Target-Aware Adaptive Tracking for Unsupervised Video Object Segmentation

Tianfei Zhou[1] , Wenguan Wang[1] , Yazhou Yao[2] , Jianbing Shen[1]
[1]Inception Institute of Artificial Intelligence, UAE [2]Nanjing University of Science and Technology, China
{ztfei.debug,wenguanwang.ai}@gmail.com

Abstract

This paper addresses the task of unsupervised multi-object video segmentation. Most current approaches cast the task as a re-identification solution, which associates objects across frames by generic feature matching. However, the generic features are not reliable for characterizing unseen objects, leading to poor generalization. To address this, 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 off-the-shelf detectors, three essential strategies are adopted to achieve accurate segmentation: 1) Target-specific tracking. Each determined target is sequentially tracked using a memory-augmented appearance model, wherein the memory stores historical information for re-training the appearance model online; 2) Target-agnostic verification. The tracked segments and object proposals are backward re-identified to trace possible tracklets. Departing from the tradition of only matching proposals between adjacency frames, we conduct long-term semantic matching among distant proposals. This helps to correct the inaccurate tracked segments or drifted results; 3) Adaptive memory updating. Memories are adaptively updated using the verified segments, instead of using tracked results all the time. This favors storing high-quality target information in the memory, reducing the risk for model drifting. By these carefully designs, our approach obtains state-of-the-art performance on DAVIS$_{20}$ `test-dev` set ($\mathcal{J}\&\mathcal{F}$: 59.8%) with a fast speed (15 FPS). It finally ranked 2^{rd} place in the DAVIS$_{20}$ Unsupervised Segmentation Challenge (`test-challenge` set).

1. Introduction

Unsupervised video object segmentation targets at automatically separating primary objects from the background in dynamic videos [16]. The task has gained significant attention in recent years, due to its potential benefits for a wide variety of applications, *e.g.*, embodied question answering [9], human-object interaction recognition [10, 21]. Extensive research efforts have been devoted to learning discriminative video object patterns, by leveraging motion cues [13], addressing spatiotemporal features [14, 4], or using recurrent networks [12, 17] to capture sequential information. Though impressive results have been generally achieved, these approaches focus on foreground/background separation, which are limited in instance-aware scenarios.

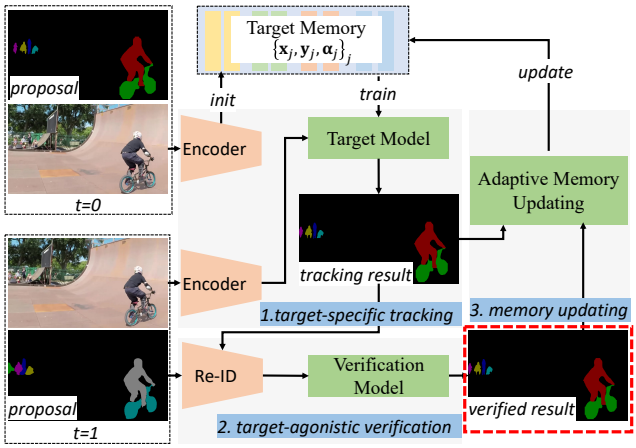

Figure 1: **Pipeline of the proposed method**, which consists of three components: target-specific tracking, target-agnostic verification and adaptive memory updating.

Video object instance segmentation is more challenging as it requires not only discovering foreground regions automatically, but also discriminating different object instances and associating instance identities across frames [1]. To address instance discrimination, image instance segmentation techniques [2] are typically applied to each single frame to generate object proposals. Then, for cross-frame identity association, matching based proposal re-identification (ReID) [8] is a nature choice. However, above two-stage paradigm easily suffers from two limitations. First, the robustness is limited. ReID networks, trained completely offline, focus more on general object appearance, while rarely

capturing fine-grained distinctive features of specific targets. Second, large amounts of data are typically needed to train the image instance segmentation and ReID modules.

In this work, we address these problems by learning robust target-specific appearance for segmentation. Inspired by [11], we utilize a light-weight discriminative appearance model to generate target-specific segmentation scores during inference. The segmentation scores then serve as guidance to achieve accurate segmentation via a boundary-aware refinement network, which is offline trained. Note that the appearance model is more prone to drift 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. The pipeline of our algorithm is depicted in Fig. 1.

With above efforts, our algorithm achieves state-of-the-art results in DAVIS$_{20}$ `test-dev` benchmark with a score of 59.8% in terms of Mean $\mathcal{J}\&\mathcal{F}$. It finally ranked 2^{rd} place in DAVIS$_{20}$ `test-challenge`. Besides, our method is efficient (15 FPS) and operates without any additional post-processing (*e.g.*, CRF).

2. Related Work

Unsupervised Video Object Segmentation (UVOS). UVOS aims to segment conspicuous video objects without any test-time human intervention. Most current research efforts focused on segmenting all primary objects together. These methods avoid the dilemma of data association, and pay more attention to enrich object representations for automatic object discovery. Specifically, they learned motion patterns to separate independent objects and camera motion [13], mined high-order contextual relationships among video frames [7, 15, 19], or exploited two-stream neural networks [4, 20]. However, in the instance-level multi-object setting, the main challenge becomes how to associate different objects across frames. Recent leading approaches [6, 8] solved this by feature matching based ReID. Though impressive, they suffer from the limited representability of generic features in characterizing specific objects, which poses great difficulties for distinguishing similar objects. In contrast, we propose to learn target-specific features for robust instance tracking, and introduce a global matching strategy to improve the tracking results.

Discrimintive Appearance Models. Appearance models have been widely explored in online visual tracking [3, 11] to capture target object appearance. Some recent efforts discriminatively learn convolution filters using efficient optimization (*e.g.*, Conjugate Gradient [3], Gauss-Newton [11]) to distinguish target from background. In this work, with a similar spirit of [11], we build a target-specific appearance model and adapt it into our instance-level unsupervised video object segmentation scenario.

3. Our Algorithm

Preliminary. Given a video sequence $\mathcal{I}$, the goal of unsupervised segmentation is to automatically generate a collection of non-overlapping segment tracks. To achieve this, our method automatically determines each important object instance and learns a discriminative appearance model to track it. Specifically, for each frame, we employ HTC [2] to generate a set of category-agnostic object proposals. In contrast to previous approaches that score the proposals only using detection confidence, we rescore them by incorporating motion-aware saliency information [20], encouraging the model to discover salient but low-confidence objects. The final score of each segment proposal is the summation of its detection confidence and saliency value. Those proposals with scores smaller than a pre-defined threshold th_{prop} are directly discarded. The remaining proposals in the first frame are treated as the initial tracking targets, and our method is capable of discovering newly appearing objects during tracking. Next, we describe our method in detail, which is mainly equipped with four components/techniques, *i.e.*, target-specific tracking, target-agnostic verification, adaptive memory updating, and a segmentation network.

Target-specific tracking. For each target, we build a target-specific appearance model to discriminate the target from background distractors. To this end, we instantiate the model with a two-layer fully convolutional network [11], *i.e.*, $D(\mathbf{x};\mathbf{w}) = \mathbf{w}_2 * (\mathbf{w}_1 * \mathbf{x})$, where $\mathbf{x}$ is the image features of frame $I \in \mathcal{I}$ and $\mathbf{w}$ denotes network parameters. Given training samples $\mathcal{M} = \{(\mathbf{x}_j, \mathbf{y}_j, \alpha_j)\}_j$, the network can be online learned by minimizing the objective:

$$\mathcal{L}(\mathbf{w};\mathcal{M}) = \sum\nolimits_j \alpha_j \|D(\mathbf{x}_j;\mathbf{w}) - \mathbf{y}_j\|^2 + \sum\nolimits_k \lambda_k \|\mathbf{w}_k\|^2, \quad (1)$$

where $\mathbf{y}_j$ denotes the target label of $\mathbf{x}_j$ and α_j is the corresponding sampling weight. The parameters λ control the regularization term. Note that the training sample set $\mathcal{M}$, or memory, is significant for model learning, especially for the unsupervised setting. In contrast to the semi-supervised setting, no ground-truth $\mathbf{y}_0$ is available for model training at the beginning; hence, the model is more prone to drifting. To address this, in our method, for each target, its segment proposal from HTC serves as the pseudo ground-truth label $\tilde{\mathbf{y}}_0$, and we augment it heavily to train the initial model. Different from [11] that regularly updates the memory using tracking results, we design heuristic strategies for online tracking verification and adaptive memory updating. These strategies help to alleviate the negative effects introduced by the noises in $\tilde{\mathbf{y}}_0$, and greatly boost the performance.

Target-agnostic verification. Let $\tilde{\mathbf{y}}_j$ and $\mathcal{T}$ denote the tracking result of a target at frame I_j and its corresponding tracklet, respectively. We aim to verify the consistency between $\tilde{\mathbf{y}}_j$ and $\mathcal{T}$, as well as find possible better candidate from the object proposal set. This is achieved by match-

ing 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 ReID network [8]. For each object proposal $\mathbf{o}$, its matching score with $\mathcal{T}$ is computed as:

$$\mathbf{s}(\mathbf{o},\mathcal{T}) = (cos(\mathbf{o},\tilde{\mathbf{y}}_j) + cos(\mathbf{o},\tilde{\mathbf{y}}_0)) * \mathbb{1}(IoU(\mathbf{o},\tilde{\mathbf{y}}_j) > 0.5), \quad (2)$$

where cos indicates the cosine similarity between two ReID embeddings, IoU denotes intersection-over-union, and $\mathbb{1}(\cdot) \in \{0,1\}$ is the indicator function. Here, we first examine the overlap between $\mathbf{o}$ with $\tilde{\mathbf{y}}_j$, which is used to truncate the ReID similarities. For more reliable matching, we compare $\mathbf{o}$ with the most recent and the most distant tracking results, *i.e.*, $\tilde{\mathbf{y}}_j$ and $\tilde{\mathbf{y}}_0$. This facilitates our model to capture long-term semantic consistency. Based on Eq. (2), we find the proposal with the highest score $\tilde{\mathbf{s}}$ with $\mathcal{T}$. If $\tilde{\mathbf{s}}$ is above a threshold th_{reid}, we replace the current tracking result $\tilde{\mathbf{y}}_j$ with the corresponding proposal; otherwise, we keep $\tilde{\mathbf{y}}_j$ unchanged. Besides, we discover new targets if the corresponding proposals have zero matching scores with all existing tracklets as well as small IoUs (<0.1) with tracking results in the current frame.

Adaptive memory updating. Once the tracking result $\tilde{\mathbf{y}}_j$ is verified, we determine to adaptively update the memory using the new sample $\{\mathbf{x}_j, \tilde{\mathbf{y}}_j, \alpha_j\}$. The sample is first given a weight $\alpha_j = (1-\eta)^{-1}\alpha_{j-1}$, where $\alpha_0 = \eta$. Besides, if $\tilde{\mathbf{s}} > th_{\text{reid}}$, we double the corresponding weight α_j so that the model can put more emphasis on relible object proposals. All the weights are then normalized to unity. During inference, if $\tilde{\mathbf{s}} > th_{\text{reid}}$, we intermediately update the appearance model at the frame; otherwise, we update the model every 8 frames.

Segmentation Network. The appearance model produces a coarse segmentation output $\mathbf{u} = D(\mathbf{x};\mathbf{w})$. It is then passed to a segmentation network S to obtain a high-resolution segmentation. Our segmentation network consists of two modules: 1) a target segmentation encoder [11] that merges the segmentation scores with backbone features; and 2) a boundary-aware refinement module [20] to produce accurate segmentation with crisp boundaries.

4. Experiments

Our approach is evaluated on DAVIS$_{20}$ `test-dev` and `test-challenge`, each containing 30 challenging sequences. Ablative experiments are conducted on DAVIS$_{20}$ `test-dev`.

Detailed Network Architecture. We use ResNet-101 [5] as the backbone network of the appearance model D and the segmentation network S. D accepts features from Res4 to produce a 1-channel coarse score map [11], while S accepts multi-scale features from Res2 to Res5, and progressively merges high-level abstract features with low-level details. The segmentation network S is offline trained on a combination of DAVIS$_{20}$ and Youtube-VOS [18] `train` splits.

Team	$\mathcal{J}\&\mathcal{F}$ Mean	$\mathcal{J}$ Mean	$\mathcal{J}$ Recall	$\mathcal{F}$ Mean	$\mathcal{F}$ Recall
Phoenix	61.6	58.4	65.0	64.7	71.1
IIAI	55.6	53.1	60.0	58.2	62.5
BLIIT	52.3	50.2	57.5	54.4	58.9
HCMUS	43.9	40.2	45.7	47.5	50.1

Table 1: **Segmentation results on DAVIS$_{20}$ `test-challenge` set** (Higher values are better). The two best scores for each metric are marked in red and blue, respectively.

Team	$\mathcal{J}\&\mathcal{F}$ Mean	$\mathcal{J}$ Mean	$\mathcal{J}$ Recall	$\mathcal{F}$ Mean	$\mathcal{F}$ Recall
IIAI	59.8	56.0	65.1	63.7	68.4
Phoenix	57.9	52.9	60.4	63.0	69.5
BLIIT	54.4	51.4	59.9	57.4	61.6
sabarim	48.6	43.9	48.0	53.3	58.2
BUAA	46.2	41.2	46.4	51.2	57.0

Table 2: **Segmentation results on DAVIS$_{20}$ `test-dev` set** (Higher values are better). The two best scores for each metric are marked in red and blue, respectively.

Variant	$\mathcal{J}\&\mathcal{F}$ Mean	$\mathcal{J}$ Mean	$\mathcal{J}$ Recall	$\mathcal{F}$ Mean	$\mathcal{F}$ Recall
Full Model	**59.8**	**56.0**	**65.1**	**63.7**	**68.4**
w/o. target verification	54.2	50.0	57.2	58.3	62.3
w/o. memory updating	56.0	51.8	59.7	60.2	64.3
w/o. saliency rescoring	58.9	53.9	61.3	62.0	67.6

Table 3: **Ablation study on DAVIS$_{20}$ `test-dev` set**.

The ReID model in [8] is used to compute embeddings for object proposals. We set $th_{\text{prop}} = 0.3$ and $th_{\text{reid}} = 0.8$ in all the experiments. Our model is implemented in PyTorch, and trained on eight NVIDIA Tesla V100 GPUs.

Results on DAVIS$_{20}$ Challenge. Table 1 and Table 2 summarize the results of top teams in `test-challenge` and `test-dev` benchmarks, respectively. Our approach (***IIAI***) achieves state-of-the-art results on `test-dev` with 59.8% in terms of $\mathcal{J}\&\mathcal{F}$ Mean. Finally, our method ranks the 2^{rd} place in `test-challenge`. Fig. 2 shows qualitative results of representative sequences in `test-challenge`. We can observe that our approach handles well various difficulties, *e.g.*, object deformation, scale variations, *etc*.

Ablation Study. We further study the impacts of essential components. As seen from Table 3, target-agnostic verification and adaptive memory updating provide substantial performance gains. Proposal rescoring also consistently boosts performance over all metrics.

5. Conclusion

This work introduces a new target-aware adaptive tracking approach for automatic segmentation of multiple object instances in videos. It demonstrates superior performance on the DAVIS$_{20}$ challenge. The contributions of essential components are examined in the ablation study.

Acknowledgements This work was sponsored by Zhejiang Lab's Open Fund (No. 2019KD0AB04), Zhejiang Lab's International Talent Fund for Young Professionals, and CCF-Tencent Open Fund.

Figure 2: Our results on DAVIS$_{20}$ `test-challenge`. From top to bottom: *dribbling*, *monster-trucks*, *surfer* and *table-tennis*.

References

[1] 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:1905.00737*, 2019. 1

[2] 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. 1, 2

[3] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ECO: Efficient convolution operators for tracking. In *CVPR*, 2017. 2

[4] 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

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[6] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *ECCV*, 2018. 2

[7] 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. 2

[8] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. UnOVOST: Unsupervised offline video object segmentation and tracking. *arXiv preprint arXiv:2001.05425*, 2020. 1, 2, 3

[9] Haonan Luo, Guosheng Lin, Zichuan Liu, Fayao Liu, Zhenmin Tang, and Yazhou Yao. SegEQA: Video segmentation based visual attention for embodied question answering. In *ICCV*, 2019. 1

[10] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 1

[11] Andreas Robinson, Felix Järemo Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Learning fast and robust target models for video object segmentation. In *CVPR*, 2020. 2, 3

[12] 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

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

[14] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *ICCV*, 2017. 1

[15] 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. 2

[16] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE TPAMI*, 40(1):20–33, 2017. 1

[17] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *CVPR*, 2019. 1

[18] 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. 3

[19] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *ICCV*, 2019. 2

[20] 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. 2, 3

[21] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. Cascaded human-object interaction recognition. In *CVPR*, 2020. 1