Target-Aware Object Discovery and Association for Unsupervised Video Multi-Object Segmentation

Tianfei Zhou[1], Jianwu Li[2]*, Xueyi Li[2], Ling Shao[3]
[1] Computer Vision Laboratory, ETH Zurich, Switzerland
[2] School of Computer Science and Technology, Beijing Institute of Technology, China
[3] Inception Institute of Artificial Intelligence, UAE

Abstract

This paper addresses the task of unsupervised video multi-object segmentation. Current approaches follow a two-stage paradigm: 1) detect object proposals using pre-trained Mask R-CNN, and 2) conduct generic feature matching for temporal association using re-identification techniques. However, the generic features, widely used in both stages, are not reliable for characterizing unseen objects, leading to poor generalization. To address this, we introduce a novel approach for more accurate and efficient spatio-temporal segmentation. In particular, to address ***instance discrimination****, we propose to combine foreground region estimation and instance grouping together in one network, and additionally introduce temporal guidance for segmenting each frame, enabling more accurate object discovery. For* ***temporal association****, we complement current video object segmentation architectures with a discriminative appearance model, capable of capturing more fine-grained target-specific information. Given object proposals from the instance discrimination network, three essential strategies are adopted to achieve accurate segmentation: 1) target-specific tracking using a memory-augmented appearance model; 2) target-agnostic verification to trace possible tracklets for the proposal; 3) adaptive memory updating using the verified segments. We evaluate the proposed approach on DAVIS$_{17}$ and YouTube-VIS, and the results demonstrate that it outperforms state-of-the-art methods both in segmentation accuracy and inference speed.*

1. Introduction

Unsupervised video object segmentation aims at automatically segmenting primary object(s) from the background in unconstrained videos, which is a fundamental vision task. This task has become increasingly popular due to its potential values in a wide range of real-world applications, *e.g.*, video compression [17], autonomous driving [14], and human-centric understanding [52, 67]. It also plays an essential role in collecting large-scale annotated dataset [34, 63]. However, the task is challenging due to the lack of prior knowledge about the target objects, as well as the challenging factors (*e.g.*, occlusions, cluttered background, diverse motion patterns) carried by video data.

Towards better segmenting the prominent foreground objects, early studies typically exploit saliency cues [50, 11] or objectness priors [25, 64, 56, 66, 27] for identifying them. More recently, with the advent of deep neural networks, many learning-based models have been proposed to learn more discriminative video object patterns, by leveraging motion cues [43], addressing spatiotemporal features [23, 68], exploring multi-frame contextual information [32, 48, 61] or using recurrent networks to capture sequential information [42]. Though impressive results have been achieved, these approaches mainly focus on foreground/background separation, hindering their applications in more practical multi-object scenarios.

Unsupervised video multi-object segmentation, with an elegant and formal definition in [4], is more challenging as it requires not only discovering instance-agnostic, foreground regions automatically, but also discriminating different object instances and associating the same identities over the entire sequence. To tackle this task, existing methods [33, 42, 49, 65] generally follow the conventional tracking-by-detection paradigm which performs in a top-down fashion to employ image-aware instance segmentation networks (*e.g.*, Mask R-CNN [18], SOLO [54]) to detect object candidates in individual frames, and associate them over consecutive frames based on object tracking or proposal re-identification (ReID). In addition, to avoid the negative impact of background objects, many studies [42, 49, 65] also rely on a foreground/background separation step to remove background proposals. Even though these approaches demonstrate compelling performance, they still suffer several limitations. **1)** Directly using image-level instance segmentation networks is insufficient

*Corresponding author: *Jianwu Li*.

since they are trained on static images, neglecting the informative temporal context in videos. **2)** Instance segmentation and foreground estimation are often separately considered by different networks, incurring high computational expense. **3)** ReID-based matching networks, trained completely offline, focus more on general object appearance, while rarely capturing distinctive fine-grained features of specific targets.

In this work, we propose a novel approach for unsupervised multi-object segmentation in unconstrained videos. To address points **1)** and **2)**, we introduce an instance discrimination network (D-Net) for video object proposal. The network performs in a bottom-up fashion and takes video temporal information into account to achieve better segmentation accuracy and efficiency. In particular, the D-Net includes two branches: a foreground estimation branch follows the typical design of fully convolutional networks to segment attention-grabbing objects, and an instance segmentation branch learns to predict the instance center as well as the offset from each pixel to its corresponding center for instance grouping. Rather than processing each frame independently, we consider segmentation of previous frames as an important guidance for segmenting the current frame. In this way, we integrate instance-agnostic and instance-aware segmentation together into one network for discovering temporal coherent object proposal.

To address point **3)**, we design a target-aware tracking network (T-Net) for associating object proposals of the same identities over each image sequence. Different from previous ReID-based matching techniques, we aim to learn target-specific appearance features for more robust object association. More specifically, the T-Net learns a discriminative appearance model for each object instance during the inference stage to predict a coarse but robust segmentation score of the target object. Note that the appearance model is more prone to drifting due to the lack of ground-truth annotations. Therefore, we further propose a target-agnostic backward verification module to examine the tracking results. The verified results are used as new training samples to update the appearance model online.

With above efforts, our algorithm achieves state-of-the-art results on the DAVIS_{17} benchmark for video multi-object segmentation. It also demonstrates compelling performance for video instance segmentation on YouTube-VIS. In addition, our approach obtains a better trade-off between segmentation accuracy and inference efficiency, running at about 10 FPS on images with 480p resolution.

To sum up, the contributions of the proposed approach are three-fold: **First**, we propose a novel bottom-up instance discrimination network which takes advantage of temporal context information in videos for more accurate segmentation. The network couples foreground discovery and instance grouping together, benefiting from multi-tasking and improving the inference efficiency. **Second**, we introduce a target-aware tracking model for online matching of object proposals. Compared with target-agnostic approaches, our method can better capture the appearance information of the target objects, yielding more robust association. **Third**, our approach achieves compelling performance on the popular DAVIS_{17} and YouTube-VIS benchmarks. Furthermore, its high inference speed enables our method to support a wide variety of practical applications.

2. Related Work

Unsupervised Video Object Segmentation. The task of U-VOS aims to segment conspicuous and eye-catching objects without any human intervention. Most current research efforts focus on segmenting the prominent foreground objects in unconstrained videos. Earlier methods typically rely on motion analysis, *e.g.*, extracting motion information from sequential images to understand object movement. For example, a large number of works [3, 12, 35, 10] employ sparse point trajectories to capture long-term motion information and segment the objects which are moving significantly in relation to the background. However, these methods are not robust because they rely heavily on optical flow estimation and feature matching, and thus may easily fail in the presence of occlusions, fast motion or appearance changes. To address this limitation, later methods employ saliency cues [11, 50] and object proposal ranking [25, 64, 36] to better identify the main objects. These *non-learning* methods are confined by the limited representative ability of handcrafted features.

More recently, with the renaissance of artificial neural networks in computer vision, *deep learning* based solutions are now dominant in this field. For example, AGS [49] proposes a dynamic visual attention-driven model for video object segmentation, while [48, 32] mine higher-order relations between video frames, resulting in more comprehensive understanding of video content and more accurate foreground estimation. Moreover, many approaches discover the motion patterns of objects [43] as complementary cues to object appearance. This is typically achieved within two-stream networks [23, 44, 10, 22, 68, 26], in which an RGB image and the corresponding flow field are separately processed by two independent networks and the results are fused to produce the final segmentation. To avoid the expensive computation of optical flow, some methods [51, 31] directly feed consecutive frames into the networks and automatically learn spatiotemporal feature representations.

Unsupervised Video Multi-Object Segmentation. Unlike the aforementioned U-VOS approaches that pay more attention to learning powerful object representations for foreground object discovery, in the multi-object setting, the challenges become how to discover and segment each object that captures human attention, and how to associate the

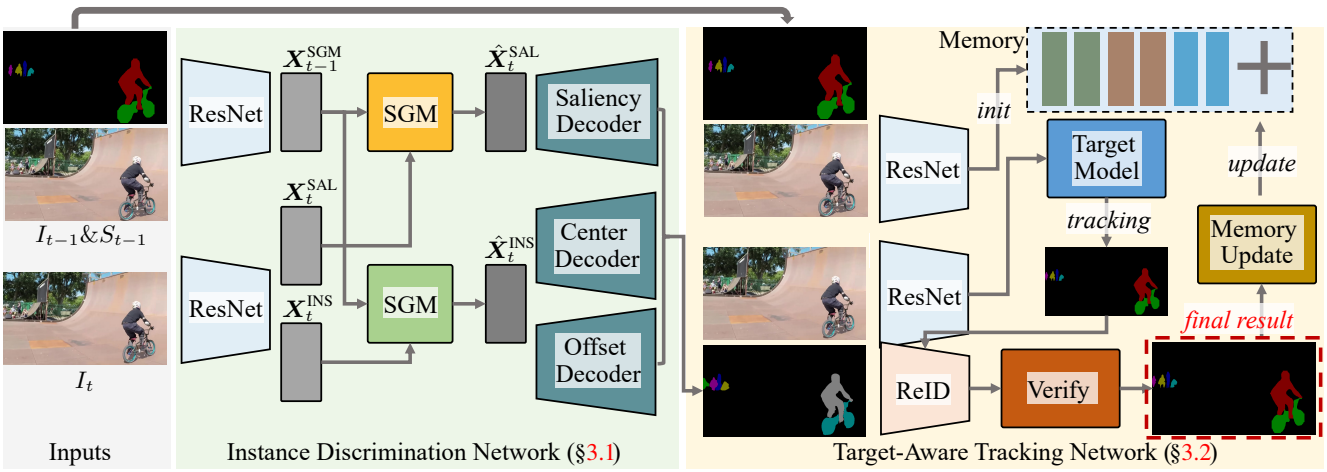

Figure 1: **Detailed illustration of our approach** for video multi-object segmentation.

objects through the whole sequence. RVOS [45] delivers an end-to-end recurrent neural network, in which the spatial recurrence helps to discover object instances in each frame, while the temporal recurrence finds the matching between instances in different frames. This work represents an early attempt towards end-to-end learning for unsupervised video multi-object segmentation. However, RVOS is weak in the segmentation performance due to its limited capability in instance discrimination. Current trends [33, 65] follow a two-stage pipeline, in which object proposals are first discovered using Mask R-CNN[18], and the association is conducted using greedy- or ReID-based matching techniques.

Instance Segmentation in Images and Videos. In recent years, image-level instance segmentation has attracted great research interests, which extends semantic segmentation [30, 28, 53, 62] to assign different labels for separate instances of objects belonging to the same class. Driven by the success of R-CNN [15], current dominant instance segmentation methods follow a detect-then-segment framework. Earlier methods[37, 8] learn to propose segment candidates, and then classify them by Fast R-CNN [15]. These methods conduct segmentation before recognition, which are slow and less accurate. Mask R-CNN[18] introduces an extra ROI segmentation head into Faster R-CNN [38] and a new assignment operator, *i.e.*, ROIAlign, to better align the ROI features with inputs. Along this line, some works improve the performance by employing cascade inference [5], low-level feature enhancement [29], and multi-tasking [6]. However, for complicated scenarios with many instances, the inference time of two-stage methods is unacceptably high, since it is proportional to the number of instances. The resolution of ROI features and resulting masks are coarse, resulting in poor segmentation of object boundaries.

To cope with these drawbacks, many recent works favor bottom-up instance segmentation. These approaches are often box-free and thus not restricted by anchor locations and scales, naturally benefiting from the inherent advantages of fully convolutional networks. For example, [9, 13] learn discriminative embeddings to group the pixels into an arbitrary number of object instances. SOLO [54] introduces a direct instance segmentation method that can predict instance segmentation in one shot without additional grouping post-processing. AdaptIS [41] first generates point proposals as representations of instances, and then sequentially predicts the corresponding segmentation mask for each detected proposal. PolarMask [57] utilizes the polar representation to encode masks and transforms per-pixel mask prediction to distance regression.

Though these methods only focus on image-level segmentation, we emphasize that they have motivated a number of video analysis tasks, such as video object/instance segmentation [59, 21, 2, 69] and multi-object tracking [70, 47]. In this work, we further propose a novel bottom-up approach for segment proposal generation in videos. Instead of frame-by-frame segmentation, we take advantage of the segmentation results in previous frames as temporal guidance, yielding more robust results.

3. Methodology

Given a video $\mathcal{I} = \{I_t\}_{t=1}^{N}$ with N frames $I_t \in \mathbb{R}^{3\times h\times w}$ with spatial size $h \times w$, the goal of unsupervised video multi-object segmentation is to automatically generate a collection of non-overlapping segment tracks, each for an individual instance. As shown in Fig. 1, we decompose the problem into two sub-tasks: 1) discover object instances using the D-Net (§3.1) and 2) associate all instances of the same identity over the entire sequence with the T-Net (§3.2).

3.1. Instance Discrimination Network (D-Net)

The D-Net consists of four major components: 1) a backbone network for feature extraction; 2) a segmentation guidance module to employ previous segmented masks to enrich the feature representations; 3) a foreground estimation head

for primary object prediction; and 4) an instance segmentation head for instance-level prediction.

Feature Extraction. Given the video frame I_t at time t, we use a backbone CNN model to extract convolutional features $\boldsymbol{X}_t^{\text{SAL}} \in \mathbb{R}^{W\times H\times C}$, $\boldsymbol{X}_t^{\text{INS}} \in \mathbb{R}^{W\times H\times C}$, where W, H and C represent the width, height and channel number of the 3D tensors, respectively. $\boldsymbol{X}_t^{\text{SAL}}$ and $\boldsymbol{X}_t^{\text{INS}}$ indicate two task-specific features that are responsible for salient foreground estimation and instance-aware segmentation, respectively. To achieve this, we take the convolutional blocks of ResNet-50 [19] as the backbone, and modify the last residual block with an atrous convolution with rate 2 to enlarge the receptive field. Furthermore, in order to extract task-specific features, we augment the backbone network with two parallel atrous spatial pyramid pooling (ASPP) modules [7]. ASPP applies several parallel atrous convolutions with different rates to further increase the receptive field. In our model, we design each ASPP to have 1) one 1×1 convolution and three 3×3 convolutions with astrous rates $=(6, 12, 18)$ and 2) a global average pooling layer on the last feature map of the backbone to obtain global context information. The resulting features from all branches are then concatenated and passed through an extra 1×1 convolutional layer to obtain $\boldsymbol{X}_t^{\text{SAL}}$ and $\boldsymbol{X}_t^{\text{INS}}$. Both of these tensors have $C=256$ channels and output strides of 16.

Segmentation Guidance Module (SGM). Rather than directly using the image-level feature representations (*i.e.*, $\boldsymbol{X}_t^{\text{SAL}}$ and $\boldsymbol{X}_t^{\text{INS}}$) for segmentation prediction, we propose to exploit the inherent correlation among video frames for better results. Particularly, the segmentation mask $S_{t-1}$of the previous frame I_{t-1} is leveraged as a guidance to improve the representation in the current frame. We introduce an extra convolutional branch whose input is the concatenation of I_{t-1} and S_{t-1}. The input is processed with a similar backbone and ASPP module to obtain feature $\boldsymbol{X}_{t-1}^{\text{SGM}}$. Then, two segmentation guidance modules are used to enrich the feature representations as follows:

$$\hat{\boldsymbol{X}}_t^{\text{SAL}} = \mathcal{F}^{\text{SGM}}(\boldsymbol{X}_t^{\text{SAL}}, \boldsymbol{X}_{t-1}^{\text{SGM}}), \tag{1}$$

$$\hat{\boldsymbol{X}}_t^{\text{INS}} = \mathcal{F}^{\text{SGM}}(\boldsymbol{X}_t^{\text{INS}}, \boldsymbol{X}_{t-1}^{\text{SGM}}). \tag{2}$$

Each guidance module $\mathcal{F}^{\text{SGM}}$ has a squeeze-and-excitation structure [20]. In particular, we first squeeze the global spatial information of each feature into channel-wise statistics using a squeeze operation $\mathcal{F}^{\text{SQ}}$:

$$\boldsymbol{c}_t^{\text{SAL}} = \mathcal{F}^{\text{SQ}}(\boldsymbol{X}_t^{\text{SAL}}) \in \mathbb{R}^C, \tag{3}$$

$$\boldsymbol{c}_t^{\text{INS}} = \mathcal{F}^{\text{SQ}}(\boldsymbol{X}_t^{\text{INS}}) \in \mathbb{R}^C, \tag{4}$$

where $\mathcal{F}^{\text{SQ}}$ is a global average pooling layer. Then, the two channel-wise descriptors are concatenated together to obtain $\boldsymbol{c} \in \mathbb{R}^{2C}$, which is then processed by an excitation operation $\mathcal{F}^{\text{EX}}$:

$$\boldsymbol{z} = \mathcal{F}^{\text{EX}}(\boldsymbol{c}; \boldsymbol{W}) = \text{softmax}(\text{reshape}(\boldsymbol{W}_2\delta(\boldsymbol{W}_1\boldsymbol{c}))) \in \mathbb{R}^{2\times C}, \tag{5}$$

where $\boldsymbol{W}_{\{1,2\}}$ denotes two fully connected layers, δ refers to the ReLU function. Note that, after the fully connected layers, we reshape the corresponding vector into $2\times C$ (which consists of two vectors $\alpha \in \mathbb{R}^C$ and $\beta \in \mathbb{R}^C$). Then, we apply a *softmax* function to ensure $\alpha+\beta=1$. Finally, we obtain the features $\hat{\boldsymbol{X}}_t^{\text{SAL}}$ and $\hat{\boldsymbol{X}}_t^{\text{INS}}$ as follows:

$$\hat{\boldsymbol{X}}_t^{\text{SAL}} = \alpha_1\boldsymbol{X}_t^{\text{SAL}} + \beta_1\boldsymbol{X}_{t-1}^{\text{SGM}}, \tag{6}$$

$$\hat{\boldsymbol{X}}_t^{\text{INS}} = \alpha_2\boldsymbol{X}_t^{\text{INS}} + \beta_2\boldsymbol{X}_{t-1}^{\text{SGM}}, \tag{7}$$

where $\alpha_{1,2}$, $\beta_{1,2}$ denote the attention vectors to weight contributions of different features.

Salient Object Estimation Head. Given $\hat{\boldsymbol{X}}_t^{\text{SAL}}$, we propose a simple yet effective decoder for salient object estimation. Specifically, we first bilinearly upsample $\hat{\boldsymbol{X}}_t^{\text{SAL}}$ by a factor of 2 and then concatenate it with the corresponding low-level features from the backbone network with the same spatial resolution (*i.e.*, `res3`). The upsampled features are further processed by a 5×5 convolutional layer, and then upsampled again by a factor of 2. After concatenating then with the features in `res2`, we process them with two consecutive 5×5 convolutional layers and one 1×1 convolutional layer to obtain the foreground estimation result S_t. Finally, the cross entropy loss is employed to evaluate the result against the corresponding ground-truth.

Instance-Aware Segmentation Head. The instance segmentation head has a similar architecture to the foreground estimation head, only differing in that it predicts two outputs, *i.e.*, an object center heatmap and a pixel offset field. Inspired by recent point-aware object representations [71, 24], we represent each object instance by its center. For dense prediction, we additionally predict the offset of each pixel to its corresponding instance center. During training, ground-truth instance centers are encoded by a 2-D Gaussian with a standard deviation of 10 pixels. We adopt the mean squared error (MSE) loss to minimize the distance between the predicted heatmaps and Gaussian-encoded ground-truths. For the offset learning, we employ the l_1 loss for optimization, which is only activated at pixels belonging to foreground object regions. During inference, predicted foreground pixels are grouped to the closest object center based on the predicted offset field, completing the instance grouping.

3.2. Target-Aware Tracking Network (T-Net)

Target-Specific Tracking. For each object instance, we build a target-specific appearance model to discriminate the target from background distractors. Specifically, we instantiate the model with a two-layer fully convolutional network as in [39]:

$$\boldsymbol{S} = \mathcal{T}(\boldsymbol{X}; \boldsymbol{W}) = \boldsymbol{W}_2 * (\boldsymbol{W}_1 * \boldsymbol{X}), \tag{8}$$

where $\boldsymbol{X}$ denotes the image feature of frame $I \in \mathcal{V}$, $\boldsymbol{W} = \{\boldsymbol{W}_1, \boldsymbol{W}_2\}$ are the network parameters of the two convolu-

tional layers, and $*$ indicates the convolution operator. $\boldsymbol{S}$ is the output of the T-Net $\mathcal{T}$, which indicates coarse segmentation score prediction. For the semi-supervised VOS task [39], Eq. (8) is trained over a set of m training samples $\mathcal{S} = \{(\boldsymbol{X}_j, \boldsymbol{y}_j)\}_{j=1}^{m}$ collected from the ground-truth annotations in the first frame, by minimizing the following objective function:

$$\mathcal{L}(\boldsymbol{W}; \mathcal{S}) = \sum_j \alpha_j \|\mathcal{T}(\boldsymbol{X}_j; \boldsymbol{W}) - \boldsymbol{y}_j\|^2 + \sum_k \lambda_k \|\boldsymbol{W}_k\|^2, \quad (9)$$

where $\boldsymbol{y}_j$ denotes the target label of $\boldsymbol{X}_j$ and $\alpha_j \geq 0$ is the weight of $\boldsymbol{X}_j$, controlling the impact of the sample on the objective. The parameter λ balances the contributions of the objective term as well as the regularization term. Note that the training sample set $\mathcal{S}$ is critical for robust model learning, especially in the unsupervised setting. In contrast to the semi-supervised setting, no ground-truth $\boldsymbol{y}_0$ is available for model training in the first video frame, and directly training on the segment proposals generated from D-Net is more prone to drifting. To address this, in our approach, the segment proposal from D-Net for each target serves as the pseudo ground-truth label $\check{\boldsymbol{y}}_0$ for initial model learning. Unlike [39], which regularly updates the training set $\mathcal{S}$ using tracked segments, we design heuristic strategies for the backward verification of tracking results, and adaptively update $\mathcal{S}$. This enables our model to be robust to the noises in $\check{\boldsymbol{y}}_0$, and greatly boosts the performance.

Target-Agnostic Verification. Let $\check{\boldsymbol{y}}_j$ and $\mathcal{Y}$ denote the tracking result of a target in frame I_j and its corresponding tracklet, respectively. We aim to verify the consistency between $\check{\boldsymbol{y}}_j$ and $\mathcal{Y}$, as well as find a better possible candidate from the object proposal set. This is achieved by matching the object proposals in the current frame with historical tracking results. To promote the reliablity of verification, we conduct the matching in a target-agnostic manner, using a pre-trained ReID network [33]. For each object proposal p, its matching score with $\mathcal{Y}$ is computed as:

$$s(p, \mathcal{Y}) = (\cos(p, \check{\boldsymbol{y}}_j) + \cos(p, \check{\boldsymbol{y}}_0)) * \mathbb{1}(\text{IoU}(p, \check{\boldsymbol{y}}_j) > 0.5), \quad (10)$$

where $\cos(\cdot, \cdot)$ indicates the cosine similarity between two ReID embeddings, $\text{IoU}(\cdot, \cdot)$ measures the intersection-over-union between two segments, and $\mathbb{1}(\cdot) \in \{0, 1\}$ is the indicator function. In Eq. (10), we first examine the overlap ratio between p and $\check{\boldsymbol{y}}_j$, which is used to truncate the ReID similarities. For more reliable matching, we compare p with the most recent and the most distant tracking results, *i.e.*, $\check{\boldsymbol{y}}_j$ and $\check{\boldsymbol{y}}_0$. This allows our model to conduct sequential modeling, while at the same time dealing with long-term semantic consistency, leading to more robust matching results. Based on Eq. (10), we find the proposal with the highest score s with $\mathcal{Y}$. If s is above a threshold th_{reid} (*e.g.*, 0.6), we replace the current tracking result $\check{\boldsymbol{y}}_j$ with the corresponding proposal; otherwise, we keep $\check{\boldsymbol{y}}_j$ unchanged. Besides, we discover new targets if all corresponding proposals have zero matching scores with all existing tracklets as well as very small IoUs (< 0.1) with tracking results in the current frame. This provides high flexibility to our model in dealing with occlusions and discovering newly-appearing objects.

Adaptive Memory Updating. Once the tracking result $\check{\boldsymbol{y}}_j$ is verified, we add a new sample $\{\boldsymbol{X}_j, \check{\boldsymbol{y}}_j\}$ into the training set $\mathcal{S}$ in order to guide the learning of latest appearance features. The sample is first assigned a weight $\alpha_j = (1 - \gamma)^{-1}\alpha_{j-1}$, where $\alpha_0 = \gamma$. For proposals with $m > th_{\text{reid}}$, we double the corresponding weight α_j so that the model can put more emphasis on reliable object proposals. Then, we normalize all weights in the training set to unity. During inference, if $m > th_{\text{reid}}$, we intermediately update the appearance model in the frame; otherwise, we update the model every eight frames.

4. Experiment

In this section, we present the experimental results of our approach. We first elaborate on the datasets, training and testing settings in §4.1. Then, we investigate the performance of our method for the unsupervised video multi-object segmentation task in §4.2 and video instance segmentation task in §4.3, respectively. Visual comparison results are presented in §4.4. We conduct detailed ablative experiments in §4.5. Finally, we provide run time analysis to quantify the efficiency of the system in §4.6.

4.1. Experimental Settings

Datasets. We conduct experiments on two popular datasets:

- **DAVIS**$_{17}$ [4] for *video multi-object segmentation*. The dataset consists of 120 high-quality videos in total. These videos are further split into 60 for `train`, 30 for `val` and 30 for `test-dev`. In our experiments, we train our models only on the `train` split, without any additional data. Then, we evaluate the performance of our approach on `val` and `test-dev`.
- **YouTube-VIS** [58] for *video instance segmentation*. It contains 2,883 high-resolution videos collected from YouTube, covering more than 131K object instances. Different from DAVIS$_{17}$ in which objects are category-agnostic, objects in YouTube-VIS are labeled with one semantic category out of 40 categories. Therefore, the task in YouTube-VIS not only requires the algorithms to segment consistent objects but also assign each object a category label. We use this dataset to examine the performance of our model in more challenging scenarios.

Training Phase. We train the D-Net on the training set of DAVIS$_{17}$ and YouTube-VIS. During training, each sample is randomly augmented with a scaling factor of [0.8, 1.5] and horizontal flipping, and is then cropped to 640×640. For optimization, we use the standard SGD solver, with a

Dataset	Method	Pub.	$\mathcal{J}\&\mathcal{F}$ Mean↑	$\mathcal{J}$ Mean↑	$\mathcal{J}$ Recall↑	$\mathcal{J}$ Decay↓	$\mathcal{F}$ Mean↑	$\mathcal{F}$ Recall↑	$\mathcal{F}$ Decay↓
val	RVOS[45]	$CVPR_{19}$	41.2	36.8	40.2	0.5	45.7	46.4	*1.7*
	‡OF-Tracker[1]	-	54.6	53.4	60.9	-1.3	55.9	63.0	1.1
	‡RI-Tracker[1]	-	56.9	55.5	63.3	2.7	58.2	64.4	6.4
	PDB[42]	$ECCV_{18}$	55.1	53.2	58.9	4.9	57.0	60.2	6.8
	AGS[49]	$CVPR_{19}$	57.5	55.5	61.6	7.0	59.5	62.8	9.0
	ALBA[16]	$BMVC_{20}$	58.4	56.6	63.4	7.7	60.2	63.1	7.9
	MATNet[65]	TIP_{20}	58.6	56.7	65.2	-3.6	60.4	68.2	1.8
	AGNN[48]	$ICCV_{19}$	61.1	58.9	65.7	11.7	63.2	67.1	14.3
	STEm-Seg[1]	$ECCV_{20}$	64.7	61.5	70.4	**-4.0**	67.8	75.5	1.2
	*UnOVOST[33]	$WACV_{20}$	**67.9**	**66.4**	**76.4**	-0.2	**69.3**	**76.9**	**0.0**
	Ours	-	65.0	63.7	71.9	6.9	66.2	73.1	9.4

Dataset	Method	Pub.	$\mathcal{J}\&\mathcal{F}$ Mean↑	$\mathcal{J}$ Mean↑	$\mathcal{J}$ Recall↑	$\mathcal{J}$ Decay↓	$\mathcal{F}$ Mean↑	$\mathcal{F}$ Recall↑	$\mathcal{F}$ Decay↓
test-dev	RVOS[45]	$CVPR_{19}$	22.5	17.7	16.2	1.6	27.3	24.8	1.8
	PDB[42]	$ECCV_{18}$	40.4	37.7	42.6	4.0	43.0	44.6	3.7
	AGS[49]	$CVPR_{19}$	45.6	42.1	48.5	2.6	49.0	51.5	2.6
	MSP[40]	$DAVIS_{20}$	57.9	52.9	60.4	16.7	63.0	**69.5**	20.5
	*UnOVOST[33]	$DAVIS_{19}$	58.0	54.0	62.9	3.5	62.0	66.6	6.6
	Ours	-	**59.8**	**56.0**	**65.1**	7.8	**63.7**	68.4	11.0

Table 1: **Quantitative video multi-object segmentation results on the `val` and `test-dev` sets of $DAVIS_{17}$** in terms of region similarity $\mathcal{J}$ and boundary accuracy $\mathcal{F}$. '$DAVIS_{19}$' and '$DAVIS_{20}$' indicate the unsupervised tracks of the DAVIS 2019 and 2020 challenges, respectively. ‡: baseline methods implemented in [1]. *: methods has complex heuristic post-processing.

momentum of 0.9 and weight decay of 5e-4. We utilize the polynomial annealing procedure to schedule the learning rate. For the T-Net, we use ResNet-101 as the backbone network of the appearance model. The layer $\boldsymbol{w}_1$ is a 1×1 convolutional layer that reduces the channel of input features to 96 while $\boldsymbol{w}_2$ is a 3×3 convolutional layer with one output channel. The two layers are optimized online using the Gauss-Newton algorithm[39] with the default settings, which leads to significantly faster convergence than other gradient descent-based approaches.

Testing Phase. Given a test video, we run our instance discrimination network and target-aware adaptive tracking network to process each frame sequentially. The input image sizes for the two networks are separately set to 480×854 and 473×473. Our model requires no additional post-processing components (*e.g.*, CRF), which guarantees high efficiency (10 FPS) against state of the arts.

Evaluation Metrics. We follow the standard evaluation settings used in each dataset for evaluation. 1) For $DAVIS_{17}$, we report the performance in terms of region similarity $\mathcal{J}$, boundary accuracy $\mathcal{F}$, and the overall metric $\mathcal{J}\&\mathcal{F}$. The evaluation scores on the `test-dev` set are obtained from the evaluation server of the $DAVIS_{20}$ challenge, since the ground-truths of the set are private. 2) For YouTube-VIS, we follow [58] to use average precision (AP) and average recall (AR) as the metrics, which are adapted from the image instance segmentation task to the video instance segmentation task.

Method	mAP	AP_{50}	AP_{75}	AR_1	AR_{10}
‡DeepSORT[55]	26.1	42.9	26.1	27.8	31.3
‡FEELVOS[46]	26.9	42.0	29.7	29.9	33.4
‡OSMN[60]	27.5	45.1	29.1	28.6	33.1
MaskTrack R-CNN[59]	30.3	51.1	32.6	31.0	35.5
SeqTracker[59]	27.5	45.7	28.7	29.7	32.5
STEm-Seg[1]	35.0	56.0	38.6	34.4	41.7
Ours	**37.1**	**57.1**	**40.9**	**34.8**	**43.2**

Table 2: **Quantitative video instance segmentation results on Youtube-VIS `val`**, in terms of AP and AR. The baselines denoted with ‡ were implemented by the authors in[59].

4.2. Performance Comparison on $DAVIS_{17}$

We compare our approach with state-of-the-art video multi-object segmentation methods on the $DAVIS_{17}$ benchmark. In addition to recently published works (*e.g.*, RVOS [45], PDB [42], AGS [49] , ALBA [16], MATNet [65]), we also include some top-ranked solutions (*i.e.*, UnOVOST [33], MSP[40]) from the unsupervised tracks of the DAVIS-2019 and DAVIS-2020 VOS challenges. This leads to a more comprehensive examination of the proposed approach. As reported in Table 1, on $DAVIS_{17}$ `val`, our approach achieves the second-best overall results across most metrics. It is slightly worse than UnOVOST, the champion solution in DAVIS-2019 VOS challenge. However, we emphasize that UnOVOST is computationally expensive, requiring not only Mask R-CNN for instance proposal generation, but also needing to compute optical flow for motion estimation. Complex post-processing and heuristics also make the method unsuitable for many practical appli-

Figure 2: Qualitative results of multi-object segmentation masks on DAVIS$_{17}$ `test-dev`. From top to bottom: *basketball-game*, *bmx-rider*, *ducks*, *mascot*, and *snowboard-race*.

Model	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN[18]	48.1	70.0	51.1	29.9	51.8	62.2
SOLO[54]	46.8	68.6	49.3	28.7	50.4	60.9
D-Net *w/o* SGM	50.3	71.9	53.3	28.9	53.6	65.3
Full D-Net	**52.0**	**73.3**	**54.9**	**30.3**	**55.4**	**67.5**

Table 3: **Ablation study of D-Net on DAVIS$_{17}$ `val`.** We report the AP scores with and without the segmentation guidance module (*i.e.*, SGM). For comparison, we report the performance of Mask R-CNN[18] and SOLO[54], which are representative methods for top-down and bottom-up instance segmentation models, respectively. All the models use ResNet-50 as the backbone. See §4.5.

Variant	$\mathcal{J}\&\mathcal{F}$ Mean	$\mathcal{J}$ Mean	$\mathcal{J}$ Recall	$\mathcal{F}$ Mean	$\mathcal{F}$ Recall
w/o. target verification	61.3	58.2	63.5	61.9	66.3
w/o. memory updating	61.2	59.4	66.3	62.8	69.2
Full Model	**65.0**	**63.7**	**71.9**	**66.2**	**73.1**

Table 4: **Key component analysis of the proposed T-Net on DAVIS$_{17}$ `val`.** See §4.5 for details.

cations. In addition, we see that our approach outperforms all other comparative approaches.

On DAVIS$_{17}$ `test-dev`, our method outperforms all competitors, including all top solutions in the challenges. Keeping in mind that `test-dev` is more challenging than `val` and the ground-truths are kept private, the good performance over this set can better support our approach.

4.3. Performance Comparison on YouTube-VIS

We further examine the performance of the proposed approach on YouTube-VIS, which requires not only segmenting and tracking objects, but also recognizing their semantic categories. To this end, we modify the foreground estimation head in the D-Net to predict semantic labels of each pixel (instead of original binary labels) following general semantic segmentation networks [30, 7]. We train all the networks on the training data of YouTube-VIS. As reported in Table 2, our approach outperforms all the comparative methods with respect to all metrics. We improve the AP by **+2.1%** in comparison with the most recent model STEm-Seg [1]. Our approach also significantly outperforms existing two-stage methods, like MaskTrack R-CNN [59], demonstrating its superiority.

4.4. Qualitative Result

In Fig. 2, we show the qualitative segmentation results of our approach on `test-dev`. Different colors are used to indicate different object instances. From the figures, we can see the remarkable performance of the proposed approach in 1) accurately discovering distinct objects in complex scenarios (*e.g.*, low-light illumination in *gold-fish*), as well as 2) producing robust and temporally coherent object tracking across the sequence. Moreover, our approach shows

Method	Instance Proposal	Foreground Estimation	Tracking / Matching	Total Time (s)
RVOS [45]	-	-	0.07	0.07
PDB [42]	0.74	0.70	0.03	1.47
AGS [49]	0.74	0.10	0.03	0.87
MATNet [65]	0.74	0.75	0.03	1.52
UnOVOST [33]	0.74	0.20	0.08	1.02
Ours	0.05	-	0.06	0.11

Table 5: **Runtime analysis** (second/frame) on DAVIS$_{17}$ `val`. Note that our approach is much faster than existing two-stage methods. Although slightly slower than RVOS, our approach has a better tradeoff between segmentation accuracy and efficiency.

good performance in dealing with various challenging factors, such as, occlusions, scale variations, fast motion.

4.5. Diagnostic Experiment

Segmentation Guidance Module. To demonstrate the superiority of the D-Net in comparison with other counterparts, we compare it with two baseline methods (*i.e.*, Mask R-CNN [18] and SOLO [54]) in terms of category-agnostic instance segmentation on DAVIS$_{17}$ `val`. We also examine the performance of the D-Net with and without the SGM. For fair comparison, we follow the standard setting [18, 54] to use mAP as the metric for evaluation. As reported in Table 3, the D-Net *w/o* SGM achieves obvious performance improvement against the two baselines ($+2.2\%$ in terms of AP). By incorporating the SGM, our full model further improves the AP by **+1.7%**, thereby demonstrating the effectiveness of the SGM module.

Key Components in T-Net. We further conduct experiments to verify the essential components (*i.e.*, target verification and memory updating modules) in T-Net. We examine the performance by discarding each module once at a time. As summarized in Table 4, the performance drops significantly after removing each module compared with the full model, proving their efficacy.

4.6. Runtime Comparison

In addition to segmentation accuracy, runtime efficiency is also an important dimension for evaluating the usability of U-VOS algorithms. For this reason, we conduct a runtime analysis on DAVIS$_{17}$ `val` for a more comprehensive comparison. Five representative methods are used for comparison, including RVOS [45], PDB [42], AGS [49], MATNet [65], and UnOVOST [33]. For each model, we report the inference speeds in terms of three components, *i.e.*, instance proposal, foreground estimation (or salient object estimation), and instance tracking (or matching). Note that most comparative methods simply claim to use MASK R-CNN for instance proposal generation without revealing too many details (*e.g.*, backbones). Thus, we directly use the value (0.74 s) reported in UnOVOST [33] for all the methods as reference. The analysis results are summarized in Table 5. We observe that, since our approach formulates instance proposal and foreground estimation in a unified framework, it requires much less time to generate instance proposals. Further, with the efficient target-aware tracking network, our approach can run at about 10 FPS, taking 0.11 s to process one image with 480p resolution. Though it is slightly slower than RVOS (0.07 s), we have seen from Table 1 that our approach is able to produce considerably more accurate segmentation results. Fig. 3 depicts a visualization of the trade-off between accuracy and efficiency of representative algorithms on the validation set of DAVIS$_{17}$. As can be seen, our approach achieves the best trade-off.

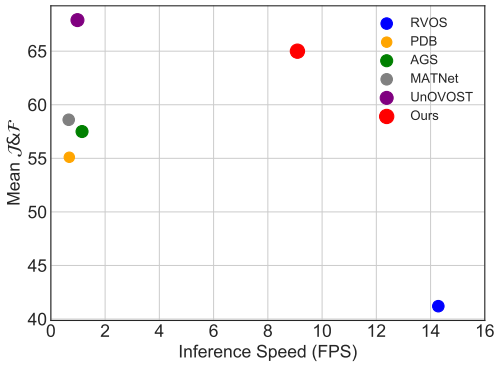

Figure 3: Trade-off between inference speed (x-axis) and segmentation accuracy (y-axis) on DAVIS$_{17}$ `val`. Our approach demonstrates compelling performance with high efficiency.

5. Conclusion

Unsupervised video object segmentation is significant in empowering machines to automatically understand dynamic real-world scenarios. In this paper, we present a novel approach for multi-object segmentation in unconstrained videos. First, we propose an instance discrimination network to discover salient instance segments in a bottom-up manner. By introducing previously well-segmented masks as guidance for segmenting later frames, the network is able to produce accurate and temporally coherent segments. Second, based on the instance proposals, we design a target-aware adaptive tracking framework to associate the proposals of the same identity across the sequence. By building a target-aware appearance model for each object, our model achieves more robust matching than previous ReID-based methods. Third, we have conducted extensive experiments on two popular benchmarks, *i.e.*, DAVIS$_{17}$ and YouTube-VIS, and the results demonstrate that our approach achieves higher segmentation accuracy against state-of-the-art methods, while running at a faster inference speed.

Acknowledgment This work was supported in part by the Beijing Natural Science Foundation under Grant L191004 and L202002.

References

[1] Ali Athar, Sabarinath Mahadevan, Aljoša Ošep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020. 6, 7

[2] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *CVPR*, 2020. 3

[3] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, 2010. 2

[4] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019. 1, 5

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: high quality object detection and instance segmentation. *TPAMI*, 2019. 3

[6] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 3

[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. *TPAMI*, 40(4):834–848, 2017. 4, 7

[8] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In *ECCV*, 2016. 3

[9] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017. 3

[10] Muhammad Faisal, Ijaz Akhter, Mohsen Ali, and Richard Hartley. Exploiting geometric constraints on dense trajectories for motion saliency. *WACV*, 2019. 2

[11] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, volume 2, page 8, 2014. 1, 2

[12] Katerina Fragkiadaki, Geng Zhang, and Jianbo Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *CVPR*, 2012. 2

[13] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. Ssap: Single-shot instance segmentation with affinity pyramid. In *ICCV*, 2019. 3

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

[15] Ross Girshick. Fast R-CNN. In *ICCV*, 2015. 3

[16] Shreyank N Gowda, Panagiotis Eustratiadis, Timothy Hospedales, and Laura Sevilla-Lara. Alba: Reinforcement learning for video object segmentation. In *BMVC*, 2020. 6

[17] Hadi Hadizadeh and Ivan V Bajić. Saliency-aware video compression. *TIP*, 23(1):19–33, 2013. 1

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 3, 7, 8

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 4

[21] Yuan-Ting Hu, Jia-Bin Huang, and Alexander Schwing. Maskrnn: Instance level video object segmentation. In *NIPS*, 2017. 3

[22] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *ECCV*, 2018. 2

[23] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, 2017. 1, 2

[24] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *CVPR*, 2020. 4

[25] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *ICCV*, 2011. 1, 2

[26] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *ICCV*, 2019. 2

[27] Jianwu Li, Tianfei Zhou, and Yao Lu. Learning to generate video object segment proposals. In *ICME*, 2017. 1

[28] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. In *AAAI*, 2021. 3

[29] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018. 3

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 3, 7

[31] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *ECCV*, 2020. 2

[32] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *CVPR*, 2019. 1, 2

[33] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. Unovost: Unsupervised offline video object segmentation and tracking. In *WACV*, 2020. 1, 3, 5, 6, 8

[34] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *CVPR*, 2018. 1

[35] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *TPAMI*, 36(6):1187–1200, 2013. 2

[36] Federico Perazzi, Oliver Wang, Markus Gross, and Alexander Sorkine-Hornung. Fully connected object proposals for video segmentation. In *ICCV*, 2015. 2

[37] Pedro OO Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *NIPS*, 2015. 3

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 3

[39] Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Learning fast and robust target models for video object segmentation. In *CVPR*, 2020. 4, 5, 6

[40] S. Kumar S. Garg, V. Goel. Unsupervised video object segmentation using online mask selection and space-time memory networks. *The 2020 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2020. 6

[41] Konstantin Sofiiuk, Olga Barinova, and Anton Konushin. Adaptis: Adaptive instance selection network. In *ICCV*, 2019. 3

[42] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, 2018. 1, 6, 8

[43] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *CVPR*, 2017. 1, 2

[44] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *ICCV*, 2017. 2

[45] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. RVOS: End-to-end recurrent network for video object segmentation. In *CVPR*, 2019. 3, 6, 8

[46] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019. 6

[47] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *CVPR*, 2019. 3

[48] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, 2019. 1, 2, 6

[49] Wenguan Wang, Jianbing Shen, Xiankai Lu, Steven C. H. Hoi, and Haibin Ling. Paying attention to video object pattern understanding. *TPAMI*, 2020. 1, 2, 6, 8

[50] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *CVPR*, 2015. 1, 2

[51] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *TIP*, 27(1):38–49, 2017. 2

[52] Wenguan Wang, Tianfei Zhou, Siyuan Qi, Jianbing Shen, and Song-Chun Zhu. Hierarchical human semantic parsing with comprehensive part-relation modeling. *TPAMI*, 2021. 1

[53] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. *arXiv preprint arXiv:2101.11939*, 2021. 3

[54] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. *arXiv preprint arXiv:1912.04488*, 2019. 1, 3, 7, 8

[55] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 6

[56] Fanyi Xiao and Yong Jae Lee. Track and segment: An iterative unsupervised approach for video object proposals. In *CVPR*, 2016. 1

[57] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *CVPR*, 2020. 3

[58] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, 2018. 5, 6

[59] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 3, 6, 7

[60] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *CVPR*, 2018. 6

[61] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, 2019. 1

[62] Yazhou Yao, Tao Chen, Guosen Xie, Chuanyi Zhang, Fumin Shen, Qi Wu, Zhenmin Tang, and Jian Zhang. Non-salient region object mining for weakly supervised semantic segmentation. 2021. 3

[63] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. 2021. 1

[64] Dong Zhang, Omar Javed, and Mubarak Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *CVPR*, 2013. 1, 2

[65] Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. MATNet: Motion-attentive transition network for zero-shot video object segmentation. *TIP*, 29:8326–8338, 2020. 1, 3, 6, 8

[66] Tianfei Zhou, Yao Lu, Huijun Di, and Jian Zhang. Video object segmentation aggregation. In *ICME*, 2016. 1

[67] Tianfei Zhou, Siyuan Qi, Wenguan Wang, Jianbing Shen, and Song-Chun Zhu. Cascaded parsing of human-object interaction recognition. *TPAMI*, 2021. 1

[68] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*, 2020. 1, 2

[69] Tianfei Zhou, Wenguan Wang, Yazhou Yao, and Jianbing Shen. Target-aware adaptive tracking for unsupervised video object segmentation. *The 2020 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2020. 3

[70] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020. 3

[71] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *arXiv preprint arXiv:1904.07850*, 2019. 4