Exploiting Better Feature Aggregation for Video Object Detection

Liang Han*
Stony Brook University
liahan@cs.stonybrook.edu

Pichao Wang*
Alibaba Group
pichao.wang@alibaba-inc.com

Zhaozheng Yin†
Stony Brook University
zyin@cs.stonybrook.edu

Fan Wang
Alibaba Group
fan.w@alibaba-inc.com

Hao Li
Alibaba Group
lihao.lh@alibaba-inc.com

ABSTRACT

Video object detection (VOD) has been a rising topic in recent years due to the challenges such as occlusion, motion blur, etc. To deal with these challenges, feature aggregation from local or global support frames is verified effective. To exploit better feature aggregation, in this paper, we propose two improvements over previous works: a **class-constrained spatial-temporal relation network** and a **correlation-based feature alignment module**. For the class constrained spatial-temporal relation network, it operates on object region proposals, and learns two kinds of relations: (1) the dependencies among region proposals of the same object class from support frames sampled in a long time range or even the whole sequence, and (2) spatial relations among proposals of different objects in the target frame. The homogeneity constraint in spatial-temporal relation network not only filters out many defective proposals but also implicitly embeds the traditional post-processing strategies (e.g., Seq-NMS [15]), leading to a unified end-to-end training networks. In the feature alignment module, we propose a correlation based feature alignment method to align the support and target frames for feature aggregation in the temporal domain. Our experiments show that the proposed method improves the accuracy of single-frame detectors significantly, and outperforms previous temporal or spatial relation networks. Without bells or whistles, the proposed method achieves state-of-the-art performance on the ImageNet VID dataset (84.80% with ResNet-101) without any post-processing methods.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**.

KEYWORDS

video object detection; relation network; class constraint; feature alignment

ACM Reference Format:
Liang Han, Pichao Wang, Zhaozheng Yin, Fan Wang, and Hao Li. 2020. Exploiting Better Feature Aggregation for Video Object Detection. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413927

1 INTRODUCTION

Object detection in still images has achieved significant progress due to the deep convolutional networks. Strong backbones [18, 42], advanced detection paradigms [4, 12, 16, 28, 30, 33], and large scale datasets [23, 25] jointly push object detection forward to the limit. However, when directly applying those image-based object detectors on a frame-by-frame basis to a video-based object detection task, the performance is often unsatisfactory due to the deteriorated appearance caused by occlusion, motion blur, out-of-focus camera, and rare poses in video capturing, etc.

An intuitive way to deal with these challenges in video object detection (VOD) is to leverage the temporal information inherently encoded in videos, which is absent in still images, to enhance the object features in the current detecting frame. Decoupled from the training stage, several post-processing methods [15, 21, 22] have been proposed. These methods explore bounding box association rules across nearby frames to refine the per-frame detection results, considering the spatial and temporal coherence in videos. Unfortunately, those post-processing methods are not jointly optimized with the proposed detection networks and do not benefit from the temporal information in the phase of training.

Recently, to overcome the sub-optimal problem of post-processing, several end-to-end feature aggregation methods propose to exploit the temporal information encoded in nearby frames to help the object detection in the current frame. For instance, optical flow [38, 44] and feature correlation [9] are used to capture the motion information, which helps to calibrate and aggregate features from nearby frames onto the current frame. Note that these methods can only exploit features in local frames to perform feature aggregation, which inevitably ignores the supportive information contained in faraway frames.

*Both authors contributed equally to this work.
†Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413927

To break this limitation, relation network is proposed to explore the long-term dependencies among video frames for feature aggregation, and dominates the research direction of VOD. Relation network is able to aggregate supportive object information from non-local frames to objects in the current frame. Current relation based feature aggregation works [6, 36, 40] usually adopt a two-stage detection strategy, i.e., first generating object proposals for target and support frames, and then performing feature aggregation for target object proposals with the generated object proposals in the support frames by weighted adding the support proposal features, where the adding weights are obtained by computing similarities between the target and support proposals.

There are three problems for most of the current relation-based feature aggregation methods of VOD. Firstly, they only consider the temporal dependencies among the objects, and neglect the spatial relations, which has been proved to be very useful in still image detection [20]; secondly, all of these methods directly aggregate all the proposals from support frames in the temporal domain, without consideration of whether they belong to the same class or not, making it inevitably bring defective proposals from irrelevant classes; thirdly, they aggregate the features from both support and target proposals directly without feature alignment, leading to unaligned feature for the following regression and classification. To exploit better feature aggregation, we propose a class-constrained spatial-temporal feature aggregation network and a correlation based feature alignment module. In the class constrained spatial-temporal feature aggregation network, we first encode the information both in spatial and temporal domains; furthermore, by modifying the RPN network, a class homogeneity constraint is introduced, by only leveraging the proposals with the same (predicted) class label to enhance target proposal feature in the temporal feature aggregation. With these improvements, we are able to learn the dependencies among region proposals from support frames of the same object in a long (even sequence-level) temporal duration, and also learn spatial relations among other proposals of different objects in the same target frame. Lastly, in the feature alignment module, we propose a correlation based feature alignment method for better feature aggregation in the temporal domain. It aligns the support and target proposals which may have quite different poses,shapes, etc., making it more suitable for the following regression and classification step.

The main contributions of this work are summarized as follows:

- We treat video object detection as a spatial-temporal proposal relation mining process, by taking consideration of both appearance features in temporal domain and topological relations in spatial domain.
- The proposed class homogeneity constraint will only bring in the most related object proposals in temporal domain, to reduce the number of defective region proposals in feature aggregation, and filter out invalid information from other classes to get more accurate aggregated feature.
- The proposed correlation based feature alignment module further improves the quality of feature aggregation.
- The proposed methods are evaluated on the large scale ImageNet VID dataset and achieved state-of-the-art results with mAP at 84.80% (with ResNet-101), without any post-processing methods.

2 RELATED WORK

2.1 Object Detection in Still Images

With the development of deep learning, state-of-the-art methods for still image object detection has achieved great progress. Generally, the still image object detection approach could be divided into two paradigms, two stage and single stage.

Two stage detectors usually generate region proposals first, and then the proposals are refined by the classification and regression through the Regions with Convolutional Neural Networks (R-CNN) stage [13]. To speedup, ROI pooling is introduced in SPP-Net [17] and Fast R-CNN [12]. Region Proposal Network (RPN) is proposed in Faster R-CNN [33] to generate region proposals. R-FCN [4] replaces the ROI pooling on the intermediate feature maps with position-sensitivity ROI pooling on the final score maps. Feature Pyramid Networks (FPN) [24] brings an inherent multiscale, pyramidal hierarchy of deep convolution networks to build feature pyramids. Mask RCNN [16] proposes the ROI align operation to replace ROI pooling to further improve the detection accuracy.

In contrast, one stage object detector directly predicts the bounding box of interest based on the extracted feature map from CNN, without the region proposal generation step, thus, it is usually faster than two-stage counterpart. YOLO [30], YOLO9000 [31], YOLOV3 [32], SSD [28] and DSSD [10] are representative works.

Recently, relation networks [20] and non-local networks [39] are proposed to explore the appearance and geometry relations among object proposals within a still image. The topological relations of objects are captured in these detectors and the performance is improved. Our work borrows the idea from this line of work, and additionally models the dependencies among region proposals in the temporal domain.

2.2 Object Detection in Videos

Most existing methods for video object detection leverages temporal information as additional constraint, and this line of research mainly falls into two categories. The first is focused on post processing [15, 21, 22] that explores bounding box association rules across nearby frame to refine the per-frame detection results, taking the spatial and temporal coherence into consideration. Those methods are sub-optimal because they are highly dependent on the quality of initial detector which is trained without any temporal information. As contrast, the other category of methods [1, 2, 5, 6, 9, 14, 36, 38, 40, 41, 43–45] takes advantages of the temporal information in videos during training stage. Optical flow [8] is widely used in feature aggregation and warping [38, 44, 45]. However, the optical flow can only exploit the temporal consistence in a local window and it does not work well in the case of occlusion. To explore longer dependencies in the temporal domain, Xiao and Lee [41] adopt the LSTM and propose a spatial-temporal memory networks (STMN) to model long-term temporal appearance and motion dynamics. Thanks to the relation networks [37], Shvets et al. [36] and Wu et al. [40] explore the temporal relation in a long range or across the whole sequence. Our proposed method not only models the long-term dependencies in temporal domain, but also explores the geometric relations in the spatial domain. Deng et al. [6] propose the RDN to model the spatial-temporal relations for VOD. However, they do not constrain the dependencies among the same object

class in different frames, and relations among different objects in the same frame, thus to improve the speed and accuracy, they still largely depend on the distillation network and post-processing. In contrast, we explicitly include the class homogeneity constrain, so that it not only largely reduces the number of region proposals involved to compute relations, but also makes a unified end-to-end framework without any post-processing step.

Another category of VOD is to use temporal information to reduce the computation cost, which is beyond the scope of this paper. For example, Zhu et al. [43, 45] adopt optical flow to propagate the features of key frames to other frames to save the expensive feature extraction cost. Chen et al. [2] design a time-scale lattice to improve the speed with an extra classifier to re-score the bounding boxes. Liu et al. [26, 27] adopt Bottleneck-LSTM with MobileNet [19, 35] as backbone and SSD as detector to improve the speed on the mobile devices. Guo et al. [14] introduce progressive sparse local attention (PSLA) to replace optical flow for fast computation.

3 PROPOSED METHOD

In this section, we present the details of how we tackle the problem of video object detection by exploring both the temporal and spatial information of the video objects under the class homogeneity constrain, and how we design the feature alignment module.

3.1 Overview

As mentioned in section 1, video object detection is challenging especially for those frames with object appearance degradation caused by occlusion, motion blur, and camera defocus, etc. Therefore, an intuitive solution is to aggregate features of similar object from other frames to enhance the object appearance feature in the current frame. Some previous works exploit optical flow to aggregate object features across frames on position level, however, the estimation of optical flow is still unreliable because of the above-mentioned degraded object appearance and dramatically changing pose. Therefore, instead of aggregating features on pixel location level, we propose to perform feature aggregation on object level by first generating object proposals for each frame with a Region Proposal Network (RPN).

Let the frame we are currently working on to generate detection results as a ***target frame***, the frames that are used to extract additional appearance features for aggregation be ***support frames***, a proposed object proposal in the target frame be ***target proposal***, and the proposals in the support frames be ***support proposals***.

Figure 1 illustrates the idea of the proposed algorithm. First, features are extracted for each frame by a backbone (e.g., ResNet-101), and a RPN is adopted to generate object proposals with the extracted features. Then, a simple classifier is designed to predict the class label for each object proposal. After that, the temporal aggregation module (TAM) is designed to enhance target object proposal features by aggregating object proposal features with the same class label from support frames. Finally, the spatial relation module (SRM) is designed to model object topological relations by analyzing the interactions among objects in the same frame, and further aggregate features for the target proposal with the object proposals in the same target frame. Feature alignment module is plugged in the temporal aggregation module for better feature aggregation. With the aggregated feature, object classification and location regression can be performed for the target object proposal.

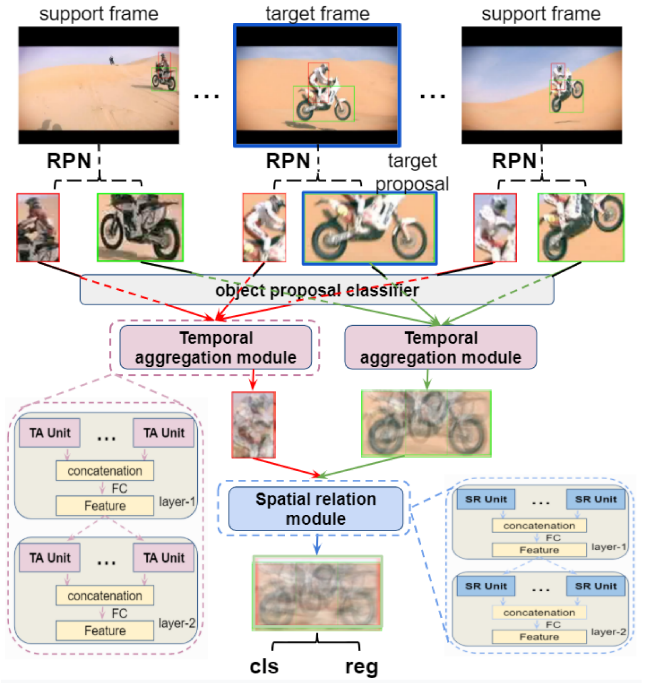

Figure 1: Flowchart of the proposed video object detection method. Feature aggregation is performed on proposal level. First, region proposal network (RPN) is used to obtain proposals for each frame. Then, the temporal aggregation module (TAM) aggregates features for proposals in the target frame by leveraging proposals coming from the same class in support frames. Feature alignment is preformed for support proposals before aggregating features in the TAM. Finally, spatial relation module (SRM) further enhances features for target proposals by exploring the spatial relation of objects in the target frame. The multi-head, multi-layer architecture of TAM and SRM are zoomed in (purple dashed box and blue dashed box, respectively).

3.2 Temporal aggregation module

The goal of the temporal aggregation module is to enhance the appearance feature of the target proposal with the ones of the support proposals from the support frames, so the most informative object appearance features should be picked out for feature enhancement.

Similar to [11], we design a multi-head and multi-layer architecture for the temporal aggregation module. There are two layers in our module and each layer consists of H (H is 16 in our experiments) Temporal Aggregation Unit (TA Unit). Figure 2 shows the architecture of the designed TA Unit. Transformer mechanism [37] is adopted in the temporal aggregation module to pick the most informative support proposals for feature aggregation of the target proposal. For a target proposal p^t, let $P^s = \{p_1^s, p_1^s, ..., p_n^s\}$ be the support proposal set. The appearance feature vectors of the target

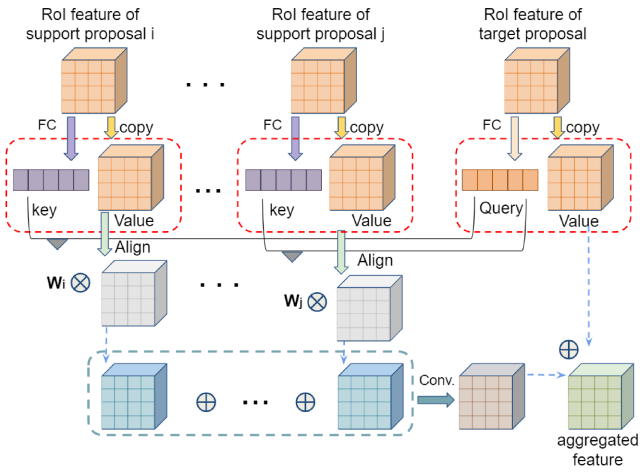

Figure 2: Architecture of the temporal aggregation unit (TA Unit). The target proposal feature is projected to the query by a fully connection (FC), and the support proposals are projected to the keys by another FC. Values are obtained by directly copy the originally extracted proposal feature, and feature alignment (Align) is performed for support values based on the target value. Feature aggregation is achieved by the weighted sum of Values, while the weights come from the similarities between the query and the corresponding keys.

proposal and the support proposal extracted by a trunk are $\mathbf{X}^t$ and $\mathbf{X}_i^s$ $(i = 1, 2, ..., n)$, respectively. In each transform unit, two fully connections ϕ and ψ are first applied on target and support proposal features to obtain the query $\phi(\mathbf{X}^t)$ and keys $\psi(\mathbf{X}_i^s)$. In our scenario, the target proposal feature is projected to the query, and the support proposals are projected to the keys. Then, for each proposal pair (p^t, p_i^s), the appearance similarity between this proposal pair is measured with the cosine similarity of the query and the key:

$$w_i^a = cos(\phi(\mathbf{X}^t), \psi(\mathbf{X}_i^s)) = \frac{\phi(\mathbf{X}^t) * \psi(\mathbf{X}_i^s)}{||\phi(\mathbf{X}^t|| \cdot ||\psi(\mathbf{X}_i^s)||}, \tag{1}$$

where cos denotes the cosine similarity of two vectors, $*$ is the dot product of two vectors, and $||\cdot||$ means the norm of a vector. Considering that our goal is to enhance the appearance feature of the target proposal with the ones of the support proposals, instead of projecting the original appearance feature with a linear projection, here we take the original appearance feature values of the support proposals as the support features ("value" in Figure 2), but feature alignment is performed on the support features by a proposed feature alignment module (this module will be introduced in details in the following subsection 3.5) to align the features of the support proposals to the ones of the target proposal. After calculating the similarity of the proposal pair and obtaining the aligned support feature value, the feature aggregation is performed as a weighted summation of the aligned support feature values with the proposal pair similarities as summation weights. To preserve the magnitude of the aggregated feature, a softmax function is applied across all the support proposals to normalize the calculated similarities. Mathematically, the aggregated appearance feature is represented as:

$$\begin{aligned} A(\mathbf{X}^t) &= \sum_{i=1}^{n} softmax(w_i^a) \cdot Align(\mathbf{X}_i^s) \\ &= \sum_{i=1}^{n} \frac{exp(\frac{\phi(\mathbf{X}^t)^T \psi(\mathbf{X}_i^s)}{\sqrt{D_1}})}{\sum_{j=1}^{n} exp(\frac{\phi(\mathbf{X}^t)^T \psi(\mathbf{X}_j^s)}{\sqrt{D_1}})} \cdot Align(\mathbf{X}_i^s). \end{aligned} \tag{2}$$

where $Align(\mathbf{X}_i^s)$ is the aligned feature of the original support feature $\mathbf{X}_i^s$ by our proposed feature alignment module, $A(\mathbf{X}^t)$ is the aggregated feature for the target proposal. Borrowing the idea of ResNet, here we apply a linear projection (LP) on the aggregated feature and add it to the original appearance feature of the target proposal to get the final aggregated appearance feature of the target proposal $A^{'}(\mathbf{X}^t)$:

$$A^{'}(\mathbf{X}^t) = \mathbf{X}^t + LP(A(\mathbf{X}^t)). \tag{3}$$

To keep the temporal aggregation module in-place (i.e., input and output with the same feature dimension), at the very beginning of each layer, we first use a 1×1 convolutional layer *conv* to reduce the input feature dimension to $\frac{1}{H}$ of the original dimension before feeding it into each TA Unit. The outputs of each TA Unit are then concatenated together to get back the original dimension, and the concatenated feature is then fed into the next layer. This operation is performed in each layer of the temporal aggregation module.

3.3 Spatial relation module

With the feature-aggregated target frame, it seems that object detection now can be done by applying a base detector (e.g., Faster R-CNN, R-FCN, etc.) on each target proposal individually. However, it has been well believed in computer vision community that relation between objects can help object recognition [3, 7, 20, 29]. Therefore, we introduce a spatial relation module to further explore the spatial topological relation of objects by embedding the additional position and shape information of the proposals besides its appearance feature to facilitate the video object detection.

The spatial relation module shares the same architecture with the temporal aggregation module (i.e., multi-head and multi-layer). The transformer mechanism is also adopted in the spatial relation module. However, the Spatial Relation Unit (SR Unit) in this module is an extension of the TA Unit in the temporal aggregation module. Besides capturing the appearance similarity between a proposal pair with a appearance similarity weight w_i^a as we do in the TAM, a geometric weight w_i^g is also calculated to capture the topology relation between object proposals in the same target frame by using the shape and location information of the proposals:

$$w_i^g = \frac{\varphi(\mathbf{G}^t)^T \varphi(\mathbf{G}_i^s)}{\sqrt{D_2}}, \tag{4}$$

where φ is a position embedding operation; D_2 is the dimension of the geometry features obtained by the position embedding, $\mathbf{G}^t$ and $\mathbf{G}_i^s$ are the geometry information of the target proposal and support proposal i, respectively, which are defined as

$$\begin{aligned} \mathbf{G}^t &= (x^t, y^t, h^t, w^t) \\ \mathbf{G}_i^s &= (x_i^s, y_i^s, h_i^s, w_i^s), \end{aligned} \tag{5}$$

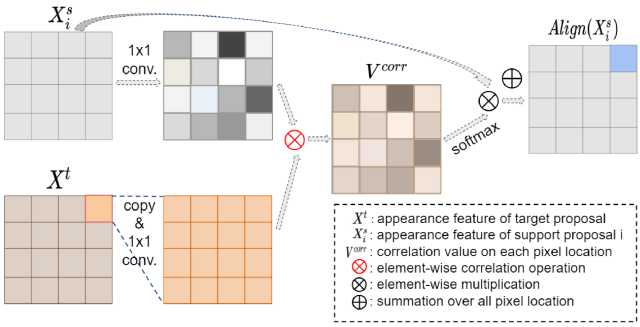

Figure 3: Idea of the proposed feature alignment module.

where x^t and y^t are the location of the target proposal bounding box center, h^t and w^t represent the height and width of the target proposal bounding box, respectively. Symbols mean the same for support proposal i. These location and shape information of the proposal bounding box can be obtained from the RPN. The new geometric weight w^g is designed to model the spatial relation of objects and only consider the relative geometric relationship between objects, which can guarantee that the SR Unit is invariant to scale transformation.

The final similarity w_i between the target proposal p^t and the support proposal p_i^s is computed by combining the geometric similarity w_i^g with the original appearance similarity weight w_i^a

$$w_i = \frac{w_i^g \cdot exp(w_i^a)}{\sum_{i=1}^{n} w_i^g \cdot exp(w_i^a)} \tag{6}$$

which enables the SR Unit in the spatial relation module to capture both the spatial relation of objects and the appearance similarity of objects in the same frame. Note that the spatial feature aggregation is only performed in the target frame.

3.4 Feature alignment module

When performing feature aggregation for target proposal with features of support proposals, it is highly possible that the objects in the target proposal and support proposal have quite different poses, shapes, etc., which makes the appearance features of these two proposals misaligned, and further degrades the feature aggregation. Therefore, appearance feature alignment is crucial for better feature aggregation. Different from FGFA [44] and MANet [38] which adopt FlowNet [8] to align feature, and STMM [41] which adopts a local "MatchTrans" for feature alignment, we design a feature alignment module which tries to align the support proposal feature to the target one globally.

Figure 3 depicts the idea of the designed feature alignment module. For better illustration, we only show the feature alignment at one pixel location in this figure. For the pixel location (m, n) in target proposal X^t, a duplication operation is performed on it's feature to generate a feature map which has the same size with the original target proposal feature map. Then, the duplicated feature map together with the feature map of the support proposal go through a 1×1 convolution layer to generate the target proposal relation feature X_r^t and support proposal relation feature X_r^s. After that, a correlation map is calculated with these two relation features:

$$V_{t,s}^{corr}(m,n,x,y) = \frac{\Big(X_r^t(m,n) - \mu(X_r^t(m,n))\Big)\Big(X_r^s(x,y) - \mu(X_r^s(x,y))\Big)}{\sigma(X_r^t(m,n)) \cdot \sigma(X_r^s(x,y))} \tag{7}$$

where $V_{t,s}^{corr}(m,n,x,y)$ represents the correlation value between the support proposal feature at location (x, y) and the target proposal feature at location (m, n), $X_r^t(m,n)$ denotes the relation feature vector at pixel location (m, n) of the target proposal, $\mu(X_r^t(m,n))$ and $\sigma(X_r^t(m,n))$ are the mean and stand variation of this feature vector, respectively, $X_r^s(x,y)$ denotes the relation feature vector at pixel location (x, y) of the support proposal, $\mu(X_r^s(x,y))$ and $\sigma(X_r^s(x,y))$ are the mean and stand variation of this feature vector, respectively. After that, a softmax operation is applied on the correlation map along the spatial location dimension to obtain the alignment weight. Finally, the alignment weight and the original support proposal feature map are multiplied, so that the aligned feature at location (m, n) is obtained:

$$Align(X_i^s)(m,n) = \sum_{x=1}^{M}\sum_{y=1}^{N} V_{t,s}^{corr}(m,n,x,y) \cdot X_i^s(x,y). \tag{8}$$

where M and N are the height and width of the proposal feature map, respectively.

3.5 Class constraint

The class constraint on the temporal aggregation module is conducted by plugging in a proposal classifier after the region proposal network. After obtaining the object proposals for each frame with the RPN, the proposal classifier will classify these proposals. Then with the label of each object proposal, the temporal aggregation module can performance feature aggregation for target proposal with the features of the support proposals which have the same class label with the target proposal. The plugged-in proposal classifier is trained with the whole framework simultaneously. The overall loss function L for training the proposed framework is

$$L = L_{reg}^{RPN} + L_{cls}^{RPN} + L_{reg}^{det} + L_{cls}^{det} \tag{9}$$

where L_{reg}^{RPN} denotes the bounding box regression loss of the region proposal network, L_{cls}^{RPN} is the classification loss of the plugged-in proposal classifier designed for class constraint. L_{reg}^{det} represents the bounding box regression loss of the final object detection, and L_{cls}^{det} stands for the classification loss of the final object detection.

3.6 Video level aggregation

After detecting the objects in all video frames, many state-of-the-art video object detection algorithms [2, 6, 38, 41, 44] adopt some kind of post-processing strategies (e.g., Seq-NMS) to refine the detection results by incorporating the video level information, and achieve huge improvement in the final performance. However, there are some constraints on these post-processing strategies which may weaken their power. Taking the Seq-NMS as an example, it requires that the bounding boxes of the same object in two adjacent frames should have an IoU higher than a certain threshold (i.e., large spatial overlap), but this assumption might not hold in cases like fast motion, sudden shot change, etc. Moreover, the post-processing

Feature Aggregation	None	TA only	SR only	TA + SR	CC + TA + SR	CC + TA + SR + FA
mAP(%)	73.4	82.6	78.8	83.4	84.4	84.8
mAP(%) slow	82.4	88.1	86.6	88.4	88.9	89.4
mAP(%) medium	71.6	82.0	76.3	82.6	83.3	83.7
mAP(%) fast	51.4	67.6	56.4	68.2	69.2	70.1

Table 1: Ablation studies on the proposed modules. 'None' means no feature aggregation is performed, 'TA only' means only temporal aggregation module is used to aggregate feature, 'SR only' means only spatial relation module is used, 'TA+SR' means both the temporal aggregation module and the spatial relation module are used, 'CC + TA + SR' means the class-constrained spatial and temporal relation module are used, and 'CC + TA + SR + FA' means the class-constrained spatial and temporal relation module + feature aggregation module are used. mAP slow/medium/fast represent the detection precision for object with slow motion, medium motion and fast motion, respectively.

strategies are not optimized with the object detection network together, which may lead to some sub-optimal results [40].

Therefore, instead of employing any post-processing strategies, we propose to capture the full-video level object information in an alternative way. Given that there is no constraint about temporal order for our temporal aggregation module (i.e., our TA module is exempt from optical flow estimation [44], bounding box shift prediction [9], etc.), there is no need to concern the temporal consistency when sampling the support frames. In other words, the support frames can be randomly selected from the video sequence without keeping the original temporal order or maintaining the frame continuity.

The random sampling strategy for support frames is adopted in our experiment setting. However, it is highly possible that the randomly-sampled support frames may contain some objects that are similar to the target proposal object in terms of appearance, but are actually in different classes. When we enhance the appearance feature of the target proposal in TA module, we should use the feature of the objects that are in the same class as the target object to perform temporal feature aggregation.

Compared with the previous post-processing strategies such as Seq-NMS, our proposed strategy of capturing video level object information has advantages in many aspects: first, the whole temporal feature aggregation module with class homogeneity constraint can be trained end-to-end, which can avoid sub-optimal results; second, the class homogeneity constraint will further feedback to the backbone network to learn more discriminative appearance features for objects in different classes; third, without any post-processing operation after detecting the object from video frames, the proposed strategy is more time efficient.

4 EXPERIMENTS

We extensively evaluate the proposed framework in this section. First, network implementation details are introduced. Then, data and evaluation metric are illustrated. After that, several ablation studies are conducted. Then we study how sampling strategy of support frames affects the performance. Finally, the comparison with state-of-the-arts is performed.

4.1 Network implementation

Backbone network We adopt ResNet-101 [18] as the backbone network to extract features for each video frame.

Region feature extraction network Region Proposal Network (RPN) [33] is applied on the extracted feature of $conv4$ of ResNet-101 to obtain the object proposals for the target and support video frames. Totally 9 anchors with 3 different scales and 3 different aspect ratios are leveraged in RPN. During both training and inference stage, we first extract 6000 proposals with highest objectness scores for each frame, and then Non-Maximum Suppression (NMS) on these proposals is performed with IoU threshold of 0.7 to finally get 300 proposals. RoI pooling followed by a fully-connected layer are applied on the $conv5$ feature to extract RoI features of each proposal.

4.2 Dataset and evaluation metric

The proposed framework is trained with an intersection of the ImageNet DET and VID datasets [34] by taking their shared 30 object classes, with the same training and validation split settings as [44]. After training, the framework is evaluated on the VID validation dataset with all 30 classes. The widely-used mean average precision (mAP)@IoU=0.5 is adopted as the evaluation metric.

4.3 Training and testing

The proposed model is trained end-to-end on 8 P100 GPUs. We first initialize the backbone network with the pre-trained weights on ImageNet classification, then all modules in the model are trained and optimized simultaneously. Note that the RPN, temporal aggregation module, spatial relation module and the final detection layers are trained from scratch. A total of 10 epochs are performed to train the model with a SGD optimizer. Batch size is set to 8 with each GPU holds one minibatch. We use an initial learning rate of $2.5e^{-4}$, which is divided by 10 after 4 epochs, and divided again after another 4 epochs. During training, every training target frame is sampled along with two random support frames in the same video sequence (identical frames for the DET dataset), When testing, for every inference frame (target frame), another N frames will be randomly sampled from the same video sequence as the support frames. In both training and testing, the video frames are resized to be with shorter dimension of 600 pixels.

4.4 Ablation study on feature aggregation

We perform several ablation studies in this subsection and the evaluation results are shown in Table 1.

Looking at the second row of Table 1, the improvement by introducing SR module only (5.4% gain, 73.4% to 78.8%) is moderate, while the gain by introducing TA module itself is much larger (9.2% gain, 73.4% to 82.6%). This is because, the spatial relation module is exploiting information within the same frame, thus it can not deal with the problems in video sequences very well, such as occlusion, motion blur, and rare pose in the target frame. The huge improvement by using TA module demonstrates that for the typical challenges in video (occlusion, motion blur and rare pose), the temporal aggregation module can leverage the feature in other frames which are not occluded, not blurred and with various poses to enhance their features and benefit the detection. Additionally, as we can see, including both TA module and SR module yields an even better performance (83.4%), indicating that both TA module and SR module are effective in our VOD task. Moreover, the class-constrained TA+SR module further improves the performance to 84.4%, with 1% improvement, which reflects that homogeneity constraint is effective to filter out those irrelevant proposals from other classes. Final, adding the proposed feature alignment module, our method achieves the accuracy of 84.8%, further justifies the usefulness to align the features.

Looking at the 3nd - 5th rows of Table 1, we could observe that, the TA module provides larger gain when the motion in video goes faster, i,e., the gain in fast motion is +16.2% (from 51.4% to 67.6%), while the gain in slow motion and medium motion is +5.7% (from 82.4% to 88.1%) and +10.4% (from 71.6% to 82.0%). This indicates that the TA module indeed does its job of exploiting temporal information across frames, especially in fast motion cases when the temporal relationship comes from frames within a long time range. On the other hand, the spatial relation module provides similar gains in cases of slow/medium/fast motion (+4.2%, +4.7% and +5.0% respectively). This is consistent with our intuition that, the SR module only cares about the relations within the same frame, thus is agnostic to the motion patterns. Interestingly, we can also see from the table that, both the class-constraint (CC) and the feature aggregation (FA) improve the fast motion largely, which indicates that the fast motion case is more challenging, and the proposed CC and FA modules are indeed helpful to tackle these challenges.

4.5 Analysis of frame sampling settings

For inference, the number of support frames N we select from the video sequence is an important parameter, and matters a lot for video object detection. Usually, sampling more support frames during testing yields better detection results [41, 44]. The effect of this parameter on the final detection performance is shown in Table 2. From the Table we can see that the detection performance improves consistently by sampling more support frames for testing, then the performance saturates when enough support frames are used (i.e., 30). The reason is that with more support samples offered, more appearance information (e.g., shape, pose, etc.) have been mined, and adding more support frames does not bring in extra information. Given the fact that more support frames means longer processing time, we set the number of support frames as 20 in our following experiments (unless otherwise noted), for the trade-off between detection precision and time efficiency.

# frames	2	6	14	20	30
mAP(%)	82.4	83.7	84.3	84.8	84.9
mAP(%) slow	87.0	88.1	88.7	89.4	88.4
mAP(%) medium	80.1	82.3	82.9	83.7	83.8
mAP(%) fast	64.6	67.8	68.9	70.1	70.2

Table 2: Effect of support frame number N on detection precision. mAP slow is the detection precision for object with slow motion, mAP medium is the detection precision for object with medium motion, and mAP fast is the detection precision for object with fast motion.

Then we take a closer look at how support frame number affects the detection of objects with different motion speed in videos. Table 2 show that objects with fast motion gain more improvement by using more support frames than objects with medium/slow motion, while objects under slow motion gain the least. This is consistent with our intuition. Usually objects under fast motion have much more appearance variation, and are more easily occluded by some other objects during video capturing. Sampling more support frames can provide various and prolific supplementary appearance information for the target proposal objects with deteriorated appearance, and therefore the detection performance is improved with more support frames. On the other hand, objects with slow motion usually have much less appearance variation in a video sequence, i.e. nearby frames are similar to each other, therefore sampling more support frames can not provide much extra information.

Next we evaluate the effectiveness of the random sampling strategy. First, we perform testing with 20 consecutive support frames (i.e., 10 consecutive frames before the target frame and 10 frames after), and the performance is shown in the second column of Table 3 ('Con W/O NMS'). Then we adopt the Seq-NMS post-processing to refine the result, which is shown in the third column of Table 3 ('Con W/ NMS'). The result of randomly sampling support frames is in the last column of Table 3.

From Table 3 we can see that consecutively sampling 20 support frames to perform feature aggregation while without any post-processing performs the worst among these three methods. This is because 20 consecutive frames capture a scene happening within 1 second, which means the object motion and appearance information are limited, especially for objects under slow motion. Moreover, for objects with fast motion, it is very possible that the object appears, disappears and re-appears in the video, but the 20 consecutive support frames are only a small portion of the whole sequence and only provide limited object information for the objects in the target frame, making the feature aggregation not ideal. When the Seq-NMS post-processing strategy is adopted, video level object information can be explored for object detection, and the detection result of the consecutive sampling can be improved (+1.6% mAP). Random sampling strategy achieves the best performance (+3.8 % mAP) over consecutive sampling (both with or without Seq-NMS), showing that the random sampling strategy can capture the object information from the whole sequence, and is more robust to deal with fast motion, sudden shot change that Seq-NMS suffers from.

Sampling strategy	Con W/O NMS	Con W/ NMS	Random
mAP(%)	81.0	82.6	84.8

Table 3: Effect of different sampling strategy on detection precision. 'Con W/O NMS' means sampling 20 consecutive support frames and without Seq-NMS post-processing, 'Con W/ NMS' means sampling 20 consecutive support frames and with Seq-NMS post-processing, 'Random' means sampling 20 support frames from the video randomly.

4.6 Comparison with state-of-the-arts

To evaluate the effectiveness of our proposed model, we compare it with some state-of-the-arts, and summarize the results in Table 4.

The comparison is performed under the circumstance that all models are with the same backbone. The results show that our model outperforms the single-frame object detection method D & T [9] (75.8% mAP) by a large margin (+9.0%). Besides, our model is remarkably better than FGFA [44] (76.3% mAP) and MANet [38] (78.1% mAP), which both aggregate features based on optical flow estimation, and the mAP improvements are +8.5% mAP and +6.8% mAP, respectively. When compared with some relation-based method (LRTRN [36] (81.0% mAP), RDN [5] (81.8% mAP), SELSA [40] (82.7% mAP)), our method also shows its superior on detection precision.

When some post-processing strategies are adopted, most of these methods gain more or less on detection precision (mAP). The D & T method even achieves a +4.0% mAP improvement by employing the tubelet rescoring post-processing. Other methods (FGFA, MANet) also improve by +1 ∼ 2% mAP. In [5], a novel post-processing technique called Box Linking with Relations (BLR) is specifically designed for their RDN method, and achieves better refined detection result (83.8mAP). However, we found that our proposed method gains no improvement by adding post-processing strategies (e.g., Seq-NMS). This is possibly because our class-constrained spatial-temporal relation network already well exploits the spatial-temporal coherence among bounding boxes.

We further compare our method with some state-of-the-arts by reporting motion-specific detection precision, which is shown in Table 5. From this table we can see that all of these methods perform well on objects under slow motion, and our proposed method outperforms other methods by a large margin on objects with medium and fast motion, especially on fast moving object detection. The reason is that when feature aggregation is performed based on optical flow estimation (MANet), the estimation of optical flow of fast moving objects is much more challenging, which in turn leads to bad feature aggregation. The relation-based feature aggregation method (LRTRN) performs feature aggregation without optical flow estimation, however, the consecutive sampling strategy limits its ability of leveraging long-term video level information for feature aggregation. Moreover, the lack of spatial object topological relation exploration also weakens its feature aggregation. SELSA also adopts the random sampling strategy to select supportive frames, which guarantees that SELSA can exploit the long-term video level information for feature aggregation, however, without considering the heterogeneity of the spatial and temporal information by separating the spatial and temporal feature aggregation and the feature alignment before feature aggregation, the performance of SELSA is worse than ours.

Method	Backbone	base detector	mAP (%)
D & T [9]	ResNet-101	R-FCN	75.8
D & T*	ResNet-101	R-FCN	79.8
FGFA [44]	ResNet-101	R-FCN	76.3
FGFA#	ResNet-101	R-FCN	78.4
MANet [38]	ResNet-101	R-FCN	78.1
MANet#	ResNet-101	R-FCN	80.3
ST-Lattice* [2]	ResNet-101	Faster R-CNN	79.6
STSN# [1]	ResNet-101+DCN	R-FCN	80.4
STMN# [41]	ResNet-101	Faster R-CNN	80.5
PSLA [14]	ResNet-101+DCN	R-FCN	80.0
PSLA#	ResNet-101+DCN	R-FCN	81.4
LRTRN [36]	ResNet-101	Faster R-CNN	81.0
RDN [6]	ResNet-101	Faster R-CNN	81.8
RDN + BLR	ResNet-101	Faster R-CNN	83.8
SELSA [40]	ResNet-101	Faster R-CNN	82.7
SELSA#	ResNet-101	Faster R-CNN	82.7
Ours	ResNet-101	Faster R-CNN	**84.8**

Table 4: Comparison with state-of-the-arts on ImageNet VID validation set. 'X+Y' means post-processing strategy Y is employed on method X. * represents +tubelet rescoring,# represents +Seq-NMS

Method	MANet [38]	LRTRN [36]	Ours
mAP(%)	78.1	81.0	**84.8**
mAP(%) slow	86.9	86.7	**89.4**
mAP(%) medium	76.9	79.5	**83.7**
mAP(%) fast	56.7	64.2	**70.1**

Table 5: More detailed comparison with state-of-the-arts on ImageNet VID validation set.

5 CONCLUSION

Feature aggregation is verified effective for video object detection. In this work, we propose a class-constrained spatial-temporal relation network and a correlation based feature alignment module to exploit a better feature aggregation. The class-constrained spatial-temporal relation network considers both the long-range temporal dependencies between objects in the same class across frames, and the spatial topological relations between different objects in the same frame. Moreover, the random sampling strategy of support frames leverages the full-video object information in an alternative way which is more robust compared with traditional post-processing strategies such as Seq-NMS. The feature alignment module aligns the target and support proposals to make the aggregated feature better. Extensive experiment results demonstrate that the proposed two improvements are effective, and it achieves state-of-the-art performance on video object detection task.

REFERENCES

[1] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. 2018. Object detection in video with spatiotemporal sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 331–346.

[2] Kai Chen, Jiaqi Wang, Shuo Yang, Xingcheng Zhang, Yuanjun Xiong, Chen Change Loy, and Dahua Lin. 2018. Optimizing video object detection via a scale-time lattice. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7814–7823.

[3] Xinlei Chen and Abhinav Gupta. 2017. Spatial memory for context reasoning in object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 4086–4096.

[4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*. 379–387.

[5] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. 2019. Object Guided External Memory Network for Video Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 6678–6687.

[6] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. 2019. Relation Distillation Networks for Video Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 7023–7032.

[7] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. 2009. An empirical study of context in object detection. In *2009 IEEE Conference on computer vision and Pattern Recognition*. IEEE, 1271–1278.

[8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 2758–2766.

[9] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2017. Detect to track and track to detect. In *Proceedings of the IEEE International Conference on Computer Vision*. 3038–3046.

[10] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. 2017. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659* (2017).

[11] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. 2019. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 244–253.

[12] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[14] Chaoxu Guo, Bin Fan, Jie Gu, Qian Zhang, Shiming Xiang, Veronique Prinet, and Chunhong Pan. 2019. Progressive Sparse Local Attention for Video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*.

[15] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. 2016. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465* (2016).

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 37, 9 (2015), 1904–1916.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[20] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. 2018. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3588–3597.

[21] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. 2017. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 10 (2017), 2896–2907.

[22] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2016. Object detection from video tubelets with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 817–825.

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982* (2018).

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[26] Mason Liu and Menglong Zhu. 2018. Mobile video object detection with temporally-aware feature maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5686–5695.

[27] Mason Liu, Menglong Zhu, Marie White, Yinxiao Li, and Dmitry Kalenichenko. 2019. Looking Fast and Slow: Memory-Guided Mobile Video Object Detection. *arXiv preprint arXiv:1903.10172* (2019).

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

[29] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 891–898.

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[31] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.

[32] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

[35] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4510–4520.

[36] Mykhailo Shvets, Wei Liu, and Alexander C Berg. 2019. Leveraging Long-Range Temporal Relationships Between Proposals for Video Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 9756–9764.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[38] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. 2018. Fully motion-aware network for video object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 542–557.

[39] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7794–7803.

[40] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. 2019. Sequence Level Semantics Aggregation for Video Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 9217–9225.

[41] Fanyi Xiao and Yong Jae Lee. 2018. Video object detection with an aligned spatial-temporal memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 485–501.

[42] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1492–1500.

[43] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. 2018. Towards high performance video object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7210–7218.

[44] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 408–417.

[45] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Deep feature flow for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2349–2358.