads to output $\hat{Y}, \hat{O}$ and $\hat{S}$. Peak detection can then be performed on the heatmap $\hat{Y}$ to obtain the predicted coordinates of the object center point. For a peak at position $(\hat{x}_i, \hat{y}_i)$, $\hat{O}_{\hat{x}, \hat{y}} = (\delta \hat{x}_i, \delta \hat{y}_i)$ is its predicted offset; $\hat{S}_{\hat{x}, \hat{y}} = (\delta \hat{w}_i, \delta \hat{h}_i)$ is the predicted width and height of object bounding box, which is represented as

$$\hat{x}_i + \delta \hat{x}_i - \hat{w}_i/2, \hat{y}_i + \delta \hat{y}_i - \hat{h}_i/2$$

$$\hat{x}_i + \delta \hat{x}_i + \hat{w}_i/2, \hat{y}_i + \delta \hat{y}_i + \hat{h}_i/2.$$ (1)

B. HEATMAP ROI EXTRACTION

CenterNet has good results as a one-stage object detector for still images. When detecting video, problems such as obscured object center point, motion blur, and rare pose may...
be encountered. Feature aggregation can augment the feature map information of key frame by extracting the feature map of nearby frames to improve the detection accuracy of the key frame. Feature aggregation is generally used in two-stage object detectors to extract object feature and remove background information by RPN. Nevertheless, CenterNet does not have the RPN structure. In light of the effectiveness of RPN, we propose to extract object feature information by heatmap ROI. Each selected heatmap ROI corresponds to a possible object heatmap region. Then, multiple heatmap ROIs of adjacent frames are aggregated to the key frame to enhance the heatmap features of candidate objects and hence improve the detection accuracy.

Video object detection requires detection of all video frames \(\{N_k\}_{k=1}^T\). We set key frame for detection to \(N_k\). Since we use dataset of VID (30 categories), The heatmap of \(N_k\) is \(H_k\), the tensor of \(H_k\) is \((128, 128, 30)\) representing 30 heatmaps with width 128 and height 128 corresponding to the 30 categories. Peak detection is performed on the \(H_k\), and 10 peak points in each heatmap are selected using \(7 \times 7\) Maxpooling. We set the peak points of each heatmap to the set \(\{P_i\}_{i=1}^I\), \(I = 10\). The five categories with the highest peaks are obtained by summing and sorting the 50 sets \(\{P_i\}_{i=1}^I\) and recording the category index values. Such peak detection is performed for all \(N_k\), and the five object categories with highest number of occurrences are used for heatmap ROI aggregation. This design can improve the model running speed while ensuring the enhanced accuracy of multi-object detection. With the position information of \(\{P_i\}_{i=1}^I\), the width and height \((\hat{w}_i, \hat{h}_i) = \hat{S}_{\hat{x}_i, \hat{y}_i}\) can be obtained, and the bounding box of heatmap ROI is calculated as

\[
(\hat{x}_i - \frac{\hat{w}_i}{2}, \hat{y}_i - \frac{\hat{h}_i}{2}, \hat{x}_i + \frac{\hat{w}_i}{2}, \hat{y}_i + \frac{\hat{h}_i}{2}).
\]

(2)

As shown in Fig. 2, a heatmap ROI is determined by the bounding box obtained by (2). To reduce the computation complexity in the following steps, we use NMS (Non-Maximum Suppression) to select salient heatmap ROIs. Each selected heatmap ROI corresponds to a potential object heatmap region.

**C. RELATION MODULE**

In two-stage network models such as [4], [5], the ROI features go through the fully connected layer and the relation module is used to obtain the relation features. Then the relation features are aggregated on ROI features to obtain the augmented features, and finally the augmented features is used to regress the bounding boxes. On the other hand, one-stage network models regress the bounding boxes directly on the feature map. The relation features are obtained as one-dimensional tensors after the fully connected layer and the relation module, so it can not to be used directly to be augmented with the original two-dimensional ROI features. To cope with this issue, We design a new relation module to calculate the relation features of two heatmap ROIs.

We denote heatmap ROI generated by key frame by \(M_t = \{m_i\}\). Considering that the feature similarity between frames that are farther away on the time series decreases, we form a group \(\{N_j\}_{j=t-\tau}^{t+\tau}\) of adjacent frames of key frames and aggregate heatmap ROIs \(L = \{M_j\}_{j=t-\tau}^{t+\tau}\) within the group. Since the size of heatmap ROI is different, the scale of its

---

**FIGURE 1.** CenterNet network flow chart, where the fully-convolutional encoder-decoder backbone network can be selected as ResNet + DCNv2.

**FIGURE 2.** Heatmap ROI extraction by peaks and regressed width and height of box encompassing peaks.
feature matrix needs to be unified. We use (3) to align the k-th heatmap ROI of the j-th adjacent frame under the same category

\[ h_{RA} = \text{ROIAlign}(h_{RA}^{i,c}(m_j^t, M_t)) \] (3)

We extract the short edge \( \delta y \) in the height and width (\( \delta x_i, \delta y_i \)) of \( m_i^t \) as the downsampling scale. We refer to the ROIAlign [10] to downsample the features of adjacent heatmap ROI so that \( h_{RA} \in [0, 1]^{\delta x_i \times \delta y_i} \). The use of \( \delta y \) as the scale of ROIAlign is to avoid the case of 0-value padding due to the size of the long edge being larger than the adjacent heatmap ROI and to effectively reduce the amount of operations.

For a single heatmap ROI we use (4) to calculate its related heatmap with adjacent heatmap ROI via dot product.

\[ h_{Rel}^{i,c,j_k} = \begin{cases} h_{RA}^{i,c}(m_j^t, M_t) \cdot h_{RA} & \text{if } \delta x_i \leq \delta y_i \\ h_{RA} \cdot h_{RA}^{i,c}(m_j^t, M_t) & \text{if } \delta x_i > \delta y_i \end{cases} \] (4)

where \( h_{Rel}^{i,c,j_k} \in [0, 1]^{\delta x_i \times \delta y_i} \), \( t_i \) denotes the i-th heatmap ROI of key frame \( t \), \( c \) denotes the category, and \( j_k \) denotes the k-th heatmap ROI of the j-th of adjacent frames.

Each feature point of \( h_{RA}^{i,c}(m_j^t, M_t) \) is dotted with and each feature point of \( h_{RA} \). The geometric meaning of the dot product of vectors is to calculate the similarity of two vectors. More similar two vectors yield larger dot product value. Larger \( h_{Rel}^{i,c,j_k} \in [0, 1]^{\delta x_i \times \delta y_i} \) value indicates that the two ROI features are more similar, and the weight of the augmentation of the key frame heatmap ROI is also larger. Therefore, \( h_{Rel}^{i,c,j_k} \in [0, 1]^{\delta x_i \times \delta y_i} \) is viewed as the relation heatmap feature of self-attention.

**D. HEATMAP ROI AGGREGATION**

As shown in Fig.3, adjacent heatmap ROIs are further aggregated to enhance the heatmap ROI of the key frame. Such an operation intends to boost object detection with augmented object heatmap features.

After calculating the relation heatmap features, to increase the nonlinearity of the aggregation module, we perform 1 × 1 convolution of \( h_{Rel}^{i,c,j_k} \), Batch Normalization and ReLU operations, to output \( h_{RA}^{i,c,j_k} \). The aggregated heatmap ROI on the key frame can be represented as

\[ h_R^i(m^t, M_t) = h^{i,c}(m_i^t, M_t) + \alpha \cdot \text{concat}(\sum_{b=1}^k \sum_{a=1}^n h_{RA}^{i,c,j_k}) \] (5)

where \( h_R^i(m^t, M_t) \) denotes the final augmented heatmap ROI. It is expected that the augmented heatmap contains more object feature information than the original heatmap \( h^{i,c}(m_i^t, M_t) \), and can effectively boost the object detection results. To solve the detection problem of multiple objects and small targets, we augment multiple heatmap ROIs for multiple classes. As shown in Fig. 4, we establish long-term heatmap ROI feature augmentation during training and testing.

**IV. EXPERIMENTS**

**A. DATASET**

We trained and tested our method on the ImageNet VID dataset. ImageNet VID dataset is a large dataset for video detection tasks, a subset of ImageNet 200 categories contains 30 categories, the training set consists of 3862 videos and the validation set contains 555 videos. We evaluated our method on the validation set using the mean accuracy (mAP) as an accuracy metric.

**B. IMPLEMENTATION DETAILS**

1) CENTERNET

In order to compare the accuracy and real-time performance with the two-stage feature aggregation algorithm, our
backbone network is selected as ResNet-101. Following the same structure as the original CenterNet paper, [21] adds three layers of deconvolution to the output layer of the residual network to obtain a clearer feature map, [24] set the channels of the three-layer deconvolution to 256, 128, and 64, and add a $3 \times 3$ variable convolution layer in front of the deconvolution layer to enhance the generalization ability of the feature map.

2) HEATMAP ROI EXTRACTION

Our method of heatmap ROI extraction improves the method in [24] by extracting peak points and incorporating NMS to achieve better performance. The object features in the heatmap are concentrated in one area, and redundant heatmap ROI are generated when the peak points are close to each other. We extract the peaks of each category independently on the heatmap, and detect whether the value at that point is larger than the $7 \times 7$ matrix centered at that point, keeping the first 10 values. The peaks of 30 categories are compared and the peak points of the top five categories with the highest peaks are retained. NMS processing with IoU = 0.3 is performed on the enclosing frame of the generated heatmap ROIs to select the essential heatmap ROIs with significant object features.

3) HEATMAP ROI AGGREGATION

Small sampling values of adjacent frames during the experiment will cause feature redundancy and high similarity of relation features in a small range, which makes it difficult to aggregate to other pose or shape features of the object, and the network robustness decreases. We take $\tau = 50$ in the training process, and sample adjacent frames within 50 frames before and after key frame. In the inference process, we reduce the computation by filtering heatmap ROI, but retaining the larger span of adjacent frames still occupy memory. The range of values of adjacent frames in the inference process is reduced by setting $\tau = 12$. The heatmap ROIs in four adjacent frames are used to enhance the key frame heatmap. The comparison of heatmap ROI aggregation with different frames on the performance is given in the following experimental part.

4) TRAINING AND INFERENCE

Following common protocols in [5], [20], [23], [25], we train our model on the intersection of the ImageNet VID and DET datasets (30 classes). The pixel size of the input image is adjusted to $512 \times 512$, the resolution of the output heatmap of CenterNet increases with the size of the input image and the accuracy of detection increases. The network model was trained on RTX 3090 GPU. We used Adam as the optimizer with a batchsize of 16 to train 100 epochs. We used the PyTorch function ReduceLROnPlateau to adaptively adjust the learning rate, with an initial learning rate of 1e-4. In the inference phase, we use the full validation set for our experiments.

C. MAIN RESULTS

We show the comparison of our model with state-of-the-art in Table 1. We compare with CenterNet baseline, which is 5ms slower in runtime and 5% better in accuracy. CenterNet HP propagates the heatmap over the time series and predicts the object center point by weighting the propagated heatmap with the heatmap of key frame. Compared with CenterNet HP, we build the object heatmap feature extraction on the time series and aggregate the object heatmap features of multiple frames on the key frame object, our method improves the accuracy by 2.1%.

In comparison with two-stage multi-frame aggregation video detectors, we discard the RPN filtering and regression of the two-stage network model and use the CenterNet as the center of the heatmap and the aspect information of the one-stage network model to extract the object features of multiple frames in the video stream and aggregate them to key frame. The vast majority of two-stage network models cannot achieve real-time video detection at 25 FPS, our model can achieve 36 FPS real-time detection, which is a minimum 4 times improvement in real-time compared to the two-stage video detector.

The accuracy of most of the video detectors can be improved by continuing post-processing. We use the post-processing method of Seq-NMS [9]. Seq-NMS can use the timing information to re-score the candidate frames and avoid using NMS which would incorrectly select the candidate frames. As shown in Table 2, our model achieves 80.5% mAP at 33 FPS with ResNet-101.

D. ABLATION STUDY

In order to examine the impact of the different modules in our model, we performed a number of comparative experiments to investigate the impact of changes in these modules on the final performance.
TABLE 1. Performance comparison between our model and state-of-the-art video object detection models on the ImageNet VID validation set. To quantify the performance all models use ResNet-101 as the backbone network. In the table, all accuracies are without post-processing, accuracies and runtime are taken from the paper. - indicates data not given in the source paper.

| Methods                                      | Backbone          | mAP50(%) | FPS  | Runtime(ms) |
|----------------------------------------------|-------------------|----------|------|-------------|
| PGFA [25]                                   | ResNet-101        | 76.3     | 1.2  | 733         |
| STSN [11]                                   | ResNet-101+DCN    | 78.9     | -    | -           |
| D&T [7]                                     | ResNet-101        | 73.8     | 6.8  | 141         |
| MEGA [4]                                    | ResNet-101        | 82.9     | 8.7  | 114.5       |
| RDN [5]                                     | ResNet-101        | 81.8     | 10.6 | 94.2        |
| CenterNet                                   | ResNet-101+DCN    | 73.8     | 44   | 22.7        |
| CenterNet+Heatmap Propagation [23]          | ResNet-101+DCN    | 76.7     | 37   | 27          |
| CenterNet+Multi-frame Heatmap ROI Aggregation(ours) | ResNet-101+DCN    | 78.8     | 36   | 27.7        |

TABLE 2. Performance comparison between our model and state-of-the-art video object detection models with post-processing methods. - indicates data not given in the source paper.

| Methods                                      | Post-Processing  | mAP50(%) | FPS  | Runtime(ms) |
|----------------------------------------------|-------------------|----------|------|-------------|
| PGFA [25]                                   | Seq-NMS           | 78.4     | 1.1  | 873         |
| STSN [11]                                   | Seq-NMS           | 80.4     | -    | -           |
| D&T [7]                                     | Viterbi           | 79.8     | 5.5  | 187         |
| MEGA [4]                                    | BLR               | 84.5     | -    | -           |
| RDN [5]                                     | BLR               | 83.8     | -    | -           |
| CenterNet                                   | Seq-NMS           | 75.8     | 41   | 24.4        |
| CenterNet+Heatmap Propagation [23]          | Seq-NMS           | 78.4     | 34   | 29          |
| CenterNet+Multi-frame Heatmap ROI Aggregation(ours) | Seq-NMS           | 80.5     | 33   | 30          |

1) SELECTION OF HEATMAP ROI

CenterNet paper uses a $3 \times 3$ maxpooling operation to extract peaks greater than or equal to the other 8 adjacent peaks on the heatmap of each category and keep the top 100 peaks. We compare the impact on model performance caused by different downsampling schemes for selecting peaks in Table 3.

We first tried the original CenterNet paper’s method of extracting peaks using the corresponding heatmap ROI generated from the peaks. Although there is a distance between the peaks, there is extensive overlap between the heatmap ROI covering each other, and only single-digit heatmap ROI remain after NMS filtering. We reduce the retained peak to 10 points with almost no loss of accuracy, and runtime is reduced by 0.4 ms. We achieve an accuracy of up to 78.8% when using $7 \times 7$ maxpooling for peak extraction, and continuing to expand the downsampling size does not improve model performance. We also tried GlobalMaxpooling for peak extraction, but the peaks are too concentrated, resulting in too high coverage of the extracted heatmap ROI, which is not conducive to the detection of small objects and multiple objects.

Considering that when multiple classes of objects appear in a video, aggregation of heatmap of only the highest-peaked classes does not enhance the detection of objects in other classes, we perform heatmap ROI aggregation for the top N classes with the highest peaks in the video. As shown in Table 4, different N values have an impact on the accuracy and real-time performance of the model.

We first designed the model considering only heatmap ROI aggregation for a single category, and the highest peak category was selected for aggregation during training or inference. However, the actual situation is that multiple categories of objects exist in a video stream, and the object recognition results of other categories are not enhanced. When we set the aggregation categories $N$ to 3, the accuracy is improved by 0.7% and the runtime is increased by 0.3 ms. When we increase the value of $N$, maximum accuracy is achieved at $N = 5$.

2) PARAMETER $\alpha$

We investigated the effect of the weight parameter $\alpha$ in (5), after as shown in Table 5. $\alpha$ as a hyperparameter determines
Y. Zhou et al.: Multiframe CenterNet Heatmap ROI Aggregation for Real-Time Video Object Detection

FIGURE 5. Examples of qualitative results in ablation experiments. The four images are taken from four discrete frames of a video. CenterNet as a video detector, (a) three frames fail to detect small objects and the detection accuracy of large objects is low. After a single-category heatmap ROI aggregation, (b) only one frame of the small object is lost and the detection accuracy of large objects is improved by 0.14 on average. After five categories of heatmap ROI aggregation, (c) although the detection accuracy of large objects is basically not improved, the detection accuracy of small objects is improved by 0.21 on average.

TABLE 5. Ablation study of parameters $\alpha$.

| $\alpha$ (default marked by *) | 0.05 | 0.1* | 0.15 | 0.2 | 0.3 |
|--------------------------------|------|------|------|-----|-----|
| mAP50(%)                       | 77.6 | 78.8 | 78.5 | 78.0 | 77.8 |

TABLE 6. Performance versus number of frames aggregated (default marked by *).

| Number of frames aggregated | 1     | 2     | 3     | 4*    | 5     |
|-----------------------------|-------|-------|-------|-------|-------|
| mAP50(%)                    | 76.0  | 77.3  | 78.1  | 78.8  | 78.9  |
| Runtime(ms)                 | 23.9  | 25.2  | 26.4  | 27.7  | 28.8  |

the proportion of relevant features that enhance the key frame heatmap ROI. $\alpha$ has a linear effect on the heatmap features, and the accuracy of the network model reaches its maximum when $\alpha = 0.1$ continuing to increase the value of $\alpha$ leads to a decrease in accuracy. The possible reasons are that the feature values within heatmap ROI are too high resulting in the prediction value $\hat{Y}_{xyc}$ being too close to 1 and the loss increases.

3) NUMBER OF FRAMES AGGREGATED

As shown in Table 6, we conducted ablation experiments with different number of the aggregated frames. It can be seen that the accuracy grows with the number of frames aggregated, and best trade-off performance is achieved when the number of frames aggregated is four. When aggregating five frames, the real-time performance deteriorates and the accuracy hardly grows.

V. CONCLUSION

In this paper, we propose a multi-frame heatmap ROI aggregation module for CenterNet to improve the accuracy of real-time video object detection. In our proposed method, peak detection and regressed box size (width and height) are used to obtain heatmap ROIs. Each selected heatmap ROI corresponds to a potential object heatmap region. Furthermore, multiple heatmap ROIs of adjacent frames are aggregated to the key frame to enhance the heatmap features of candidate objects. Experiments conducted on ImageNet VID dataset validate the effectiveness of our proposed method. Compared to CenterNet Heatmap Propagation [23], our proposal achieves 2.1% improved mAP with almost no sacrifice of frame processing time.

REFERENCES

[1] G. Bertasius, L. Torresani, and J. Shi, “Object detection in video with spatiotemporal sampling networks,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 331–346.
[2] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, “YOLOv4: Optimal speed and accuracy of object detection,” 2020, arXiv:2004.10934.
[3] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6154–6162.
[4] Y. Chen, Y. Cao, H. Hu, and L. Wang, “Memory enhanced global-local aggregation for video object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10337–10346.
[5] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, and T. Mei, “Relation distillation networks for video object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7023–7032.
[6] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, “FlowNet: Learning optical flow with convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2758–2766.
[7] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Detect to track and track to detect,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3038–3046.

[8] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.

[9] W. Han, P. Khorrami, T. Le Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, “Seq-NMS for video object detection,” 2016, arXiv:1602.08465.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., Oct. 2017, pp. 2961–2969.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[12] K. Kang, W. Ouyang, H. Li, and X. Wang, “Object detection from video tubelets with convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 817–825.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 25, 2012, pp. 1097–1105.

[14] H. Law and J. Deng, “CornerNet: Detecting objects as paired keypoints,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 734–750.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.

[16] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767.

[17] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 28, 2015, pp. 91–99.

[18] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.

[20] S. Wang, Y. Zhou, J. Yan, and Z. Deng, “Fully motion-aware network for video object detection,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 542–557.

[21] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 466–481.

[22] F. Xiao and Y. J. Lee, “Video object detection with an aligned spatial-temporal memory,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 485–501.

[23] Z. Xu, E. Hrustic, and D. Vivet, “Centernet heatmap propagation for real-time video object detection,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Springer, 2020, pp. 220–234.

[24] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” 2019, arXiv:1904.07850.

[25] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, “Flow-guided feature aggregation for video object detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 408–417.

[26] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, “Deep feature flow for video recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2349–2358.

YOU ZHOU received the B.S. degree from Shihezi University, China, in 2019. He is currently pursuing the master’s degree with Hainan University. His research interests include machine learning and computer vision.

YONG BAI (Senior Member, IEEE) received the B.S. degree from Xidian University, China, in 1992, the M.S. degree from the Beijing University of Posts and Telecommunications (BUPT), China, in 1995, and the Ph.D. degree from Rutgers—The State University of New Jersey, in 2001. He was with PacketVideo Corporation, from 2000 to 2002, Motorola, from 2002 to 2004, and CEC Wireless, from 2004 to 2005. He was a Senior Researcher with the DOCOMO Beijing Communication Laboratories, from 2005 to 2009. He has been a Professor with the School of Information and Communication Engineering, Hainan University, since 2010. His current research interests include artificial intelligence and mobile communications. He has been serving as an Associate Editor for the EURASIP Journal on Wireless Communications and Networking, since March 2017.

YONGQING CHEN received the B.S. degree from the South China University of Technology, Guangzhou, China, in 2009, and the M.S. degree in software engineering from Sichuan University, in 2015. He is currently pursuing the Ph.D. degree with Hainan University. His research interests include pattern recognition, machine learning, and computer vision.