Object-based Spatial Similarity for Semi-supervised Video Object Segmentation

Bofei Wang[1], Chengjian Zheng[1], Ning Wang[1], Shunfei Wang[1], Xiaofeng Zhang[1], Shaoli Liu[1], Si Gao[1], Kaidi Lu[1], Diankai Zhang[1], Lin Shen[1], Yukang Wang[2], Yongchao Xu[2]
[1]ZTE Corporation, [2]Huazhong University of Science and Technology

Abstract

Video object segmentation (VOS) is a fundamental task in computer vision. In this paper, we present a two-stage semi-supervised VOS method, which aims to perform VOS with the annotations of first-frame. In the first stage, we employ a state-of-the-art instance segmentation approach followed by an improved merging method, which takes Object-based Spatial Similarity (OSS) into account as well. In this way, preliminary segmentation of video sequences is generated. In the second stage, we propose a novel Adaptive Reference-frame Selection (ARS) algorithm based on OSS, which could reduce the object-ID mismatching under object occlusion and deformation. With ARS, the reliable reference frame for each object can be selected dynamically and adaptively during the tracking and segmentation process. Then, the preliminary segmentation can be further refined based on the corresponding reference frame. We evaluate the proposed algorithm on the DAVIS 2017 Video Object Segmentation Benchmark and achieve first place on the DAVIS 2019 Semi-Supervised Challenge with a $\mathcal{J}\&\mathcal{F}$ mean score of 76.7%.

1. Introduction

Semi-supervised video object segmentation is a challenging task which aims at tracking and segmenting different objects in the video sequences, according to the given first-frame annotation. Recently, the DAVIS [13, 12, 2, 3] challenge introduces a real-world dataset for VOS with complex scenarios such as small objects, occlusion, deformation and so on.

Many VOS methods have been proposed for the challenge. OSVOS [1] made efforts to segment video frames separately. However, it does not take full advantage of temporal information. MaskTrack [11] exploited optical flow information to propagate the segmentation mask from one frame to the next. Yet, it heavily relies on temporal continuity and suffers from issues like drifting, leading to the inability in handling fast moving objects. Other methods [8, 15, 10] presented at the DAVIS 2018 Challenge [2] have achieved appealing results. DyeNet [8] improved MaskTrack by tracking the target bidirectionally. The method in [15] proposed a category-agnostic method based on spatial information in addition to temporal information. These methods achieve great performance even in some occlusion and deformation scenarios. However, their robustness needs to be strengthened. PReMVOS [10], the winner of DAVIS 2018 Challenge, employed Mask R-CNN [6] for coarse object proposal generation. Then they applied Deeplab v3+ [5] for proposal refinement, followed by merging ReID features [9] and optical flow features [7] to track the objects. Yet, the optimal matching weights involved in the merge module are obtained by traversing parameter combinations, which makes it less robust. Besides, object-ID mismatching is often occurred especially in the presence of occlusion or deformation, since the object is always matched with the first frame when using ReID vectors.

To solve the above problems, we present a semi-supervised VOS method with Object-based Spatial Similarity (OSS). Our contributions are summarized as follows:

- We improve the original PReMVOS by utilizing OSS in the proposal merging step. For that, quantified color histogram features are taken into account. In this way, our method achieves better performance than the PReMVOS even using the same weight, without traversing parameter combinations.

- Adaptive Reference-frame Selection (ARS) algorithm based on OSS is proposed to adaptively select the reliable reference-frame for each object. Then the first frame is replaced by the selected frame to re-identify each object, yielding improved performance for occlusion and deformation situations.

2. Approach

Our proposed method is derived from the PReMVOS pipeline, which consists of four different neural networks to generate proposals and merge them into pixel-wise object tracks. Figure 1 illustrates our two-stage algorithm. Specifically, we first improve the PReMVOS to generate

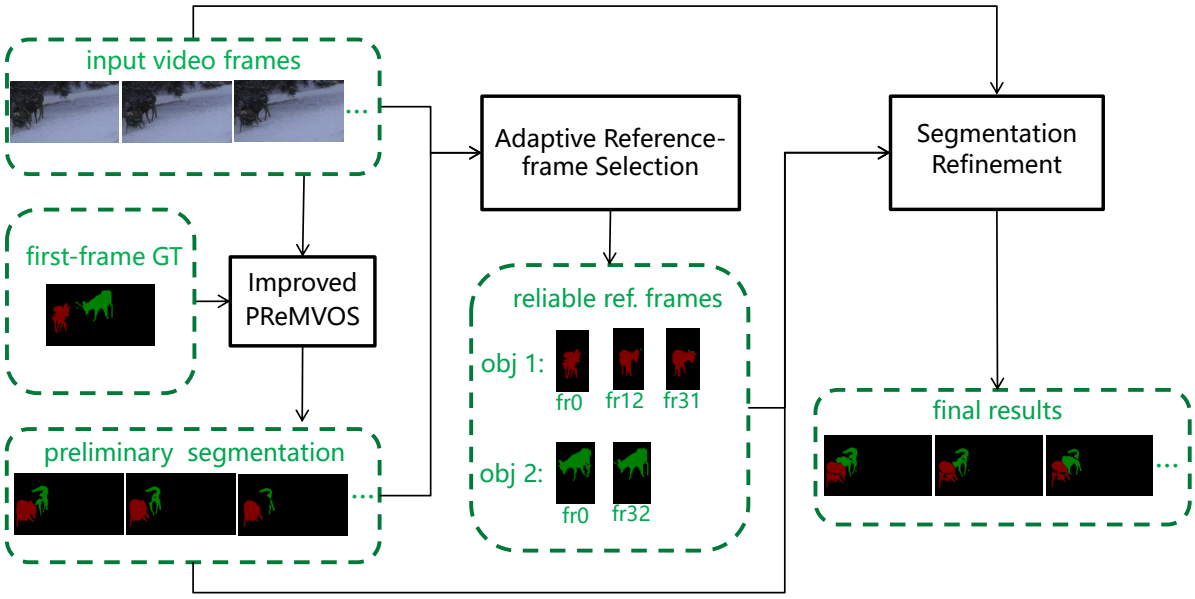

Figure 1. Pipeline of our approach. Given the video frames and the ground truth (GT) of the first frame, an improved PReMVOS is applied to obtain preliminary segmentation, followed by segmentation refinement via a novel Adaptive Reference-frame Selection.

preliminary segmentation by utilizing Object-based Spatial Similarity (OSS) and then refine the segmentation via a novel Adaptive Reference-frame Selection (ARS). We present these two stages in 2.1 and 2.2, respectively.

2.1. Preliminary Segmentation

In the first stage, preliminary segmentation is obtained based on the improved PReMVOS. We modify the original PReMVOS in two aspects, which are detailed as follows.

2.1.1 Proposal Generation

Instead of Mask R-CNN in the original PReMVOS, Cascaded Mask R-CNN [4] with a ResNeXt101 [14] backbone is adopted in our proposal generation step. As shown in Table 1, the Cascaded Mask R-CNN provides better detections than the original Mask R-CNN.

2.1.2 Proposal Merging

The proposal merging algorithm is also improved by utilizing object-based spatial similarity. For that, color histogram features of the mask regions are taken into account. We first calculate quantified color histogram for each segmentation proposal and then obtain the color histogram feature sub-score according to following expression:

$$S_{hist}(h_{gt}^{i}, h_{t}^{j}) = \sqrt{h_{gt}^{i} \cdot h_{t}^{j}}, \tag{1}$$

where h_{gt}^{i} denotes the color histogram feature for object i in the first frame and h_{t}^{j} denotes the color histogram feature for object j in another frame t. In fact, $S_{hist}(h_{gt}^{i}, h_{t}^{j})$ can be regarded as the spatial similarity between objects.

Finally, the color histogram feature sub-score and other sub-scores are combined together to merge the proposals and track the objects as depicted in [10].

2.2. Segmentation Refinement

The first frame is always the reference while the current frame could be far from it, leading to poor performance especially under object occlusion or deformation. Therefore, in the second stage, the preliminary segmentation is further refined. A novel Adaptive Reference-frame Selection (ARS) algorithm is proposed and the reference-frame of each object can be selected adaptively.

2.2.1 Adaptive Reference-frame Selection

Adaptive Reference-frame Selection (ARS) algorithm based on Object-based Spatial Similarity (OSS) is presented in Algorithm 1. We first calculate the color histogram h_{gt}^{i} for each ground truth object i in the first frame. Then the same process is applied to the other frames to obtain h_{t}^{j}. Given h_{gt}^{i} and h_{t}^{j}, the similarity of the same object is calculated according to Eq (1). If the similarity is greater than T_s and the interval of adjacent reference-frames is longer than T_t, the frame will be selected for the corresponding object.

2.2.2 Template Update

With the reliable reference frame of each object, the refined segmentation can be easily generated via template update. The process is detailed in Algorithm 2. Another proposal merging is proposed and the template is updated according to the reference frame of each object. As shown

	gts	dets	recall	precision	ap
Mask R-CNN	3929	13750	0.754	0.215	0.641
Cascaded Mask R-CNN	3929	10319	0.768	0.293	0.737

Table 1. Performance with different proposal generation networks on the DAVIS 2017 val set.

(a) Segmentation results without ARS (b) Segmentation results with ARS

Figure 2. (a) Segmentation results without ARS. Objects in red boxes are mismatched due to large occlusion and deformation. (b) Segmentation results with ARS. Reference frames are selected adaptively except for the first frame, giving rise to better performance.

Algorithm 1: Adaptive Reference-frame Selection.

Input: I, J: objects in the first frame and another frame; N: total frame number; T_s: similarity threshold; T_t: frame interval threshold.
Output: R: reference-frame of each object

function ARS()
  $t_0 = 1$
  for t *from* 1 *to* N **do**
    if $t = 1$ **then**
      foreach i *in* I **do**
        calculate h^i_{gt}
    else
      foreach $j = i$ *in* J **do**
        calculate h^j_t and $S_{hist}(h^i_{gt}, h^j_t)$
        if $S_{hist}(h^i_{gt}, h^j_t) > T_s$ *and* $t - t_0 > T_t$ **then**
          $R(i) \leftarrow R(i) \cup t, t_0 = t$
  return R

Algorithm 2: Template Update.

Input: I: objects in the first frame; N: total frame number; R: reference-frame of each object i.
Output: $Template$: template of each object

function Template_update()
  for t *from* 1 *to* N **do**
    foreach i *in* I **do**
      if $t \in R(i)$ **then**
        $Template(i) = h^i_t$
  return $Template$

in Figure 2, the qualitative results with and without the update of reference frame demonstrate that the proposed ARS effectively eliminates the object-ID mismatching and obtains more accurate segmentation.

3. Experiments

We evaluate our algorithm on the DAVIS 2017 dataset, which contains 150 high-quality video sequences with all frames annotated with pixel-wise object masks. It is worth to note that only annotations of the 90 train/val video sequences are publicly available. The evaluation metric is the J-score, calculated as the average IoU between the proposed masks and the ground truth mask, and the F score, calculated as an average boundary similarity measure between the boundary of the proposed masks and the ground truth masks, and their average value.

3.1. Experimental Results

As shown in Table 2, where rankings in each categories are placed in parentheses, the proposed approach achieves a $\mathcal{J}\&\mathcal{F}$ mean score of 76.7% on DAVIS 2017 test-challenge set and ranks first place on DAVIS 2019 Semi-Supervised Challenge. Qualitative results in Figure 2 also show that the proposed method performs better than other VOS methods in predicting accurate mask proposals, especially under occlusion and deformation.

3.2. Ablation Study

We also study the contribution of the two components, both of which rely on OSS. As shown in Table 3, the quantified color histogram features in proposal merging lead to a 1.2% improvement. The performance is further boosted by 3.1% using segmentation refinement with ARS algorithm.

Team	Global	Region J			Boundary F		
	Mean	Mean	Recall	Decay	Mean	Recall	Decay
Ours	**0.767 (1)**	0.727(2)	0.815 (3)	0.195 (3)	**0.806 (1)**	0.873 (2)	0.220 (3)
Jono	0.762 (2)	**0.729 (1)**	**0.817 (1)**	0.163 (2)	0.794 (2)	0.867 (3)	0.195 (2)
HCMUS	0.754 (3)	0.724 (4)	0.817 (2)	**0.110 (1)**	0.784 (3)	**0.876 (1)**	**0.129 (1)**
swoh	0.752 (4)	0.726 (3)	0.810 (4)	0.212(5)	0.777 (4)	0.849 (4)	0.245 (5)
H2VISION	0.731 (5)	0.701 (5)	0.773 (6)	0.248 (8)	0.761 (5)	0.840 (6)	0.283 (9)
savor-123	0.713 (6)	0.677 (7)	0.748 (7)	0.247(7)	0.750 (6)	0.812 (7)	0.275 (8)
dolfers	0.706 (7)	0.685 (6)	0.781 (5)	0.203 (4)	0.728 (7)	0.842 (5)	0.240 (4)
ByteCV	0.692 (8)	0.660 (8)	0.734 (9)	0.285 (10)	0.723 (8)	0.804 (8)	0.311(10)
sourf	0.689 (9)	0.659 (9)	0.748 (8)	0.214 (6)	0.719 (9)	0.800 (9)	0.256 (6)
AGAMErs	0.643 (10)	0.612 (10)	0.694 (10)	0.264(9)	0.674 (10)	0.777 (10)	0.273 (7)

Table 2. Performance comparison of different methods on DAVIS 2019 Semi-Supervised Challenge.

Approach	**score**	**boost**
PReMVOS (proposal generation with Cascade Mask-RCNN)	0.691	-
+ quantified color histogram features in proposal merging	0.703	1.2%
+ segmentation refinement with ARS algorithm	0.734	3.1%

Table 3. Ablation study on DAVIS 2017 test-dev set.

4. Conclusion

In this paper, we have presented a method for semi-supervised VOS by exploiting object-based spatial similarity. First, the preliminary segmentation is obtained by an improved PReMVOS, which merges proposals with additional color histogram features. Then, the segmentation is refined by a novel Adaptive Reference-frame Selection (ARS). Experiments are conducted on DAVIS 2017 test-challenge set. The proposed method achieves superior performance, ranking first place on DAVIS 2019 Semi-Supervised Challenge.

References

[1] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, pages 221–230, 2017. 1

[2] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *CoRR*, abs/1803.00557, 2018. 1

[3] 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. *CoRR*, abs/1905.00737, 2019. 1

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018. 2

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 1

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 1

[7] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 2462–2470, 2017. 1

[8] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *ECCV*, pages 90–105, 2018. 1

[9] Xiaoxiao Li, Yuankai Qi, Zhe Wang, Kai Chen, Ziwei Liu, Jianping Shi, Ping Luo, Xiaoou Tang, and Chen Change Loy. Video object segmentation with re-identification. In *CVPR Workshops*, 2017. 1

[10] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for the davis challenge on video object segmentation 2018. In *CVPR Workshops*, 2018. 1, 2

[11] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, pages 2663–2672, 2017. 1

[12] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 1

[13] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *CoRR*, abs/1704.00675, 2017. 1

[14] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. 2

[15] Shuangjie Xu, Linchao Bao, and Pan Zhou. Class-agnostic video object segmentation without semantic re-identification. In *CVPR Workshops*, 2018. 1