Multi-level Spatial-temporal Feature Aggregation for Video Object Detection

Chao Xu, Jiangning Zhang, Mengmeng Wang, Guanzhong Tian, Yong Liu

***Abstract*—Video object detection (VOD) focuses on detecting objects for each frame in a video, which is a challenging task due to appearance deterioration in certain video frames. Recent works usually distill crucial information from multiple support frames to improve the reference features, but they only perform at frame level or proposal level that cannot integrate spatial-temporal features sufficiently. To deal with this challenge, we treat VOD as a spatial-temporal hierarchical features interacting process and introduce a *Multi-level Spatial-Temporal* (MST) feature aggregation framework to fully exploit frame-level, proposal-level, and instance-level information in a unified framework. Specifically, MST first measures context similarity in pixel space to enhance all frame-level features rather than only update reference features. The proposal-level feature aggregation then models object relation to augment reference object proposals. Furthermore, to filter out irrelevant information from other classes and backgrounds, we introduce an instance ID constraint to boost instance-level features by leveraging support object proposal features that belong to the same object. Besides, we propose a *Deformable Feature Alignment* (DAlign) module before MST to achieve a more accurate pixel-level spatial alignment for better feature aggregation. Extensive experiments are conducted on ImageNet VID and UAVDT datasets that demonstrate the superiority of our method over state-of-the-art (SOTA) methods. Our method achieves 83.3% and 62.1% with ResNet-101 on two datasets, outperforming SOTA MEGA by 0.4% and 2.7%.**

***Index Terms*—Video Object Detection, Feature Alignment, Feature Interaction, Instance ID Constraint.**

I. Introduction

Video object detection [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11] is a task to automatically annotate every object with its bounding box and class label in each frame of the videos, which has promising application capabilities in visual monitoring systems and self-driving vision systems. Although object detection in a single image [12], [13], [14], [15], [16], [17], [18], [19], [20], [21] has achieved remarkable success due to the development of deep convolutional networks [22], [23], [24], [25], [26], video object detection remains a challenging problem. One key element of videos is temporal information. If image object detection frameworks are directly applied to videos frame by frame, the detection confidences of objects show dramatic changes between adjacent frames, especially the frames that suffer deteriorated appearance, *e.g.*, occlusion, motion blur, video defocus, and pose variation. We report that the proportion of degraded frames in ImageNet [27] validation set is 9.15%, which has a significant impact on the VOD performance. The left part of Figure 1 showcases some hard examples in VOD.

Since the videos inherently contain richer temporal and motion context than individual images, one direct way is to make full use of temporal information from neighboring frames to solve the object appearance deterioration problem. Specifically, the frame to be detected is called the reference frame and some neighboring frames as support frames. Most existing methods focus on distilling critical information from the support frames and fuse the distilled information into the reference frame to generate enhanced features for robust detection. The distillation and fusion operations are mainly applied at frame-level [1], [2], [3], [28], [29] or proposal-level [4], [5], [6], [30] features.

For frame-level methods, some works employ optical flow as external guidance for feature aggregation. FGFA [1] adopts a optical flow network to calculate the spatial relationship between frames, which guides the per-frame aggregation of nearby features over time. THP [3] also extracts optical flow for propagating keyframe features to non-keyframe features. However, optical flow is widely used in feature aggregation and warping, implemented by an extra model, and significantly increase model size and computation. Recently, attention-based methods have illustrated impressive results. STMM [29] proposes a novel MatchTrans module that models the displacement introduced by motion across frames to achieve accurate pixel-level spatial features over time. PSLA [28] further argues that the gap between optical flow and advanced features may prevent accurately establishing spatial correspondence. They replace optical flow with progressive sparse local attention module to propagate high-level semantic features. Although frame-level aggregation achieves fine-grained feature augmentation, it performs in a global manner, which fails to focus on critical foreground objects.

To focus on object features and fully explore their relationship in videos, Hu *et al.* [20] propose the relation network to explore the dependencies among video frames for feature aggregation. Their basic idea is to measure proposal features as the weighted sum of appearance features from other objects in the same image and other support frames. The weights reflect object similarity in terms of appearance and geometry information. Subsequently, RDN [4] proposes a two-stage framework that first generates proposals for reference and support frames and then models object relation in spatial-temporal context to boost the quality of reference feature. MEGA [5] aggregates both global and local information to key frames at the proposal level and introduces a memory module to enable key frames access to more context from

Y. Liu is the corresponding author

C. Xu, J. Zhang, M. Wang, G. Tian and Y. Liu are with Zhejiang University, Hangzhou, China (e-mail: 21832066@zju.edu.cn; 186386@zju.edu.cn; mengmengwang@zju.edu.cn; gztian@zju.edu.cn; yongliu@iipc.zju.edu.cn).

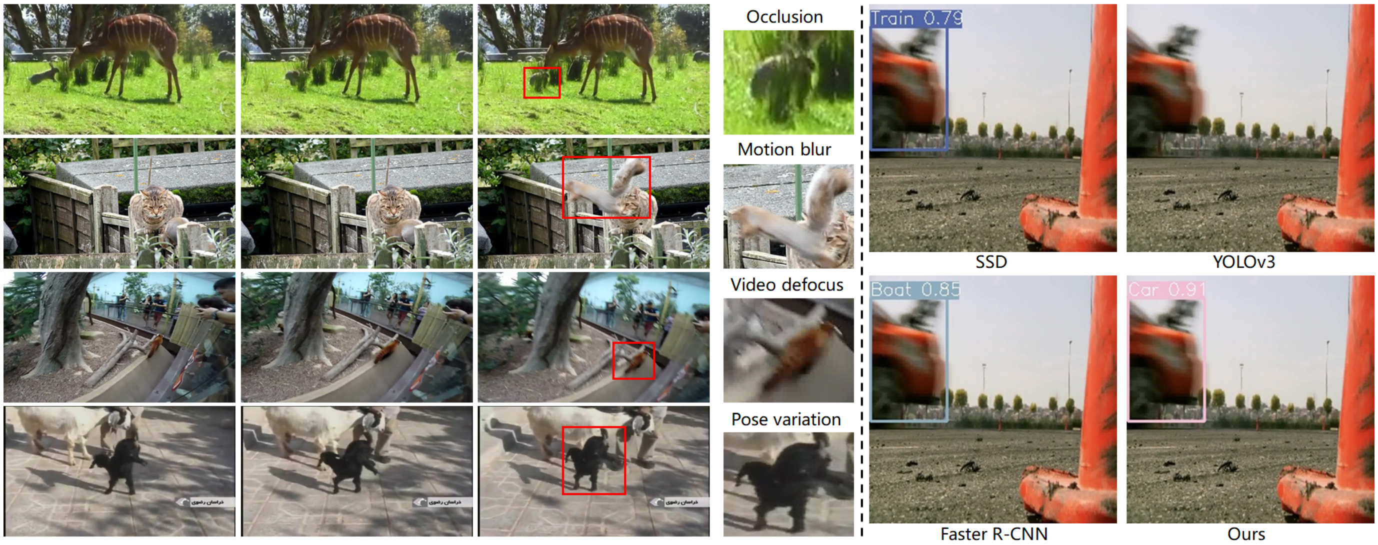


Fig. 1. The left part shows some examples of deteriorated video sequences. , The conditions are occlusion, motion blur, video defocus, and pose variation from top to bottom. We mark the low-quality foreground areas as red bounding boxes and zoom in on them. The right part shows a case that SSD [31], YOLOv3 [21], and Faster R-CNN [14] all fail to predict correct class label. Our method can predict the car with high confidence.

previous frames. The above methods directly aggregate all the proposals from support frames without considering whether they belong to the same instance or not.

In this paper, rather than perform feature aggregation at a single level, we propose a novel *Multi-level Spatial-Temporal* (MST) feature aggregation framework to effectively utilize frame-level, proposal-level, and instance-level features hierarchically for better aggregation. First, the frame-level feature aggregation enhances all input features with sufficient spatial-temporal information instead of only updating the reference features. We observe that the upgraded support features could provide more reliable cues for feature aggregation. Second, we follow the basic stage in [4] and design the proposal-level feature aggregation after the RPN to augment each object's features in the reference frame by aggregating its relation support features over the proposals. Third, to filter out the proposal features that from other instance and background, we further introduce the instance-level feature aggregation, which only leverage the proposals assigned with the same instance ID to enhance the reference object proposal features, and the proposals related to the background are not updated in this stage. In this way, we can explicitly aggregate the multi-level spatial-temporal features to generate more robust features.

Besides, current relation-based methods neglect feature alignment before feature aggregation. They aggregate the features from both support and reference frames. Such unaligned features would confuse relation networks for similarity calculation. To deal with this problem, we design a Deformable Feature Alignment (DAlign) module to aligns the support and target features with different poses and shapes. We insert it before MST. Thus the frame features and proposal features are both spatially aligned across frames.

In summary, to address the performance degradation in VOD due to appearance deterioration, we first employ DAlign to align multiple frames in the temporal domain for better feature aggregation. MST is then proposed to aggregate aligned multi-level features to augment the deteriorated features. The three modules of MST are all inspired by the relation network that is designed in the attention mechanism. As shown in the right part of Figure 1, our method could handle a frame with occlusion and motion blur that produce correct detection results. We make the following three contributions:

- We introduce a *Deformable Feature Alignment* (DAlign) module that uses deformable convolutions across space and time to align features between frames.
- We devise a *Multi-level Spatial-Temporal* (MST) feature aggregation framework that performs feature aggregation hierarchically at the frame level, proposal level, and instance level to obtain more robust aggregated features.
- The proposed method is evaluated on ImageNet VID and UAVDT datasets and achieves the superior performance of 83.3% mAP and 62.1% with ResNet-101, respectively.

II. Related Work

A. *Object Detection in Images*

Benefit of deep Convolutional Neural Networks (CNN) [22], [23], [24], [25], [26] and well-annotated dataset [32], the image object detection [12], [13], [14], [15], [16], [17], [18], [19], [21], [33], [34] have achieved remarkable improvements, which is widely applied in face detection [35], [36], [37], intelligent transportation [38], [39], [40], SAR image processing [41], [42], [43], [44], [45], [46], and so on. There are generally two directions for object detection. Two-stage detectors usually perform region proposals first, and then the proposals are refined by classification and regression. R-CNN first utilizes selective search to generate region proposals. To speedup, SPPNet [47] and Fast R-CNN [15] introduce SPP pooling and Roi pooling to handle multi-size images and avoid calculating features twice. Faster R-CNN [14] further replaces

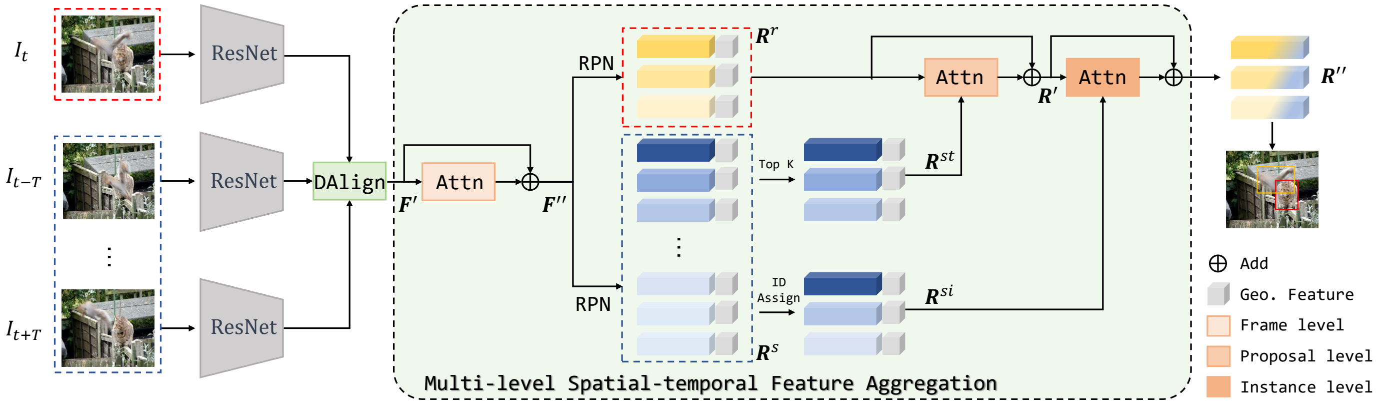


Fig. 2. The pipeline of our proposed method. Given a reference frame I_t and a set of support frames $\mathbf{I} = \{I_\tau\}_{\tau=t-T}^{t+T}$, we first extract their frame-level features using the backbone feature extractor. DAlign follows to align features from support frames to the reference frame, obtaining $\boldsymbol{F}^{'}$. Then, the aligned features go through MST. Specifically, We apply frame-level feature aggregation to enhance all input features. The enhanced frame-level features $\boldsymbol{F}^{''}$ are input to RPN to generate the high-quality proposals, obtaining reference proposals $\mathbf{R}^r$ and support proposals $\mathbf{R}^s$. Furthermore, $\mathbf{R}^r$ is updated by the top-K support proposals $\mathbf{R}^{st}$ in proposal-level feature aggregation, and the foreground proposals of $\mathbf{R}^{'}$ are updated by the associated support proposals $\mathbf{R}^{si}$ in instance-level feature aggregation sequentially. Finally, the enhanced proposal features $\boldsymbol{R}^{''}$ are fed into the detection head for classification and regression.

selective search by Region Proposal Networks, which introduces the anchor to generate region proposals more reliable.

In contrast, one-stage detectors directly output object coordinates and categories without region proposal stage. YOLO [33], [34], [21] divides the feature map into rigid grids. Each of them is responsible for detecting the objects located in the grid. Another one-stage detector, SSD [31], borrows the idea of anchor and combines different scale features to boost detection for objects in various scales and aspect ratios. Although one-stage methods faster than two-stage one, but usually has lower performance. One major reason lies in unbalanced positive and negative proposals. RetinaNet [18] designs a new loss named Focal Loss to ease unbalanced problem. GHM [48] proposes a gradient harmonizing mechanism to solve the problem of sample imbalance. Recently, point-based methods [49], [50], [51], [52], [53], [54] are designed to get rid of the limitation of anchors. In this paper, we build our method upon Faster R-CNN [14], which is one of the state-of-the-art object detectors.

B. *Object Detection in Videos*

Due to the complex manner of video variation, such as motion blur, occlusion, video defocus, and rare pose, it is not trivial to directly apply a single image detector into the video domain. One direction of video object detection is the box-level association, which associates bounding boxes from consecutive frames to generate tubelets [55], [56] by linking or tracking. For instance, D&T [57] designs a correlation network to predict the local displacement between two frames. Chen *et al.* [58] propagate and refines key frame boxes through Motion History Image (MHI). Yao *et al.* [59] directly use real-time trackers to exploit temporal information and track the bounding boxes in the next frames. Besides, some offline post-processing methods [60], [4] integrate per-frame proposals into tubelets for re-scoring to further improves the robustness of video object detection. However, these methods are challenging to correct the errors produced by the associated image object detectors.

Another common solution is feature aggregation that enhances per-frame feature by aggregating local frames or global frames. Specifically, FGFA [1] utilizes optical flow estimated by FlowNet [61] to propagate feature across frames. However, an extra model to predict flow would significantly increase the model size. PSLA [28] establishes the spatial correspondence between two feature maps to propagate high-level semantic features among them without relying on optical flow. Besides of above frame-level aggregation methods, RDN [4] based on Relation Network to learn the relation among proposals of different frames in a local range. In contrast, SELSA [6] aggregate box features in the full-sequences level to capture more discriminative and robust features. To seek the full merit of both local and global aggregation, MEGA [5] strengthen boxes features by exploiting the relation across local and global frames. Besides, SPFTN [62] learns video object detection and video object segmentation [63], [64] in a unified frame work to facilitate each other. Unlike these methods that separately enhance per-frame feature by frame-level or proposal-level aggregation, we propose a hierarchically feature aggregation strategy. Our model combines frame-level, proposal-level, and instance-level modules in a unified framework. The proposed instance-level feature aggregation follows after the proposal-level feature aggregation to further enhance reference object proposal features by the support proposals with the same instance ID.

C. *Self-attention Mechanism*

Attention [65] module plays a critical role in NLP and starts supporting other computer vision tasks, such as object detection and semantic segmentation. In particular, Hu *et al.* [20] presents relation networks to explore the relations among object proposals. Wang *et al.* [66] adds non-local module to capture contextual information within feature maps. CC-Net [67] designs criss-cross attention to obtaining contextual information more effective and efficient. Moreover, current works like [4], [5], [6], [68] extend self-attention to a temporal domain to boost video object detection. In practice, attention-based methods usually split features into separated channel-wise features with equal channels, and then the separated features are input to the multi-head attention module. The previous methods have proven the effectiveness of the attention for feature aggregation. Thus our method is also built upon the attention mechanism.

III. Method

In this section, our proposed method is first under a brief overview. Then, we introduce two key components in our method. We design DAlign to deal with the object motion and align the feature from support frames to the reference frame. MST is proposed to perform feature aggregation at the frame level, proposal level, and instance level. Each feature aggregation follows the multi-layer and multi-head design.

A. Overview

The pipeline of our method is illustrated in Figure 2. It is built on the Faster R-CNN image-based detector. Formally, given a sequence of adjacent frames $\mathbf{I} = \{I_\tau\}_{\tau=t-T}^{t+T}$, where the central frame I_t is reference frame and the whole frames $\{I_\tau\}_{\tau=t-T}^{t+T}$ are support frames, we first extract the frame-level features $\mathbf{F} = \{F_\tau\}_{\tau=t-T}^{t+T}$ by using the backbone feature extractor $\mathcal{N}_{feat}$. Then, we apply DAlign $\mathcal{N}_{ali}$ to achieve accurate pixel-level spatial alignment over time, generating aligned features $\mathbf{F}^{'} = \left\{F_\tau^{'}\right\}_{\tau=t-T}^{t+T}$. After that, the proposed frame-level feature aggregation $\mathcal{N}_{fra}$ performs on these aligned features to obtain the corresponding enhanced frame-level features $\mathbf{F}^{''} = \left\{F_\tau^{''}\right\}_{\tau=t-T}^{t+T}$, which are then input into RPN to generate reference proposals $\mathbf{R}^r$ and all support proposals $\mathbf{R}^s$. We feed $\mathbf{R}^r$ and top-K support proposals $\mathbf{R}^{st}$ into proposal-level feature aggregation $\mathcal{N}_{pro}$ to update $\mathbf{R}^r$, obtaining $\mathbf{R}^{'}$. Furthermore, an instance-level feature aggregation $\mathcal{N}_{ins}$ is followed to enhance each object proposal feature by corresponding support object proposal features $\mathbf{R}^{si}$ that are belong to the same object, the enhanced proposal features denoted as $\mathbf{R}^{''}$, which contains the upgraded instance features and other object-irrelevant features that directly copy from $\mathbf{R}^{'}$. Finally, we feed the enhanced proposal features into the detection head for object classification and bounding box regression.

B. Deformable Feature Alignment

Due to the motion dynamics, the objects in the reference frame and support frames usually present different poses and shapes. Subsequently, the features of these frames are not aligned in spatial, leading to confusing information and makes the detector fail to obtain correct recognition and accurate localization. Therefore, feature alignment is crucial for better feature aggregation. The former works, DFF [2] and FGFA [1] adopt FlowNet [61] to model inter-frame motion for feature alignment, which is explicit but quite time-consuming. PSLA [28] replaces optical flow with a progressive sparse local attention module to improve the running efficiency. However, they use the fixed sample locations without considering the diversity of object movements. In contrast, we design our DAlign using deformable convolution layers to model pixel-level spatial alignment, which could learn flexible and diverse offsets to model the object deformation.

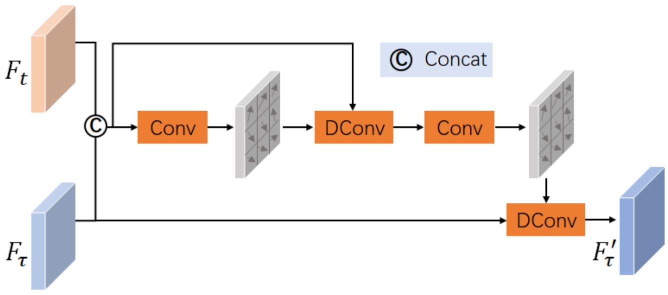


Fig. 3. The brief architecture of our DAlign with only two deformable convolution layers. Given the reference feature F_t and support feature F_τ, we concatenate F_t and F_τ, and then feed them through deformable convolution layers to produce offsets that are used to sample discriminative features from F_τ, obtaining aligned features $F_\tau^{'}$.

Specifically, reference feature F_t and support F_τ feature first concatenated into a new feature tensor that contains fine-grained features from both the reference and support frames. Then a convolution layer is followed to predict sampling parameters Θ for the features F_τ:

$$\Theta = f_\theta\left(F_t, F_\tau\right), \tag{1}$$

where $\Theta = \{\Delta p_n \mid n = 1, \ldots, |\mathcal{R}\,|\}$ refers to the offsets of the convolution kernels, $\mathcal{R} = \{(-1,-1), (-1,0), \ldots, (0,1), (1,1)\}$ denotes a regular grid of a 3×3 kernel. With Θ and F_τ, the aligned feature F_τ' can be computed by the deformable convolution, for each position p_0 on the aligned feature map F_τ', we have:

$$F_\tau'\left(p_0\right) = \sum_{p_n \in \mathcal{R}} \mathbf{w}\left(p_n\right) F_\tau\left(p_0 + p_n + \Delta p_n\right), \tag{2}$$

where $\mathbf{w}$ is the matrix that weighted summarize sampled values, p_n enumerates the locations in $\mathcal{R}$. The convolution will be operated on the irregular positions $p_n + \Delta p_n$, where the Δp_n may be fractional. To address this issue, the operation is implemented by using bilinear interpolation, which is the same as that proposed in [69].

We show a simplified version of DAlign in Figure 3. In practice, we first apply two deformable convolution layers on concatenated features sequentially to predict the offsets Θ between these two feature maps. Then, two deformable convolution layers are fed support feature F_τ and Θ to produce aligned feature $F_\tau^{'}$. The DAlign denotes as:

$$\mathbf{F}^{'} = \mathcal{N}_{\text{ali}}\left(\mathbf{F}\right) \tag{3}$$

C. Multi-level Spatial-temporal Feature Aggregation

Frame-level Feature Aggregation. Recent proposal-level feature aggregation such as RDN [4] only distill relation through proposals. They ignore that if the frame features input to the RPN are of low quality, the proposals could not cover the foreground with high scores, which will cause

This article has been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3183646

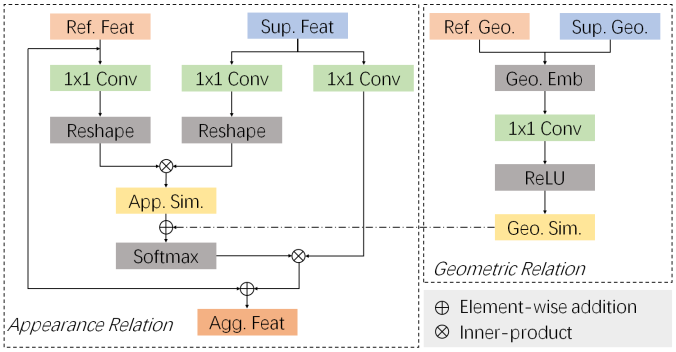


Fig. 4. The architecture of aggregation. The left paradigm shows the appearance similarity calculation and feature aggregation. The right paradigm shows the geometric similarity calculation. Note that we abbreviate reference, support, appearance, geometric, and aggregated to Ref., Sup., App., Geo., Agg., respectively. The frame-level feature aggregation follows the left paradigm, and the proposal-level and instance-level feature aggregations are follow the whole paradigm.

the subsequent proposal-level feature aggregation to easily combine some irrelevant information. In order to improve the reliability of proposal-level feature aggregation, we first perform feature aggregation at pixel level. Unlike previous works that only enhance the reference feature by itself and other support features, we argue that the enhanced support features could also provide key information for feature fusion. So we take both reference and support features as input and simultaneously enhance all of them.

Specifically, we follow the multi-head design that promotes learning coherent spatial-temporal transformations for leveraging the rich information in videos. For a more clear description, we take one attention head for example. As shown in Figure 4, given aligned frame-level features $\mathbf{F}^{'} = \left\{F_{\tau}^{'}\right\}_{\tau=t-T}^{t+T}$, we first apply three transformation layers on $\mathbf{F}^{'}$ to obtain Query, Key, and Value:

$$\mathbf{Q},(\mathbf{K},\mathbf{V}) = M_q\left(\mathbf{F}^{'}\right),\left(M_k\left(\mathbf{F}^{'}\right),M_v\left(\mathbf{F}^{'}\right)\right), \quad (4)$$

where $M_q(\cdot)$, $M_k(\cdot)$, $M_v(\cdot)$ denote the 1×1 convolution layers. After that, the reshape operation is followed and the similarities between reshaped query features and key features is calculated as follows:

$$\mathbf{A} = softmax(\frac{\mathbf{Q}\cdot\mathbf{K}^{Trans}}{\sqrt{d_k}}), \quad (5)$$

where d_k is hidden dimensions of each projection subspace. Each element of $\mathbf{A}$ indicates the relation between each spatial location in the query and key features. Thus, the value features are weighted-summed with the attention values as the summation weights, which is then element-wisely added to the original input features to generate the final updated reference and support features:

$$\mathbf{F}^{''} = \mathbf{AV} + \mathbf{F}^{'} = \mathcal{N}_{\text{fra}}\left(\mathbf{F}^{'}\right). \quad (6)$$

After the frame-level feature aggregation, each of the enhanced features can distill rich spatial-temporal information from the frame-level features of the other frames.

Proposal-level Feature Aggregation. For the proposal-level feature aggregation, we feed enhanced features $\mathbf{F}^{''}$ into RPN to generate a set of proposals $\mathbf{R}$, which consists of reference proposals $\mathbf{R}^r = \{R_1^r, R_2^r, \ldots, R_m^r\}$ and top-K support proposals $\mathbf{R}^{st} = \{R_1^{st}, R_2^{st}, \ldots, R_n^{st}\}$. The object proposals are represented with their geometric features g and appearance features a, which are both exploited to alleviate the information distortion problem caused by noise during the aggregation process. Following the RDN, we devise a multi-head and stacked architecture, aiming to enhance each proposal in the reference frame with the most informative appearance from multiple proposals in the support frames.

Formally, given the $\mathbf{R}^r$ and $\mathbf{R}^{st}$, we first apply two fully connections on reference and support proposal features to obtain the query and keys. The reference proposal feature is projected to the query, while the support proposal features are projected to the keys. Then, as shown in Figure 4, we measure each proposal pair not only based on the appearance information but also based on the geometric information, which is calculated as follows:

$$s_{i,j} = \frac{\exp\left(as_{i,j} + gs_{i,j}\right)}{\sum_{v=1}^{n}\exp\left(as_{i,v} + gs_{i,v}\right)}, \quad (7)$$

where i, j are the index of reference and support proposals, $as_{i,j}$ and $gs_{i,j}$ represent the appearance similarity and geometric similarity between R_i^r and R_j^{st}. $as_{i,j}$ is formulated as:

$$as_{i,j} = < \phi\left(a_i^r\right), \varphi\left(a_j^{st}\right) >, \quad (8)$$

where ϕ and φ denote fully connections, a_i^r is i-th reference appearance feature and a_j^{st} is j-th support appearance feature. For calculating the geometric similarity, we use scale information (the width and the height), which is more reliable than spatial locations. The $gs_{i,j}$ is formulated as:

$$gs_{i,j} = \psi\left(\varrho\left(\log\left(\frac{w_i}{w_j}\right), \log\left(\frac{h_i}{h_j}\right), \log\left(\left|\frac{w_i}{h_j} - \frac{w_i}{h_j}\right|\right)\right)\right), \quad (9)$$

where h and w are height and width of bounding box, ψ denotes as a general transformation function, ϱ is the embedding function used in [20].

After calculating the similarity of the proposal pair, the feature aggregation is performed as a weighted summation of the support proposal features with the proposal pair similarities as summation weights mathematically. The aggregated support features are then added to the reference appearance features:

$$a_i^{'} = a_i^r + \sum_{j=1}^{n} s_{ij}\cdot\left(\mathbf{W}^p\cdot a_j^{st}\right), \quad i = 1,\cdots,m, \quad (10)$$

where $\mathbf{W}^p$ denotes the transformation matrix. $a_i^{'}$ is augmented reference appearance features. The whole proposal-level feature aggregation denotes as:

$$\mathbf{R}^{'} = \mathcal{N}_{\text{pro}}\left(\mathbf{R}^r, \mathbf{R}^{st}\right). \quad (11)$$

Instance-level Feature Aggregation. As depicted in the proposal-level feature aggregation, the reference proposal is enhanced with all support proposals to effectively augment the reference proposal features. However, it is prone to introduce irrelevant cues to crucial object features. Inspired by the instance information produced by instance ID, we propose the instance-level feature aggregation to further enhance the proposal features associated with the foreground objects. In practice, we update the object proposal features by support proposal features that belong to the same objects. The background proposal features would not participate in the aggregation process. Such a design filters out invalid support proposals and reduces the computation cost for relation reasoning.

Technically, given the augmented reference proposals $\mathbf{R}^{'}$ and the support proposals $\mathbf{R}^{s}$, we first assign both reference and support proposals with ID information according to the IoU. Thus, each ID has its reference and support proposals. The instance-level aggregation strengthens the reference object proposals by the corresponding support object proposals $\mathbf{R}^{si}$. The similarity calculation and feature enhancement are the same operations in proposal-level feature aggregation. The whole instance-level feature aggregation denotes as:

$$\mathbf{R}^{''} = \mathcal{N}_{ins}\left(\mathbf{R}^{'}, \mathbf{R}^{si}\right). \tag{12}$$

Note that the final proposal features consists of $\mathbf{R}^{''}_{fg}$ that related to the foreground and $\mathbf{R}^{''}_{bg}$ that related to background, the former are updated and the latter are directly copy from $\mathbf{R}^{'}$. We exploit $\mathbf{R}^{''}$ for proposal classification and regression.

IV. Experiments

In this section, we first briefly illustrate the datasets and evaluation protocols for video object detection. Then, we present the details of the network architecture and the implementation details both at the training and testing stages. After that, we compare our method with several state-of-the-art video object detection methods on ImageNet VID [27] and UAVDT [70]. Finally, we carry out efficient ablation studies on the ImageNet VID validation set to demonstrate the effectiveness of each proposed module.

A. Dataset

ImageNet VID dataset. It is a large benchmark for video object detection, consisting of 3,862 training and 555 validation videos in 30 classes. All bounding boxes are fully annotated with the class labels, coordinates, and instance id. Because the official testing set is not publicly available, we follow the widely adopted protocols [57], [71], [29], [1], [2] in video object detection, evaluating the mAP@IoU=0.5 scores on the validation set.

Due to the redundancy of videos, the objects of each category have limited appearance diversity. Therefore, as in the previous works [57], [71], [29], [1], [2], we utilize both ImageNet VID and ImageNet object detection (DET) dataset to train our network. The ImageNet DET dataset is a still image detection dataset with 200 categories, containing 30 categories in the ImageNet VID. Thus, we use images of 30 overlapped categories in the ImageNet DET for training.

Unmanned Aerial Vehicle Benchmark (UAVDT). It is a large scale challenging UAV Detection and Tracking benchmark, consisting of 40000 annotated frames belonging to 30 training videos and 30 testing videos. These frames are manually annotated with bounding boxes and instance ID for Multiple Object Tracking (MOT). With these annotations, it is also suitable for VOD.

B. Network Architecture

Backbone network. We adopt ResNet-101 [22] or ResNeXt-101-32×4d [72] as our backbone feature extractor. Following the previous works [4], [6], we set the stride of the first convolution block in *conv*5 of convolutional layers from 2 to 1, so the total stride of *conv*5 is changed from 32 to 16, the resolution of the feature map becomes doubled. Besides, we also modify all the 3x3 convolution layers in *conv*5 by the atrous convolution with dilation rate *d*=2 to further enlarge the receptive field of the backbone network.

Detection network. We adopt Faster R-CNN [14] as our detection network and apply RPN to the output of *conv*4. We design the anchors with 3 aspect ratios $\{1:2, 1:1, 2:1\}$ and 4 scales $\left\{64^2, 128^2, 256^2, 512^2\right\}$, resulting 12 anchors for each spatial location. A non-maximum suppression (NMS) with an IoU threshold of 0.7 is adopted to reduce redundancy during training and inferencing stages, and 300 candidate boxes are generated in each frame. After that, we apply RoI-Align [17] to the output of *conv5* and a fully connected layer followed to extract the RoI feature for each box.

Multi-level feature aggregation. We apply the DAlign to the output of *conv*4, and a frame-level feature aggregation equipped with one attention layer is followed to generate the enhanced feature map. The proposal-level feature aggregation with two stacked attention layers and instance-level feature aggregation with one attention layer are followed behind RoI-Align sequentially. The enhanced proposal features are then fed into the detection head for object classification and bounding box regression.

C. Implementation Details

Our model consists of a backbone feature extractor, DAlign, frame-level feature aggregation, RPN, proposal-level feature aggregation, instance-level feature aggregation, and a detection head sequentially. The backbone is initialized with the pre-trained weights on ImageNet [73], then all modules are trained and optimized simultaneously. The input images are resized to be with the shorter dimension of 600 pixels. The whole architecture is trained on 4 GPUs by the SGD optimizer with momentum of 0.9 and weight decay of 0.0001. Each GPU holds one mini-batch, and each mini-batch contains one image. Every reference frame is sampled during training along with two random support frames in the same video sequence with a temporal spanning range $T = 18$ as RDN. The IoU threshold in instance-level for assigning foreground proposals to each object instance is set to 0.5. We optimize location loss and regression loss simultaneously. When testing, we follow the RDN [4] and process each frame with a sliding feature buffer of the nearby frames. Except for the beginning and ending 18 frames, the feature buffer's capacity remains 37. Each feature buffer is composed of 36 support frames and a reference frame. Since the instance id used in instance-level feature aggregation could not be available during testing, we sample $K = 20$ proposals with the highest objectness scores from $\mathbf{R}^{\mathbf{s}}$ as candidate foreground instances, each of them sample the corresponding $r = 20\%$ proposals with the highest similarities to enhance its object proposal features. Besides, we adopt NMS with a threshold of 0.5 IoU to suppress reduplicate detection boxes.

D. Comparison with state-of-the-arts

End-to-End models. We show the performance of different end-to-end video object detection models on ImageNet VID in Table I. For a fair comparison, we only include the state-of-the-art end-to-end methods, which learn video object detectors by enhancing per-frame features in an end-to-end fashion without any post-processing. Among them, FGFA [1] fuses the features across frames with external guidance using optical flow, which is estimated by a FlowNet. MANet [74] uses box-level calibration to further improve per-frame features. STSN [75] replaces optical flow with MathTrans to propagate and aggregate features at frame-level. The above four methods focus on aggregating frame-level features. While our method benefits from the multi-level feature aggregation, gaining much better performance than FGFA, MANet, and STSN by +7.0%, +5.2%, +4.4%.

For proposal-level feature aggregation, note that there are two structures of feature storage, *sliding window* stores raw features of several neighbor frames of the current frames, and *memory bank* utilizes recurrent temporal connections to aggregate more temporal information from additional frames, even rely on temporal coherence of the whole video for prediction. Specifically, SELSA [6] calculates the semantic similarity between two proposals, which serves as guidance for the reference proposal to aggregate features from support proposals at the full-sequence level. RDN [4] designs a multi-stage network to aggregate and propagate object relation to augment proposal features in the local frames. LSTS [76] develops a more effective sampling method to mine the local motion information. Furthermore, HVRNet [77] integrates intra-video and inter-video proposal relations hierarchically. In addition to the above methods that use sliding windows structure, MEGA [5] introduces a global-local feature aggregation method. MAMBA proposes a new memory update strategy to utilize knowledge from the whole video. Compared with these works, our method considers feature alignment and extends frame-level and instance-level feature aggregation to effectively enhance the features for VOD. As a result, our method achieves a competitive mAP of 83.3%, and the mAP improvements compared with SELSA, RDN, MEGA, LSTS, and HVRNet are +3%, +1.5%, +0.4%, +3.2%, and +0.1%, respectively. Notably, our mAP is lower than MAMBA [78]. The reason lies that MAMBA focuses on enlarging the number of visible frames and proposes an effective memory updating strategy, while our method is dedicated to sufficiently using limited neighbor frames to exploit multi-level information for better feature aggregation. Although our method has not achieved the best performance among all the competitors, we outperform other methods that use sliding windows. Moreover, as shown in Table I, compared to MAMBA, our method is an online system, which is more applicable for video applications. We further change the backbone feature extractor from ResNet-101 to a stronger one, ResNeXt-101. Our method improves the mAP from 83.3% to 84.3%, which attributes to the more powerful features extractor.

Besides, we compare our method with some SOTA methods on UAVDT. Since most of the recent methods are not releasing their source code, and they are only evaluated on ImageNet VID, we compare our method with FGFA, RDN, and MEGA. As shown in Table II, our method achieves remarkably better results than the above methods. Overall, the same basic architecture results on two datasets demonstrate that our proposed method by aligning features first and employing a multi-level feature aggregation structure exhibits better performance than all above end-to-end models.

Post-processing. In this section, we compare our method with other state-of-the-art methods by further applying post-processing. The results are summarized in Table I. There are three common post-processing techniques, including tubelet re-scoring [56], Seq-NMS [60] and BLR [4], which link detection boxes across frames and use high-scoring object detection from nearby frames to boost scores of weaker detection. Our method employs the Seq-NMS to boost its performance from 83.3% to 84.9% mAP, which still achieves the best detection precision. The +1.6% performance gain on mAP can be ascribed to the additional temporal post-processing. It confirms the effectiveness of propagating the confidence scores among high-related boxes to boost video object detection results.

E. Ablation study

1) Quantitative Analsis: We conduct several ablation studies on the ImageNet VID validation set to evaluate the effectiveness of the proposed method. As shown in Table III, the quantitative results obtained by seven variants of our methods are reported. First, we introduce these variants briefly. Method (a) is the Faster R-CNN with ResNet-101 as the image-based baseline. Method (b) only uses frame-level feature aggregation. Method (c) only uses proposal-level feature aggregation. Method (d) adds the DAlign module into (b). Method (e) adds the DAlign into (c). Method (f) employs the DAlign, frame-level, and proposal-level feature aggregations simultaneously. Method (g) is the complete version of our method.

DAlign module. By comparing the results between Table III(b) and Table III(d) or Table III(c) and Table III(e), we can see that introducing the proposed DAlign into method (b) and method (c) leads to +0.5% gain and +0.6% gain, respectively, which attribute to the DAlign that can model object motion and align the features from frame to frame. The subsequent feature aggregation modules would benefit from aligned features. Besides, we further compare DAlign with FlowNetS [61] and apply these two alignment modules on our method. As shown in Table IV, our method with DAlign

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART END-TO-END VIDEO OBJECT DETECTION MODELS ON IMAGENET VID VALIDATION SET. **BOLD** AND UNDERLINE REPRESENT OPTIMAL AND SUBOPTIMAL RESULTS. THE UP ARROW INDICATED THAT THE LARGER THE VALUE, THE BETTER THE MODEL PERFORMANCE, AND VICE VERSA.

Methods	Backbone	Feature storage	Online	Offline	mAP(%) ↑
FGFA [1]	ResNet-101	Window	✓		76.3
MANet [74]	ResNet-101	Window	✓		78.1
STSN [75]	ResNet-101 + DCN	Window	✓		78.9
OGEMN [79]	ResNet-101 + DCN	Memory		✓	80.0
SELSA [6]	ResNet-101	Window		✓	80.3
MINet [80]	ResNet-101	Window	✓		80.2
RDN [4]	ResNet-101	Window	✓		81.8
MEGA [5]	ResNet-101	Memory		✓	82.9
LSTS [76]	ResNet-101 + DCN	Window	✓		80.1
HVR [77]	ResNet-101	-		✓	83.2
MAMBA [78]	ResNet-101	Memory		✓	**84.6**
Ours	ResNet-101	Window	✓		83.3
RDN	ResNeXt-101	Window	✓		83.2
MEGA	ResNeXt-101	Memory		✓	84.1
MAMBA	ResNeXt-101	Memory		✓	**85.4**
Ours	ResNeXt-101	Window	✓		84.3
FGFA + Seq-NMS	ResNet-101	-		✓	78.4
MANet + Seq-NMS	ResNet-101	-		✓	80.3
STSN + Seq-NMS	ResNet-101 + DCN	-		✓	80.4
SELSA + Seq-NMS	ResNet-101	-		✓	80.5
RDN + BLR	ResNet-101	-		✓	83.8
MEGA + Seq-NMS	ResNet-101	-		✓	84.5
LSTS + Seq-NMS	ResNet-101 + DCN	-		✓	82.1
HVR + Seq-NMS	ResNet-101	-		✓	83.8
Ours + Seq-NMS	ResNet-101	-		✓	**84.9**

TABLE II
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART END-TO-END VIDEO OBJECT DETECTION MODELS ON UAVDT.

Methods	Backbone	mAP(%) ↑
Faster R-CNN [14]	ResNet-101	59.0
FGFA [1]	ResNet-101	59.2
RDN [4]	ResNet-101	60.2
MEGA	ResNet-101	59.6
Ours	ResNet-101	**62.3**

TABLE III
ABLATION STUDIES ON THE IMAGENET VID VALIDATION SET. THE RESULTS ARE OBTAINED BY SEVEN VARIANTS OF OUR METHOD.

Methods	(a)	(b)	(c)	(d)	(e)	(f)	(g)
DAlign				✓	✓	✓	✓
Frame-level		✓		✓		✓	✓
Proposal-level			✓		✓	✓	✓
Instance-level							✓
mAP(%)	75.1	78.0	81.8	78.5	82.4	82.8	**83.3**

TABLE IV
THE UPPER PART OF THE TABLE IS THE EFFICIENCY EVALUATION BETWEEN DALIGN AND FLOWNETS. THE BOTTOM PART IS THE COMPARISON RESULTS OF THESE TWO ALIGNMENT MODULES APPLIED ON OUR PROPOSED METHOD.

Methods	mAP ↑	Params (M) ↓	GFlops ↓
FlowNetS	-	38.680	8.90
DAlign	-	**1.956**	**0.64**
Ours-F	83.1	92.751	260.71
Ours-D	**83.3**	**65.027**	**252.45**

TABLE V
ABLATION STUDY ON EACH COMPONENT OF OUR METHOD WHEN APPLIED ON MEGA. THE RESULTS SHOWS FOUR VARIANTS OF THE MEGA.

Methods	(h)	(i)	(j)	(k)	(l)
DAlign		✓			✓
Frame-level			✓		✓
Instance-level				✓	✓
mAP(%)	82.9	83.1	83.1	83.3	83.6

(Ours-D) achieves higher mAP and lower model complexity than it with FlowNetS (Ours-F), which indicates that DAlign is more effective and computationally friendly.

Frame-level feature aggregation. Method (b) obtains a 78.0% in terms of mAP, 2.9% higher than the image-based detector Faster R-CNN. The reason is that the proposed feature aggregation is capable of fusing the frame-level information across space and time to enhance the reference frame feature.

Proposal-level feature aggregation. Table III(c) reports the results obtained by applying the proposal-level feature aggregation to the baseline detector. Compared with the baseline, the proposal-level feature aggregation achieves a significant +6.7% gain. This huge improvement demonstrates that the proposal-level feature aggregation could enhance each proposal in the reference frame with the most informative appearance from multiple proposals in the support frames. Besides, method (c) performs better than method (b). This result indicates that the proposal-level feature aggregation focuses on the proposal features that contain balance foreground and background in-

TABLE VI
PERFORMANCE AND RUN TIME COMPARISONS BY USING DIFFERENT SAMPLING NUMBER K IN INSTANCE-LEVEL FEATURE AGGREGATION.

K	10	15	20	25	30
mAP(%)	82.5	83.0	**83.3**	83.2	82.9
runtime(ms)	99.6	102.6	105.7	109.3	112.0

TABLE VII
PERFORMANCE AND RUN TIME COMPARISONS BY USING DIFFERENT SAMPLED RATIO $r\%$ IN INSTANCE-LEVEL FEATURE AGGREGATION.

r(%)	10	20	30	40	50
mAP(%)	83.1	**83.3**	82.9	82.7	82.7
runtime(ms)	101.4	105.7	109.5	113.2	117.3

formation while the frame-level feature aggregation neglects the interaction among crucial foreground regions.

Instance-level feature aggregation. Comparing with the Table III(f) and Table III(g), the instance-level feature aggregation improves the performance from 82.8% to 83.3%, with 0.5% improvement, which reflects that the instance-level feature aggregation further updates the object proposal features by support proposal features that belong to the same object. Such an operation filters out the irrelevant information and pays attention to each object proposal for feature aggregation.

The result of our pipeline is presented in Table III(f), which yields the best performance, achieving +8.2% gain compared with baseline. It demonstrates that the frame-level, proposal-level, and instance-level feature aggregation are complementary. The former is responsible for enhancing the frame features. The midden distills the relation from support proposals to strengthen the reference proposal features. The latter further strengthens the proposal features that are related to the foreground by using instance ID. Besides, DAlign is also indispensable to align features from frame to frame. Thus, our method exploits rich spatial-temporal information in videos, making the detector robust against object appearance variations, such as motion blur and occlusion. Overall, the results in Table III verify the effectiveness of each module and prove that the combination of all modules in a unified framework can better deal with the deteriorated video quality.

Effect of proposed components on a strong baseline. We further apply our proposed components on a more stronger baseline MEGA to verify the effectiveness of each module. Specifically, Method (h) is the baseline. The original MEGA is inherently equipped with proposal-level feature aggregation. Method (i) adds DAlign module into (h). Method (j) adds Frame-level feature aggregation into (h). Method (k) adds Instance-level feature aggregation into (h). Method (l) uses all components of our method. As shown in Table V, by introducing our proposed components, the performance of the MEGA is improved significantly, which demonstrates that the combination of memory bank structure and our proposed components could cover more frames and sufficiently mine critical cues for feature aggregation. Besides, the instance-level feature aggregation gains the largest improvement from 82.9% to 83.3%, which is consistent with the ablation study in Table III.

TABLE VIII
PERFORMANCE EVALUATION ON IMAGENET VID DATASET WITH AND WITHOUT APPEARANCE DETERIORATION. AD IS ABBREVIATED FROM APPEARANCE DETERIORATION.

Methods	with AD ↑	w/o AD ↑
Faster R-CNN	75.1	88.3
Ours	**83.3**	**89.6**

Effect of sampling number K and sampling ratio $r\%$ in instance-level feature aggregation. We firstly vary K from 10 to 30 to explore the relationship between the performance/run time and the sampling top K proposals. As shown in Table VI, the run time at inference is gradually increased when enlarging the sampling number. The mAP rises first and then falls gradually as the sampling number increases, and the best performance is attained when the sampling number K is 20. Next, to investigate the effect of sampling ratio $r\%$, we further compare the results of performance and run time by varying the sampling ratio from 10% to 50% in Table VII. The performance is slightly affected by the change of sampling ratio r. Besides, the run time significantly increases with the sampling ratio increases. Therefore, we set the sampling number K to 20 and the sampling ratio r to 20% experimentally.

Whether to update the support frame features in frame-level feature aggregation. The typical way to conduct frame-level feature aggregation is combining the support features to update the reference feature while the support features stay the same. We argue that the high quality of support features would benefit the frame-level feature aggregation. To verify our suppose, we design two variations: the query is all input frames, and the query is reference frame only. As shown in Table IX, the query of all input frames boosts up the mAP from 83.0% to 83.3% compared with the query of reference frames only. This improvement indicates that the quality of the supporting frame features affects the video object detection.

Effect of the degraded frames in the dataset. As shown in Table VIII, by excluding the low-quality frames from the validation set, Faster R-CNN gains +13.2% improvement on mAP while our method achieves +6.3%. The relatively low improvement illustrates that the proposed MST has corrected some false detection results in degraded frames. Besides, from columns 2 and 3, MST improves the performance by +8.2% mAP in the original dataset while only +1.3% in the manipulated dataset, which indicates that MST plays a critical role in dealing with appearance deterioration.

2) Qualitative Analysis: First, we select two challenge video sequences in the ImageNet VID validation set. One suffers from significant pose variations, and the other has serious motion blur. As shown in Figure 5, we represent the results of four variants, baseline, baseline + DAlign + frame-level, baseline + DAlign + frame-level + proposal-level, and the whole framework, respectively. The show threshold is set to 0.6. It is obvious that our method shows better detection results than baseline. For example, the lizard is presented in a rare pose for quite a while. The baseline fails to detect it in the second and third frames. In contrast, our proposed

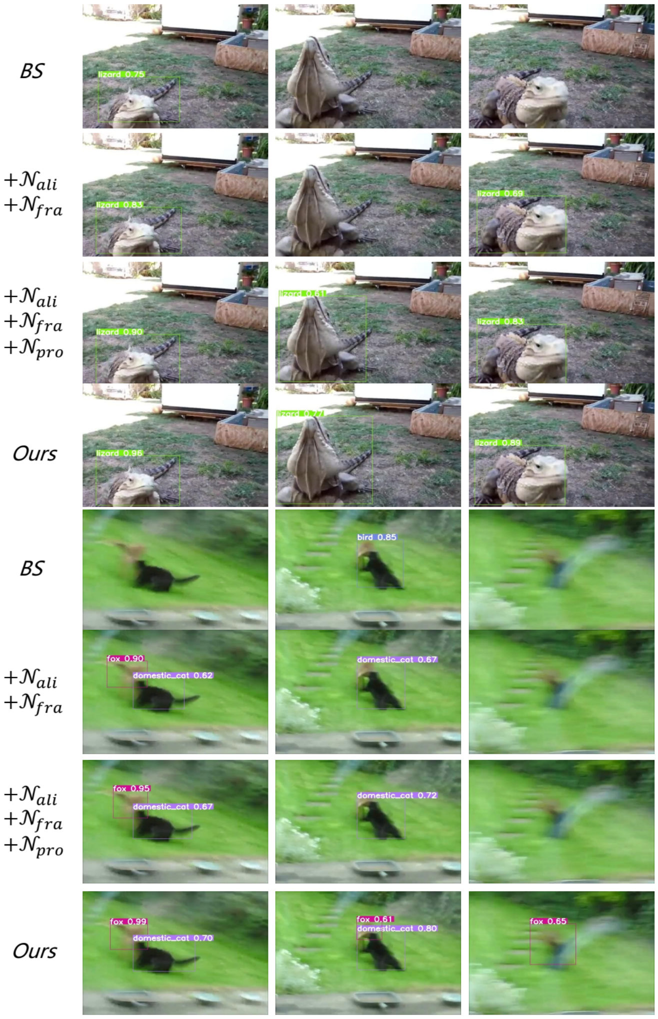

BS
$+\mathcal{N}_{ali}$ $+\mathcal{N}_{fra}$
$+\mathcal{N}_{ali}$ $+\mathcal{N}_{fra}$ $+\mathcal{N}_{pro}$
Ours
BS
$+\mathcal{N}_{ali}$ $+\mathcal{N}_{fra}$
$+\mathcal{N}_{ali}$ $+\mathcal{N}_{fra}$ $+\mathcal{N}_{pro}$
Ours

Fig. 5. The detection results of two video sequences for qualitative comparison. Please zoom in for a more clear comparison.

method could successfully detect objects with a high score in each frame. Besides, we also observe that the frame-level feature aggregation gains better performance than baseline but still fails to detect the frame in column 2. After adding the proposal-level feature aggregation, every frame could be detected correctly. The instance-level feature aggregation further improves the object scores. Benefit from the multi-level feature aggregation, our method could provide more robust results, consistent with the results in the quantitative experiment.

F. Failure Cases

We show some failure cases in Figure 6. The first row gives a wrong class label that predicts the squirrel as a monkey. The second predicts the motorcycle as a car. The third fails to produce a prediction and detects duplicated either. Although our method adopts multi-level feature aggregation, it is still difficult to detect the correct results in some extreme cases. Through analyzing the results of the entire video sequences, the reason may lie in the limited temporal spanning range.

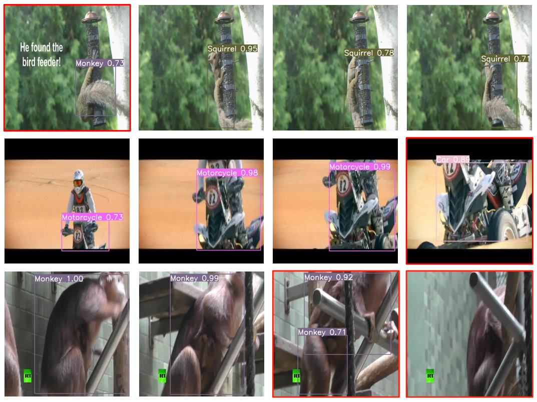


Fig. 6. Some failure cases caused by rare pose and occlusion. We mark the failure frames with red contour.

TABLE IX
PERFORMANCE COMPARISON BY TWO USAGES OF THE FRAME-LEVEL FEATURE AGGREGATION, THE DIFFERENCE BETWEEN THOSE ARE THE QUERY FEATURE OF FORMER ARE ALL INPUT FRAMES, AND THE LATTER IS REFERENCE FRAME ONLY.

Methods	Query-ref	Query-all
mAP(%)	83.0	**83.3**

V. Conclusion and Future Work

In this paper, we present a novel video detection method, which consists of a *Deformable Feature Alignment* (DAlign) module and a *Multi-level Spatial-Temporal* (MST) feature aggregation module. In particular, DAlign models object motion and aligns the features from frame to frame. Then, MST sufficiently exploits the spatial-temporal information at the frame level, proposal level, and instance level in a unified framework to generate enhanced features. Specifically, given the aligned features, the frame-level feature aggregation is designed to distill informative appearance from all frames to augment reference and support features. We devise the proposal-level feature aggregation after the RPN. Each proposal pair measures the similarity based on the appearance and geometric representation. The reference proposal features are enhanced by aggregating the support proposal features according to the proposal pair similarities. Furthermore, an instance-level aggregation is followed to enhance each object proposal feature by corresponding support object proposal features that belong to the same object. Finally, the upgraded proposal features are input into the detection head and perform bounding box classification and regression. Extensive experiments conducted on the ImageNet VID dataset and UAVDT demonstrate the superiority of our method. In future work, we want to explore a method that could run in real time in edge devices. Such a method is equipped with a light-weight structure while still maintaining powerful relation modeling ability. We hope our method will play a critical role in the field of autonomous driving.

References

[1] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, “Flow-guided feature aggregation for video object detection,” in *Proceedings of the IEEE*

International Conference on Computer Vision, 2017, pp. 408–417.

[2] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2349–2358.

[3] X. Zhu, J. Dai, L. Yuan, and Y. Wei, "Towards high performance video object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7210–7218.

[4] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, and T. Mei, "Relation distillation networks for video object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7023–7032.

[5] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global-local aggregation for video object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 337–10 346.

[6] H. Wu, Y. Chen, N. Wang, and Z. Zhang, "Sequence level semantics aggregation for video object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9217–9225.

[7] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, and T. Mei, "Single shot video object detector," *IEEE Transactions on Multimedia*, 2020.

[8] Y. Huang, Q. Jiang, and Y. Qian, "A novel method for video moving object detection using improved independent component analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 6, pp. 2217–2230, 2020.

[9] Y. Cao, Q. Tang, X. Wu, and X. Lu, "Effnet: Enhanced feature foreground network for video smoke source prediction and detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[10] N. Wang, W. Zhou, and H. Li, "Reliable re-detection for long-term tracking," *IEEE transactions on circuits and systems for video technology*, vol. 29, no. 3, pp. 730–743, 2018.

[11] L. Han, P. Wang, Z. Yin, F. Wang, and H. Li, "Class-aware feature aggregation network for video object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[12] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

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

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[15] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[16] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *arXiv preprint arXiv:1605.06409*, 2016.

[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[18] 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*, 2017, pp. 2980–2988.

[19] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4203–4212.

[20] 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*, 2018, pp. 3588–3597.

[21] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[28] C. Guo, B. Fan, J. Gu, Q. Zhang, S. Xiang, V. Prinet, and C. Pan, "Progressive sparse local attention for video object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3909–3918.

[29] F. Xiao and Y. Jae Lee, "Video object detection with an aligned spatial-temporal memory," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 485–501.

[30] M. Shvets, W. Liu, and A. C. Berg, "Leveraging long-range temporal relationships between proposals for video object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9756–9764.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[34] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[35] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[36] P. Hu and D. Ramanan, "Finding tiny faces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 951–959.

[37] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 797–813.

[38] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1758–1770, 2019.

[39] Z. Shao, L. Wang, Z. Wang, W. Du, and W. Wu, "Saliency-aware convolution neural network for ship detection in surveillance video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 781–794, 2019.

[40] C.-B. Wu, L.-H. Wang, and K.-C. Wang, "Ultra-low complexity block-based lane detection and departure warning system," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 582–593, 2018.

[41] T. Zhang and X. Zhang, "Shipdenet-20: An only 20 convolution layers and¡ 1-mb lightweight sar ship detector," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 7, pp. 1234–1238, 2020.

[42] ——, "A polarization fusion network with geometric feature embedding for sar ship classification," *Pattern Recognition*, vol. 123, p. 108365, 2022.

[43] T. Zhang, X. Zhang, X. Ke, C. Liu, X. Xu, X. Zhan, C. Wang, I. Ahmad, Y. Zhou, D. Pan *et al.*, "Hog-shipclsnet: A novel deep learning network with hog feature fusion for sar ship classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–22, 2021.

[44] T. Zhang, X. Zhang, and X. Ke, "Quad-fpn: A novel quad feature pyramid network for sar ship detection," *Remote Sensing*, vol. 13, no. 14, p. 2771, 2021.

[45] T. Zhang and X. Zhang, "Squeeze-and-excitation laplacian pyramid network with dual-polarization feature fusion for ship classification in sar images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[46] T. Zhang, X. Zhang, J. Shi, S. Wei, J. Wang, J. Li, H. Su, and Y. Zhou, "Balance scene learning mechanism for offshore and inshore ship detection in sar images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2020.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[48] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8577–8584.

[49] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[50] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 850–859.

[51] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 840–849.

[52] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 9627–9636.

[53] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi, "Foveabox: Beyond anchor-based object detector," *arXiv preprint arXiv:1904.03797*, 2019.

[54] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[55] K. Kang, W. Ouyang, H. Li, and X. Wang, "Object detection from video tubelets with convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 817–825.

[56] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang *et al.*, "T-cnn: Tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, 2017.

[57] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Detect to track and track to detect," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3038–3046.

[58] K. Chen, J. Wang, S. Yang, X. Zhang, Y. Xiong, C. Change Loy, and D. Lin, "Optimizing video object detection via a scale-time lattice," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7814–7823.

[59] C.-H. Yao, C. Fang, X. Shen, Y. Wan, and M.-H. Yang, "Video object detection via object-level temporal aggregation," in *European Conference on Computer Vision*. Springer, 2020, pp. 160–177.

[60] W. Han, P. Khorrami, T. L. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. S. Huang, "Seq-nms for video object detection," *arXiv preprint arXiv:1602.08465*, 2016.

[61] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[62] D. Zhang, J. Han, L. Yang, and D. Xu, "Spftn: A joint learning framework for localizing and segmenting objects in weakly labeled videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 475–489, 2018.

[63] P. Huang, J. Han, N. Liu, J. Ren, and D. Zhang, "Scribble-supervised video object segmentation," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 2, pp. 339–353, 2021.

[64] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang, "Reinforcement cutting-agent learning for video object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9080–9089.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[66] 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*, 2018, pp. 7794–7803.

[67] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612.

[68] L. Lin, H. Chen, H. Zhang, J. Liang, Y. Li, Y. Shan, and H. Wang, "Dual semantic fusion network for video object detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1855–1863.

[69] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[70] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 370–386.

[71] K. Kang, H. Li, T. Xiao, W. Ouyang, J. Yan, X. Liu, and X. Wang, "Object detection in videos with tubelet proposal networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 727–735.

[72] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[73] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[74] S. Wang, Y. Zhou, J. Yan, and Z. Deng, "Fully motion-aware network for video object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 542–557.

[75] G. Bertasius, L. Torresani, and J. Shi, "Object detection in video with spatiotemporal sampling networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 331–346.

[76] Z. Jiang, Y. Liu, C. Yang, J. Liu, P. Gao, Q. Zhang, S. Xiang, and C. Pan, "Learning where to focus for efficient video object detection," in *European conference on computer vision*. Springer, 2020, pp. 18–34.

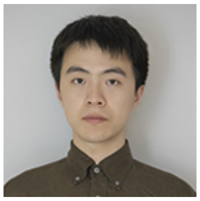
[77] M. Han, Y. Wang, X. Chang, and Y. Qiao, "Mining inter-video proposal relations for video object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 431–446.

[78] G. Sun, Y. Hua, G. Hu, and N. Robertson, "Mamba: Multi-level aggregation via memory bank for video object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2620–2627.

[79] H. Deng, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan, "Object guided external memory network for video object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6678–6687.

[80] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, and T. Mei, "Minet: Meta-learning instance identifiers for video object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 6879–6891, 2021.

Chao Xu received the B.S. degree in electrical engineering and its automation from Nanchang University, Nanchang, China, in 2018. He is currently working toward the doctor degree in control science and engineering with the School of Control Science and Engineering, Zhejiang University, Hangzhou, China. His major research interests include computer vision and deep learning.



Jiangning Zhang received the B.S. degree from Wuhan University, Wuhan, China, in 2017. He is currently working toward the doctor degree in control science and engineering with the School of Control Science and Engineering, Zhejiang University, Hangzhou, China. His major research interests include low-level computer vision, generative adversarial network, and neural architecture design.

Mengmeng Wang received the B.S. degree and M.S. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2015 and 2018. She is currently working toward the doctor degree in control science and engineering with the School of Control Science and Engineering, Zhejiang University, Hangzhou, China. Her major research interests include computer vision and deep learning.

Guanzhong Tian (Member, IEEE) received the B.S. degree from Harbin Institute of Technology, Harbin, China, in 2010, and the Ph.D. degree in from Zhejiang University, Hangzhou, China, in 2021. He is currently a Research associate with Ningbo Research Institute, Zhejiang University. His research interests include computer vision, model compression, embedded AI.

Yong Liu received the B.S. degree in computer science and engineering and the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2001 and 2007, respectively. He is currently a Professor with the Department of Control Science and Engineering, Zhejiang University. His latest research interests include machine learning, robotics vision, information processing, and granular computing.