Identity-Consistent Aggregation for Video Object Detection

Chaorui Deng[1], Da Chen[2], Qi Wu[1,*]
[1]Australia Institute of Machine Learning, University of Adelaide
[2]Department of Computer Science, University of Bath

Abstract

In Video Object Detection (VID), a common practice is to leverage the rich temporal contexts from the video to enhance the object representations in each frame. Existing methods treat the temporal contexts obtained from different objects indiscriminately and ignore their different identities. While intuitively, aggregating local views of the same object in different frames may facilitate a better understanding of the object. Thus, in this paper, we aim to enable the model to focus on the identity-consistent temporal contexts of each object to obtain more comprehensive object representations and handle the rapid object appearance variations such as occlusion, motion blur, etc. *However, realizing this goal on top of existing VID models faces low-efficiency problems due to their redundant region proposals and nonparallel frame-wise prediction manner. To aid this, we propose ClipVID, a VID model equipped with Identity-Consistent Aggregation (ICA) layers specifically designed for mining fine-grained and identity-consistent temporal contexts. It effectively reduces the redundancies through the set prediction strategy, making the ICA layers very efficient and further allowing us to design an architecture that makes parallel clip-wise predictions for the whole video clip. Extensive experimental results demonstrate the superiority of our method: a state-of-the-art (SOTA) performance (84.7% mAP) on the ImageNet VID dataset while running at a speed about 7× faster (39.3 fps) than previous SOTAs.*

1. Introduction

Video Object Detection (VID) aims to recognize and localize the objects in all frames given a video clip. It is a challenging task as it must handle the complex appearance variations of video objects, caused by motion blur, occlusion, rotation, unusual poses, and deformable shapes, *etc*. To tackle these issues, prior works [16, 26, 27, 53] utilize a set of support frames (*e.g*., neighboring frames of the target frame), which provide rich temporal contexts, to guide the object detection in the target frame. For example, [52, 43, 3] build grid-level relations between the feature maps of the support frames and the target frame to propagate temporal contexts. More recent SOTA methods [45, 8, 19, 10, 28] adopt object-level relation modules [24] to leverage the region proposals extracted from long-range support frames to enhance the object representations in the target frame.

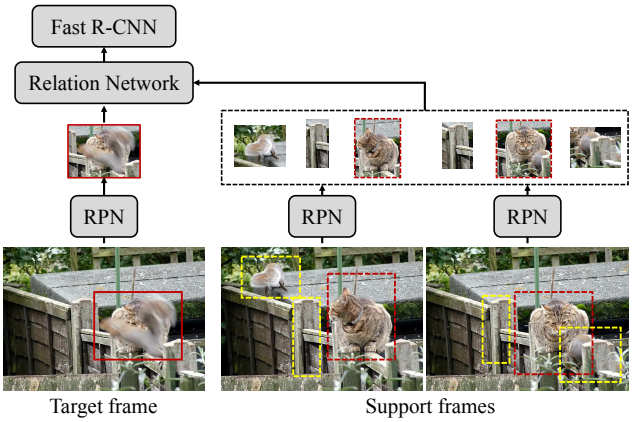

Figure 1. Illustration of the temporal context aggregation in a typical VID method [45]. The region proposals from the support frames (dashed boxes) are treated indiscriminately regardless of their object identities. However, to detect the cat in the target frame, the region proposals with red dashed boxes should provide more relevant information, as they are obtained from the same cat.

However, as shown in Figure 1, the temporal contexts from the support frames usually contain irrelevant and noisy information that may negatively affect the object representations. *E.g.*, when detecting the cat in the target frame, the yellow boxes in the support frames only provide information from different objects and even from the background. On the other hand, the red boxes are different local views of the same cat, showing it from various perspectives. Intuitively, incorporating these local views into a unified representation could lead to a more comprehensive understanding of the object, and further facilitate the model to deal with the rapid variations of the object appearance. Unfortunately, existing methods make no distinction between these two kinds of temporal contexts. In light of this, we propose

*Corresponds to qi.wu01@adelaide.edu.au. Code is available at https://github.com/bladewaltz1/clipvid.

an Identity-Consistent temporal context Aggregation (ICA) approach, which aims to discover and utilize the local views of each object to learn its global view to guide the detection.

To achieve this, a prerequisite is to ensure that the region proposals extracted from support frames have a high recall rate for the video objects, so that each object in the target frame can find its identity-consistent temporal contexts from them. This requires existing methods to extract a large number of region proposals (*e.g.*, 300) from each support frame due to the redundant predictions made by their base detectors [11, 35]. However, the computation complexity of the object relation module is usually quadratic to the total number of region proposals in the video. Therefore, existing methods have to lower the number of region proposals extracted from the support frames to make the computation feasible, decreasing the recall rate of the video objects and hampering the ICA process. Worse still, the low recall rate also leads to a low detection performance on the support frames. Thus, given an input video clip, existing methods only make predictions for one target frame, while the support frames are merely used as guidance. This non-parallel behavior further hampers the model efficiency.

For these reasons, we build our VID model based on the DETR [5] framework. Specifically, instead of generating a large number of redundant object candidates, we represent each object in a video frame with a learnable embedding, *a.k.a.* "object query", and iteratively incorporate its object-related visual content from the frame representation and its identity-consistent temporal contexts from the video into the object query. Hungarian algorithm is used to perform one-to-one bipartite matching between the object queries and the ground-truth objects. In this way, the total number of object queries can be typically less than 100 per frame, which is an order of magnitude smaller than the number of region proposals in previous methods. Thus, the whole detection process can be achieved efficiently, further allowing our model to perform parallel clip-wise predictions.

The proposed model, termed ClipVID, adopts a clean backbone + Transformer decoder architecture. It first extracts features from each frame separately using a CNN-based [22] backbone. Then, the object queries for all input frames are adaptively generated and are fed into a Transformer decoder jointly to propagate the temporal contexts. To perform identity-consistent temporal context aggregation, we assign each object query with an object identity, and additionally predict an identity embedding for it, which is then adopted to select the object queries from other frames that are close in the embedding space. These selected object queries are considered to have the same object identity and are fed into an ICA module to maintain the identity consistency of the video objects. Finally, the object queries from all frames are fed into the detection head jointly to obtain their predictions in parallel.

When evaluated on the ImageNet VID dataset [37], ClipVID achieves a significant performance improvement in fast-moving objects, which are the type of objects that suffers mostly from the appearance variations in a video, *e.g.*, motion blur, occlusion, and deformation. This further leads to a state-of-the-art overall performance (84.7% mAP) without the need for post-processing. Moreover, our model is able to run at 39.3 fps, which is about $7\times$ faster than recent SOTAs. In summary, our contributions are three-fold:

1. We propose the ClipVID model which is able to leverage the identity-consistent temporal contexts to obtain comprehensive representations for the video objects, leading to a SOTA performance on the VID task.

2. The proposed ClipVID makes clip-wise predictions for the VID task, *i.e.*, detects the objects on all input frames simultaneously, which is significantly faster than previous frame-wise prediction methods.

3. We conduct extensive experiments to analyze the performance of the proposed ClipVID model.

2. Related Works

Images Object Detection. Object Detection methods in the image domain have developed rapidly over the years [35, 17, 34, 11, 32, 29, 30, 41]. Among them, Faster RCNN [35] is one of the most popular object detectors. In Faster RCNN, a backbone CNN extracts the image representation and then feeds it into a region-proposing stage to generate a large number of region proposals, followed by a detection stage to classify and refine the proposals. Non-maximum suppression (NMS) is required in both stages to remove redundancy. Deformable convolution [12] and Relation Network [24] are two useful approaches to boost object detection performance. Specifically, deformable convolution samples feature from dynamic locations to facilitate a more aligned receptive field. Relation Networks applies self-attention [42] among the region proposals to enhance their features with contextual information.

Video Object Detection. Leveraging temporal contexts as guidance has been proven to be beneficial for frame-wise VID. In FGFA [52], estimated flow fields [15] are used to wrap the features of neighbor frames to enhance the feature of the target frame. In STSN [3], deformable convolutions [12] are used to sample features from neighbor frames to boost the target frame representation. STMM [46] and PSLA [18] propagate temporal contexts from neighbor frames through a recurrent feature map memory, which communicates with the target frame through grid-level attention. To acquire more semantically diverse temporal contexts, [45, 38] adopt long-range support frames as guidance, by feeding the region proposals in the target frame and support frames together into a Relation Network [24].

MEGA [8] and OGEMN [13] propose memory mechanisms to utilize temporal contexts from both long-range and neighbor support frames. HVR-Net [19] further considers cross-video proposal relations to exploit more diverse temporal cues. TF-Blender [10] inserts a grid-level fusion layer into MEGA to leverage fine-grained and diverse temporal contexts. In contrast to previous methods where they treat the temporal contexts from different objects indiscriminately, we propose to capture the identity-consistent temporal contexts for each object explicitly.

There have also been attempts to accelerate the frame-wise VID methods through sparse computation [6, 53, 50, 25], *i.e.*, assigning more computation budgets (*e.g.*, using heavier backbones or larger frame resolutions) to the keyframes in the video, and assigning fewer computation budgets to the non-key frames. These strategies can also be applied to our model to further accelerate its speed.

End-to-End Detectors for Images/Videos. The end-to-end object detector DETR [5] has drawn great attention recently. The core idea is to use the Hungarian algorithm to perform a one-to-one label assignment between the ground-truth objects and a set of learnable object queries during training. The DETR framework effectively removes the need for many hand-crafted components like NMS and anchor generation. The main limitation of DETR is its low convergence speed. To aid this, [51, 40] adopt guided attention which only selects a subset of locations from the feature map to perform cross attention according to learned reference points/boxes. Moreover, these reference points/boxes are updated iteratively at each decoder layer.

The DETR architecture has been applied to the video domain. TransVOD [23] is a recently proposed DETR-based model which is able to perform end-to-end video object detection. However, it is still a frame-wise prediction method and contains complex sub-modules to process the support and the target frames, respectively. Differently, the proposed ClipVID is a clean backbone + Transformer decoder architecture that performs clip-wise prediction, which is simple and efficient. VisTR [44] for Video Instance Segmentation (VIS) proposes an Instance Sequence Matching strategy that addresses the object linking problem in VIS. it imposes an assumption that the differences between input frames are mild, which generally does not hold in practice. Different from VisTR, the proposed ICA can be used to link objects across frames, and is especially effective in dealing with large appearance variations.

Multi-Object Tracking (MOT) is also an extensively studied problem in computer vision. SOTA methods in MOT are dominated by the tracking-by-detection paradigm [9, 47, 4]. *I.e.*, the objects on each frame are first detected using object detectors like Faster RCNN, then associated together. In other words, the bounding boxes of these objects are pregiven and the tracker only needs to solve the association task. Differently, VID focuses on generating high-quality detection results on each frame with the help of temporal contexts. Some MOT methods [2, 49, 48] perform detection and tracking jointly, where the detected object in previous frames are used to detect and associate the objects in the current frame. Following this, a more recent work [39] adopts the DETR framework and uses object queries to detect objects at the current frame and associate them with existing tracklets. However, the dependence on previous frames enforces these methods to perform frame-wise prediction, which is inefficient.

3. Method

The architecture of the proposed method is shown in Figure 2. Given the input video clip, the proposed ClipVID first extracts the frame features using a backbone network. Then, ClipVID generates object queries adaptively for each frame and inputs them into a Transformer decoder where the self-attention operations process object queries from all frames in a unified manner to propagate temporal contexts. Then in the cross-attention operations, the object queries attend to their corresponding frame features to retrieve object-related visual contents. In the last several decoder layers, Identity-consistent Aggregation is performed to aggregate identity-consistent temporal contexts into each object query. A feed-forward network (FFN) with a detection head is applied to all object queries to make parallel clip-wise predictions.

3.1. Clip-wise Video Object Detection

ClipVID is an end-to-end video object detector with a backbone + Transformer decoder architecture. We detail its basic components as follows.

Backbone. Given a video clip of T input frames $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^{T}$, $\boldsymbol{x}_i \in \mathbb{R}^{H_0 \times W_0 \times 3}$, the backbone model extracts a lower-resolution feature map from each frame, following by a 1×1 convolution layer to reduce its dimension to d. We denote the frame features as $\boldsymbol{f} = \{\boldsymbol{f}_i\}_{i=1}^{T}$, where $\boldsymbol{f}_i \in \mathbb{R}^{H \times W \times d}$ is for each frame. Unlike DETR, we do not adopt a transformer encoder to further encode $\boldsymbol{f}$ since it is memory-consuming to process the long feature sequences ($T \times H \times W$) of the video clip.

Adaptive Object Queries. To handle input video clips with arbitrary frame lengths, a naive approach is to share a fixed set of object queries across all frames, which may lead to a low semantic diversity. Instead, we generate L object queries adaptively for each frame conditioned on a learnable embedding matrix $\boldsymbol{e} \in \mathbb{R}^{L \times d}$, resulting in $T \times L$ object queries in total, denoted by $\boldsymbol{q} = \{\boldsymbol{q}_{ij} \in \mathbb{R}^{1 \times d} | i \in$

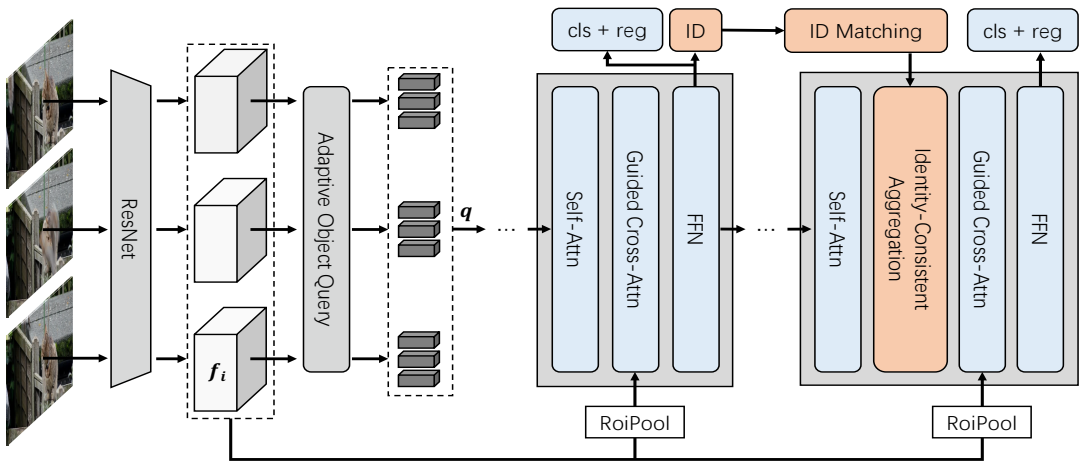

Figure 2. The overview of ClipVID. $\boldsymbol{f}_i$ and $\boldsymbol{q}$ indicate the frame feature and the object queries, respectively. "cls", "reg", and "ID" indicate the classification branch, localization branch, and identity embedding branch of the prediction head, respectively. Two transformer layers are shown. In the first one, the identity embedding branch predicts the identity embedding of each object query. Then in the second one, a Matching module is used to group the object queries according to the distances of their identity embeddings, and an ICA module is used to perform identity-consistent temporal context aggregation. For simplicity, only three frames and three queries per frame are shown.

$[1, T], j \in [1, L]\}$. Specifically, for i-th frame, its j-th adaptive object queries $\boldsymbol{q}_{ij}$ is obtained by

$$\boldsymbol{q}_{ij} = \text{softmax}(\boldsymbol{e}_j \boldsymbol{m}_i^\top)\boldsymbol{m}_i, \tag{1}$$

where $\boldsymbol{m}_i \in \mathbb{R}^{s^2 \times d}$ is obtained by firstly down-sampling the spatial dimensions of $\boldsymbol{f}_i$ to $s \times s$ and then flattening its spatial dimensions, and $\boldsymbol{e}_j \in \mathbb{R}^{1 \times d}$ is the j-th embedding.

Self Attention. The self-attention operation is performed over all $T \times L$ object queries, propagating temporal contexts among all frames to make parallel predictions for the input video clip. In self-attention, the object query is updated by:

$$\begin{aligned} \boldsymbol{q}_{ij} \longleftarrow & \text{LN}(\boldsymbol{q}_{ij} + \text{MHA}(\boldsymbol{q}_{ij}, \boldsymbol{q}, \boldsymbol{q})), \\ & \forall i \in [1, T], \forall j \in [1, L] \end{aligned} \tag{2}$$

where $\text{MHA}(q, k, v)$ indicates the multi-head attention [42] operation using the query q, key k, and value v. LN indicates the layer normalization operation [1]. After the self-attention layer, each object query is enhanced with the contextual information from the whole clip.

Guided Cross Attention. In the cross attention operation, different from DETR where an object query attends to all locations in the feature map, here we adopt the idea of guided attention [51, 40] to accelerate the convergence speed, *i.e.*, the object query $\boldsymbol{q}_{ij}$ only performs cross attention with the locations inside a reference box b_{ij}. Specifically, given the frame feature $\boldsymbol{f}_i$, RoIPooling [21] is performed over $\boldsymbol{f}_i$ based on b_{ij}, to extract a $s \times s$ feature map. The feature map is further flattened along the spatial dimensions, denoted by $\boldsymbol{k}_{ij} \in \mathbb{R}^{s^2 \times d}$. Moreover, we propose to enhance $\boldsymbol{k}_{ij}$ with the semantic information from its matched object query through an element-wise adaptation operation:

$$\boldsymbol{k}_{ij} \longleftarrow \boldsymbol{k}_{ij} + \text{reshape}(\boldsymbol{q}_{ij}\boldsymbol{W}^p), \tag{3}$$

where $\boldsymbol{W}^p \in \mathbb{R}^{d \times s^2 d}$ is a learnable parameter and $\boldsymbol{q}_{ij}\boldsymbol{W}^p$ is reshaped to $s^2 \times d$ which is added to $\boldsymbol{k}_{ij}$ element-wisely. This process is important as it fuses the grid-level and instance-level features of an object (obtained from the backbone and the decoder, respectively) in a fine-grained manner, leading to an improved object representation. Then, cross-attention is performed using $\boldsymbol{k}_{ij}$ as the source and $\boldsymbol{q}_{ij}$ as the query:

$$\boldsymbol{q}_{ij} \longleftarrow \text{LN}(\boldsymbol{q}_{ij} + \text{MHA}(\boldsymbol{q}_{ij}, \boldsymbol{k}_{ij}, \boldsymbol{k}_{ij})). \tag{4}$$

Through this guided cross-attention, each object query attends to a specific region in the corresponding frame, acquiring its object-related visual contents.

Detection Head. A feed-forward network with a detection head is append after each decoder layer to iteratively refine the detection results. Taking an object query $\boldsymbol{q}_{ij}$ as input, the detection head adopts a localization branch to predict a box offsets δ_{ij} to update the reference box b_{ij}; and a classification branch to predict the class logits p_{ij}. Denote by $y_i = \{(p_{ij}, b_{ij})\}_{j=1}^L$ the predicted objects on the i-th frame and $y_i^* = \{(c_{ij}, b_{ij}^*)\}_{j=1}^L$ the corresponding ground-truth objects padded with $\varnothing$. Following [5], ClipVID applies the set prediction loss which first finds an optimal bipartite

matching between y_i and y_i^* by searching for a permutation of L elements $\sigma \in S_L$ with the lowest cost:

$$\sigma = \arg\min_{\sigma \in S_L} \sum_{j=1}^{L} \mathcal{L}_{match}(y_{ij}^*, y_{i\sigma(j)}), \tag{5}$$

where $\mathcal{L}_{match}$ is defined as

$$\begin{aligned} \mathcal{L}_{match}(y_{ij}^*, y_{i\sigma(j)}) &= \lambda_{cls}\mathcal{L}_{cls}(c_{ij}, p_{i\sigma(j)}) \\ &+ \lambda_{giou}\mathcal{L}_{giou}(b_{ij}^*, b_{i\sigma(j)}) \\ &+ \lambda_{L1}\mathcal{L}_{L1}(b_{ij}^*, b_{i\sigma(j)}). \end{aligned} \tag{6}$$

Here, $\mathcal{L}_{cls}$ indicates the focal loss [30], $\mathcal{L}_{giou}$ and $\mathcal{L}_{L1}$ are the GIoU loss [36] and L1 loss, respectively. λ_* are coefficients of the loss terms. Then, the training objective is defined to have the same form as Eq. (6), but it is only applied to the matched pairs. The final loss is the sum of all matched pairs normalized by the number of objects inside the whole video clip.

3.2. Identity-Consistent Aggregation

The ICA module is applied in the last several layers of the transformer decoder. It consists of a Matching step, which introduces an additional identity embedding branch to the detection head of the previous decoder layer; and an Aggregation step, where an identity-consistent aggregation layer is inserted between the self-attention and cross-attention operations of the current decoder layer.

Matching. The identity embedding branch is a two-layer MLP followed by an L_2 normalization layer that projects the object query $\boldsymbol{q}_{ij}$ into an object identity embedding $\boldsymbol{h}_{ij} \in \mathbb{R}^d$. Then, given the n-th object query on the m-th frame, $\boldsymbol{q}_{mn}$, we select its most similar object query in each of the rest frames according to their dot-product similarity. The selected object queries are considered to have the same identity with $\boldsymbol{q}_{mn}$, *i.e.*, identity-consistent object queries of $\boldsymbol{q}_{mn}$, denoted by $\{\boldsymbol{q}_{iJ_{mn}(i)} | i \in I_m\}$ where:

$$\begin{aligned} J_{mn}(i) &= \arg\max_{j \in [1,L]} \boldsymbol{h}_{mn} \cdot \boldsymbol{h}_{ij}, \\ \forall i \in I_m,\ I_m &= \{i | i \in [1,T], i \neq m\}. \end{aligned} \tag{7}$$

To train the identity embedding branch, suppose the set I^* and J^* contain the frame indexes and the query indexes of all object queries that are assigned to the same ground-truth video object according to Eq. (6), respectively. Then, any two indexes i and m in I^* will give us a pair of object queries $\boldsymbol{q}_{iJ^*(i)}$ and $\boldsymbol{q}_{mJ^*(m)}$ that are consistent in their object identity, which should have a relatively small distance in the embedding space. To achieve this, we use an additional contrastive loss to train the parameters for identity-consistent feature aggregation:

$$\begin{aligned} \mathcal{L}_{con} &= -\sum_{\forall m,i \in Z} \log \frac{\exp(\boldsymbol{h}_{mJ^*(m)} \cdot \boldsymbol{h}_{iJ^*(i)})}{\sum_{j=1}^{L} \exp(\boldsymbol{h}_{mJ^*(m)} \cdot \boldsymbol{h}_{ij})}, \\ Z &= \{m, i | m, i \in I^*, m \neq i\}. \end{aligned} \tag{8}$$

The final contrastive loss is the sum of $\mathcal{L}_{con}$ for all video objects, normalized by the number of all matched pairs.

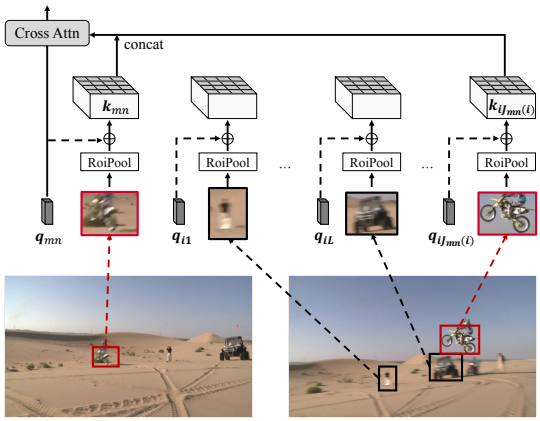

Figure 3. Illustration of the ICA process. In this example, $\boldsymbol{q}_{mn}$ and $\boldsymbol{q}_{iJ_{mn}(i)}$ indicate the same motorbike on the m-th frame and i-th frame, respectively. Thus, to obtain the "global view" of the object query $\boldsymbol{q}_{mn}$, only its identity-consistent temporal contexts, *i.e.*, the temporal contexts extracted from $\boldsymbol{q}_{iJ_{mn}(i)}$, are used for aggregation. The dashed lines indicate the positional embedding generated from the object queries.

Aggregation. For $\boldsymbol{q}_{mn}$ and its identity-consistent object queries $\{\boldsymbol{q}_{iJ_{mn}(i)} | i \in I_m\}$, we reuse their corresponding region features $\boldsymbol{k}_{mn}$ and $\{\boldsymbol{k}_{iJ_{mn}(i)} | i \in I_m\}$ obtained in Eq. (3), and stack them into a joint representation $\boldsymbol{K}_{mn} \in \mathbb{R}^{Ts^2 \times d}$. Note that, $\boldsymbol{K}_{mn}$ consists of the fine-grained grid-level feature representations of the same video object (ideally) from multiple frames, which is adopted as the identity-consistent temporal contexts for $\boldsymbol{q}_{mn}$. Finally, a cross-attention operation is performed using $\boldsymbol{q}_{mn}$ as the query and $\boldsymbol{K}_{mn}$ as the source, where $\boldsymbol{q}_{mn}$ attends to all $T \times s^2$ elements in $\boldsymbol{K}_{mn}$ to obtain its identity-related information at a fine-grained level, results in a more comprehensive "global view" of the corresponding video object. Taking this global view as input, the guided cross-attention layer then retrieves its local view from the region feature $\boldsymbol{k}_{mn}$. A detailed illustration of the ICA process is given in Figure 3.

4. Experiments

Dataset. We conduct experiments on the widely used benchmark dataset ImageNet VID [37]. It contains 30 object categories and has 3,862 training videos and 555 validation videos. Mean Average Precision (mAP) is

Methods	Backbone	mAP	fps
FGFA [52]	R-101	76.3	-
PSLA [18]	R-101	77.1	30.8
MANet [43]	R-101	78.1	-
STSN [3]	R-101	78.9	-
OGEMN [13]	R-101	79.3	-
LRTRN [38]	R-101	81.0	9.5
RDN [14]	R-101	81.8	10.6
TransVOD [23]	R-101	81.9	-
SELSA [45]	R-101	82.7	9.6
MEGA [8]	R-101	82.9	5.8
HVRNet [19]	R-101	83.2	-
TF-Blender [10]	R-101	83.8	< 5.8
DSFNet [28]	R-101	84.1	< 5.8
*STMM [46] + SeqNMS [20]	R-101	80.5	~ 1
*RDN [14] + BLR [14]	R-101	83.8	~ 1
*HVRNet [19] + SeqNMS [20]	R-101	83.8	~ 1
*MEGA [8] + BLR [14]	R-101	84.5	~ 1
ClipVID	R-101	**84.7**	39.3
RDN [14]	X-101	83.2	-
MEGA [8]	X-101	84.1	5.3
SELSA [45]	X-101	84.3	-
HVRNet [19]	X-101	84.8	-
*RDN [14] + BLR [14]	X-101	84.7	~ 1
*MEGA [8] + BLR [14]	X-101	85.4	~ 1
*HVRNet [8] + SeqNMS [20]	X-101	85.5	~ 1
ClipVID	X-101	**85.8**	25.1

Table 1. Comparisons with state-of-the-art methods on ImageNet VID dataset. * indicates the use of sequence-level post-processing methods like SeqNMS and BLR.

adopted as the evaluation metric.

Implementation Details. We use ResNet-101 [22] with dilated convolutions [7] in the last stage as the backbone for analysis. The transformer decoder has 6 layers, 8 attention heads, and a hidden dimension of $d = 384$. The number of object queries for each frame is set to 72. The reference boxes are initialized as the frame size, and the output size of the RoIPooling operation is 7. By default, the identity-consistent feature aggregation is only performed in the last two decoder layers, within the top-10 object queries that have the largest classification scores of each frame. λ_{cls}, λ_{giou}, and λ_{L1} are set to 2, 2, 5, respectively, as in [5].

Following the common practice [8, 14], we utilize both ImageNet VID and ImageNet DET datasets to train our model. During training, we randomly sample $T = 3$ frames from the same video. During inference, we use $T = 30$ frames by default. The frames are resized to a shorter side of 600 pixels. The backbone network is initialized with the ImageNet pre-trained weights, the rest model parameters are randomly initialized. The training process is separated into two stages. In the first stage, we train all model parameters except those used by the ICA modules for 180k iterations using the AdamW [33] optimizer with a total batch size of 4. The initial learning rate is set to 1e-5 and is divided by 10 at the 120k-th iteration. We then train the whole model for another 60k iterations, using an initial learning rate of 1e-6, and divide it by 10 at the 40k-th iteration.

Methods	mAP(%)	slow (%)	medium (%)	fast (%)
SELSA [45]	82.7	88.0	81.4	67.1
MEGA [8]	82.9	89.4	81.6	62.7
HVRNet [19]	83.2	88.7	82.3	66.6
DSFNet [28]	84.1	90.0	82.6	67.0
ClipVID w/o ICA	83.3	89.0	82.7	66.1
ClipVID	84.7	89.9	83.9	68.5
ClipVID (oracle ICA)	85.8	90.8	84.5	72.3

Table 2. Detailed performance comparisons on ImageNet VID. "slow/medium/fast" indicates the mean Average Precision for video objects with slow/medium/fast moving speed.

Encoder	Adaptive	Extend	Guided	1 frame		5 frames		30 frames	
Free	Query	SA	CA	mAP	fps	mAP	fps	mAP	fps
				61.8	24.4	N/A		N/A	
✓			✓	76.7	45.2	N/A		N/A	
		✓		63.2	24.4	63.5	14.2	OOM	
✓		✓	✓	78.1	43.5	79.6	43.1	82.4	41.9
✓	✓	✓	✓	78.3	43.5	80.1	43.1	83.3	41.7

Table 3. Performance analysis based on ClipVID w/o ICA. "SA" and "CA" are short for Self-Attention and Cross-Attention, respectively. The first row and the last row indicate the performance of the original DETR and the ClipVID w/o ICA, respectively. The mAP values are shown in percentage.

4.1. Comparisons with state of the arts

We first compare the proposed ClipVID with previous SOTA methods. As shown in Table 1, ClipVID achieves significantly faster (about 7×) inference speed than recent SOTA methods like TF-Blender [10] and DSFNet [28], while also outperforms them by a large margin in terms of mAP, *e.g.*, 84.7% vs. 84.1%. Compared with PSLA [18], a VID model specifically optimized for efficient inference in real-world scenarios, our model still outperforms its speed clearly (39.3 fps vs 30.8 fps). More importantly, PSLA uses sparse computation techniques to reduce computation, which is also applicable to our model to further improve the inference speed. In terms of performance, ClipVID outperforms PSLA by 6.2%, making it a better choice for deployment. Note that, the proposed ClipVID model is fully end-to-end trainable. Still, it outperforms previous SOTA approaches that are equipped with sequence-level post-processing techniques like Seq-nms [20] and BLR [14], while being nearly 40 times faster[1]. Compared with

[1] measured based on our implementations.

Methods	ClipVID		Sparse RCNN		Deformable DETR	
	w/ ICA	w/o ICA	w/ ICA	w/o ICA	w/ ICA	w/o ICA
mAP (%)	84.7	83.3	84.1	82.5	83.9	82.2
FPS	39.3	41.7	35.0	36.8	33.1	34.7

Table 4. Results of ClipVID with different query-based object detectors.

Method	Last layer	Last 2 layers	Last 3 layers	All
mAP (%)	84.4	84.7	84.5	83.9

Table 5. Performance analysis on the number of decoder layers that adopt the Identity-consistent Aggregation module.

Method	Top-5	Top-10	Top-15	All
mAP (%)	84.2	84.7	84.5	83.6
fps	39.8	39.3	38.5	30.3

Table 6. Performance and speed analysis on the number of object queries that perform the identity-consistent temporal context aggregation. Top-k denotes selecting the k object queries with the highest classification scores.

# infer frames	1	10	15	25	30	45
mAP (%)	78.3	82.1	82.7	83.2	83.3	83.4
fps	43.5	42.7	42.4	41.9	41.7	37.3

Table 7. Performance and speed analysis on the number of inference frames in ClipVID w/o ICA.

# object queries	48	64	72	80	100
mAP (%)	80.6	82.4	83.3	82.9	82.1
fps	45.7	44.4	41.7	40.0	37.2

Table 8. Performance and speed analysis on the number of object queries per frame in ClipVID w/o ICA.

# decoder layers	4	5	6	7
mAP (%)	81.9	82.8	83.3	83.3
fps	45.6	44.1	41.7	38.0

Table 9. Performance and speed analysis on the number of decoder layers in ClipVID w/o ICA.

TransVOD [23], an end-to-end VID model, our model obtains a much better performance (84.7% vs. 81.9%) with a much simpler network architecture. We obtain similar observations when using a stronger backbone ResNext-101-32x8d for ClipVID, where it outperforms SOTA methods in terms of both speed and accuracy. These results demonstrate the effectiveness of the proposed method.

4.2. Performance Analysis of ClipVID

We first analyze the performance of the proposed Identity-consistent Aggregation approach. As shown in Table 2, there is a clear performance degradation (from 84.7% to 83.3%) when removing ICA from the ClipVID model. This degradation is more obvious for fast-moving objects, where mAP fast drops significantly by 2.4%. Moreover, compared with the previous method, our ClipVID is generally much more accurate in detecting fast-moving objects. These results show the effectiveness of our ICA module in handling large appearance variations. Also note that ClipVID w/o ICA is only slightly faster than ClipVID, indicating that applying identity-consistent aggregation in ClipVID is very efficient. We further conduct an experiment using the oracle matching process, *i.e.*, for each object query, we find its identity-consistent object query by choosing those that are assigned to the same video object. We find that the performance improves clearly with the oracle query matching, especially for the fast-moving objects, where the mAP is booted by 3.8%.

Besides the proposed ICA module, ClipVID makes several modifications to the DETR framework, including 1) removing the transformer encoder; 2) using adaptive object queries; 3) extending the self-attention across frames; and 4) using guided cross-attention. Here, we analyze how our model benefits from them in Table 3. Note that we use ClipVID w/o ICA as the base model to better reveal the contributions of these modifications. From the table, the first two rows indicate performing image-level object detection without considering the temporal contexts. We find that the original DETR (the first row) produces much worse performance compared to its "encoder-free + guided cross-attention" counterpart (row 2), which may be largely due to its low convergence speed. Besides, the encoder also slows down the inference speed by nearly 2 times. In terms of video object detection, DETR trained with extended self-attention (row 3) improves the performance clearly over its original results in the single-frame inference setting. In the multi-frame setting, however, the performance is only improved marginally (from 63.2% to 63.5%) when using 5 inference frames, while suffering from a severe slowdown of the inference speed. We believe that the extremely-long feature sequences in the video setting not only dramatically increase the computation complexity of the encoder but also hampers the learning of the encoder and prevent it from leveraging useful information from the temporal contexts. Moreover, using more inference frames failed in this setting due to memory limitation.

On the other hand, our ClipVID w/o ICA model (the last row) enjoys having more inference frames, where its performance is boosted from 78.3% in the single-frame setting to 80.1% in the 5-frame setting and is further increased significantly to 83.3% when using 30 inference frames. This observation verifies the effectiveness of extended self-attention in propagating temporal information. Meanwhile, the inference speed of our model is consistently fast (more than 40 fps) in all inference settings thanks to the elimina-

Figure 4. Visualization of the detection results. "w/o ICA" indicates the ClipVID w/o ICA model. For simplicity, we only show the detection results with confidence scores higher than 0.1.

tion of the transformer encoder and its clip-wise prediction manner. Comparing the last two rows shows that the adaptive object queries perform better than fixed object queries, especially when the number of input frames is large.

Lastly, we show the performance of using different query-based object detection methods for ClipVID in Table 4. We consider Sparse RCNN [40] and Deformable DETR [51], which are widely used and have shown strong performance in still image object detection. However, we find that they perform inferior to our simple detector modified from DETR, and are also slightly slower. Notably, both Sparse RCNN and Deformable DETR adopt a heavy decoder (*i.e.*, having more parameters and incurring more computation cost) which may lead to overfitting in the relatively small ImageNet VID dataset.

4.3. Hyper-parameter Analysis

In this section, we analyze the hyper-parameters in our model design. We first analyze the effect of applying the Identity-consistent Aggregation for different numbers of decoder layers (Table 5), as well as the effect of using it for different numbers of object queries (Table 6). From Table 5, our model achieves the best result when applying ICA in the last two decoder layers. Moreover, using it for the last one or three decoder layers yields similar performances to the best setting. When using the ICA module for all decoder layers, the performance degrades to 83.9, showing that the early decoder stages may not be able to capture the object identity. From Table 6, applying ICA on the top-5, top-10, and top-15 object queries all bring clear gains over the baseline model ClipVID w/o ICA. Among them, choosing the top-10 scored object queries to perform ICA achieves the best result. When applying identity-consistent temporal context aggregation for all object queries, a clear performance degradation is observed. We hypothesize that the object queries with low classification scores may hamper the matching step of ICA and further introduce noises to the temporal contexts. Besides, the inference speed is also decreased clearly in this setting.

We provide more hyper-parameter analysis based on the ClipVID w/o ICA model. Specifically, we analyze how the number of inference frames per input clip influences the performance. From Table 7, our model benefits from larger temporal receptive fields, which is aligned with previous works. Thanks to the clip-wise prediction manner of our method, the inference speed is only mildly reduced when increasing the number of inference frames from 1 to 45. Then, we study the effect of the number of object queries per frame on the model performance. As shown in Table 8, with 72 object queries, our model yields the best performance. This is different from DETR which benefits from having more than 100 object queries. The reason could be

that the MSCOCO [31] dataset used by DETR is much more complex than ImageNet VID, in terms of the object categories and the number of objects per image. Thus, DETR requires more object queries to increase its object representation capacity. Lastly, we show the performance of our model with different numbers of transformer decoder layers in Table 9. From the table, 6 decoder layers are sufficient for ClipVID to achieve strong performance. Using fewer decoder layers can increase the inference speed but at the cost of clear performance degradation.

4.4. Visualization

We further visualize some detection results of the proposed methods in Figure 4. From the figures, the proposed ICA module qualitatively improves the detection performance. On some hard cases, ClipVID w/o ICA fails to make accurate predictions, *e.g.*, in the first column and last two columns of the first example, ClipVID w/o ICA mistakenly recognizes the squirrel as a monkey due to occlusion and the unusual pose of the squirrel. Besides, for some objects, ClipVID w/o ICA makes low confident predictions, like the cat in the third column of the first example and the squirrel in the fourth column of the second example, due to occlusion and motion blur, respectively.

5. Conclusion

Existing VID models usually treat the temporal contexts from different video objects indiscriminately despite their different identities. This may hamper the learning of object representations due to the irrelevant and noisy information contained in the temporal contexts. In this paper, we aim to perform Identity-Consistent temporal context Aggregation (ICA) to enhance the video object representations. To achieve this, we first need to reduce the redundancies in the temporal context so that ICA can be done efficiently. Thus, we proposed a VID model called ClipVID which is based on the DETR framework. ClipVID is able to perform identity-consistent aggregation, while also effectively removing the redundancies and making predictions for all input frames simultaneously, making the model very efficient. In the experiment, our ClipVID model outperforms previous SOTAs on the benchmark ImageNet VID dataset in terms of both speed and accuracy.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019.

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

[4] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6247–6257, 2020.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020.

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

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[8] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10337–10346, 2020.

[9] Peng Chu and Haibin Ling. Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6172–6181, 2019.

[10] Yiming Cui, Liqi Yan, Zhiwen Cao, and Dongfang Liu. Tf-blender: Temporal feature blender for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[11] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, pages 379–387, 2016.

[12] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017.

[13] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Object guided external memory network for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6678–6687, 2019.

[14] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7023–7032, 2019.

[15] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van

Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.

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

[17] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[18] Chaoxu Guo, Bin Fan, Jie Gu, Qian Zhang, Shiming Xiang, Veronique Prinet, and Chunhong Pan. Progressive sparse local attention for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3909–3918, 2019.

[19] Mingfei Han, Yali Wang, Xiaojun Chang, and Yu Qiao. Mining inter-video proposal relations for video object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 431–446. Springer, 2020.

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

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[23] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. End-to-end video object detection with spatial-temporal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1507–1516, 2021.

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

[25] Zhengkai Jiang, Yu Liu, Ceyuan Yang, Jihao Liu, Peng Gao, Qian Zhang, Shiming Xiang, and Chunhong Pan. Learning where to focus for efficient video object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 18–34. Springer, 2020.

[26] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. Object detection in videos with tubelet proposal networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 727–735, 2017.

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

[28] Lijian Lin, Haosheng Chen, Honglun Zhang, Jun Liang, Yu Li, Ying Shan, and Hanzi Wang. Dual semantic fusion network for video object detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1855–1863, 2020.

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

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

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

[32] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37. Springer, 2016.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

[36] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

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

[38] Mykhailo Shvets, Wei Liu, and Alexander C Berg. Leveraging long-range temporal relationships between proposals for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9756–9764, 2019.

[39] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

[40] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan

Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021.

[41] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9627–9636, 2019.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

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

[44] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021.

[45] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9217–9225, 2019.

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

[47] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3988–3998, 2019.

[48] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129(11):3069–3087, 2021.

[49] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020.

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

[51] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

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

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