A semantics-guided warping for semi-supervised video object instance segmentation

Q.F. Wang, Lu Zhang, K. Kpalma

A semantics-guided warping for semi-supervised video object instance segmentation

Qiong WANG[1,2], Lu ZHANG[2], and Kidiyo KPALMA[2]

[1] College of Computer Science and Technology, Zhejiang University of Technology, 310023 Hangzhou, China
[2] Univ Rennes, INSA Rennes, CNRS, IETR (Institut d'Electronique et de Télécommunication de Rennes) - UMR 6164, F-35000 Rennes, France

Abstract. In the semi-supervised video object instance segmentation domain, the mask warping technique, which warps the mask of the target object to flow vectors frame by frame, is widely used to extract target object. The big issue with this approach is that the generated warped map is not always of high accuracy, where the background or other objects may be wrongly detected as the target object. To cope with this problem, we propose to use the semantics of the target object as a guidance during the warping process. The warping confidence computation firstly judges the confidence of the generated warped map. Then a semantic selection is introduced to optimize the warped map with low confidence, where the target object is re-identified using the semantics-labels of the target object. The proposed method is assessed on the recently published large-scale Youtube-VOS dataset and compared to some state-of-the-art methods. The experimental results show that the proposed approach has a promising performance.

Keywords: Warping flow · Semantics · Semi-supervised video object instance segmentation.

1 Introduction

Video object segmentation aims to segment objects from backgrounds. It assigns object IDs to the pixels belonging to objects, and assigns 0 values to other pixels. It has numerous applications in autonomous driving, video surveillance, object recognition, etc.

According to the object to be segmented, video object segmentation can be roughly classified into three categories: video foreground segmentation, video semantic object segmentation and video object instance segmentation. Video foreground segmentation aims at segmenting all probable objects. For real-world scenes, the detected region may contain multiple objects. Decomposing the detected region into different objects is more meaningful and is better for video understanding. Video semantic object segmentation segments the region based on the semantic label. The objects belonging to the same semantic label are grouped together. In the output map of video object instance segmentation, the

pixels are grouped into multiple sets and assigned to consistent object IDs. Pixels within the same set belong to the same object.

Video object instance segmentation attracts more interests and has not been fully investigated. One popular way for video object instance segmentation is called as Semi-supervised video object segmentation. Human-guidance is adopted to define the objects that people want to segment. It is usually delineated in the frame that the object appears in the first time. By propagating the manual labels to the rest of the video sequence, the object instance is segmented in the whole video sequences. Semi-supervised video object segmentation can be regarded as a tracking problem but with the mask output. This paper focuses on semi-supervised video object instance segmentation.

For semi-supervised video object segmentation based on the human-guidance, one challenge is how to segment a pre-defined object in a video based on its provided mask of the frame in which the object appears at the first time. An initial way for semi-supervised video object segmentation is to firstly train the parent network which detects all foreground objects (also called as offline learning), secondly fine-tune the parent network for the particular object using the manual label (also called as online learning), as in state-of-the-art method Segflow [2]. However, it is very time-consuming. Segmenting the target object just from each static frame is not sufficient.

Most works adopt "Mask warping", which combines the necessary appearance information and the temporal context together, to generate the warped map. "Mask warping" is faster than online learning, which benefits the video object segmentation. However, the warped map generated in this way is vulnerable to lighting changes, deformations, etc. The wrongly detected regions in one frame can be propagated to the following ones, thus more background is warped.

The semantics label of the object instance in the first frame is another useful cue for semi-supervised video object segmentation. In the method [11], a semantics instance segmentation algorithm is leveraged to obtain the semantics label of the target object in the first frame, and then the semantics label is propagated to the following frames. In the method [6], objects are divided into human and non-human object instances which are propagated using different networks.

Mask warping and semantics label guidance are not mutually exclusive, and could be taken simultaneously. Few studies combine the advantages of two aforementioned cues. In order to take merits of mask warping and semantics label guidance, a novel semi-supervised video object instance segmentation is proposed with following contributions:

- we propose a new method, named Warping Confidence Computation (WCC), to differentiate the warped maps by classifying them into low-confidence or high-confidence.
- Semantics selection is introduced when a low-confidence warped map is detected. With the semantics label, the optimized warped map is generated through re-identifying the target object. Different from [11], the temporal information (optical flow) is also used for the optimization of the mask warping.

The remaining of this paper is organized as follows. Section 2 introduces an overview of state-of-the-art methods. Section 3 presents the proposed method in detail. In section 4, we show and discuss the performance of the proposed method. Section 5 concludes the paper.

2 An overview of state-of-the-art methods

Recent works are introduced based on the way to use the human-guidance.

2.1 Online-offline learning

The methods [1, 12] employ the combination of offline and online learning strategies. Caelles *et al.* [1] design a network to learn the foreground object, which consists of a foreground branch and a contour branch. Compared with OSVOS [1], OnAVOS [16] updates the result based on online selected training example. It aims at adapting the changes in appearance. Cheng *et al.* [2] propose a network which has two branches: the segmentation branch and the flow branch to predict the foreground objects. MaskTrack [12] predicts the segmentation mask with a rough estimated map of the previous frame.

2.2 Mask warping

The initial way of mask warping is to directly warp the mask of the target object to the optical flow vectors to generate warped map frame by frame [8, 7, 19, 15, 17]. Leibe *et al.* [9] propose to optimize the generated warped map in each step with an objectness score; Khoreva *et al.* [5] propose to optimize the generated warped map by removing the possibly spurious blobs.

3 Proposed method

The proposed algorithm (semantic-guided warping for semi-supervised video object segmentation (SWVOS)) consists of three main steps: (1) according to the provided pixel-wise mask of the first frame, target object is firstly segmented using mask warping technique, where warped maps are generated; (2) the warping confidence is computed for each warped map, which is then divided into high-confidence map and low-confidence map; (3) the warped map with high-confidence is directly used as the final segmentation map, while the low-confidence warped map is optimized using a semantics selection. The block-diagram of the proposed algorithm is shown in Fig.1.

3.1 Mask warping

The optical flow vectors between pairs of successive frames are generated using the Flownet [4]. Then the warped map of each frame is obtained by warping the

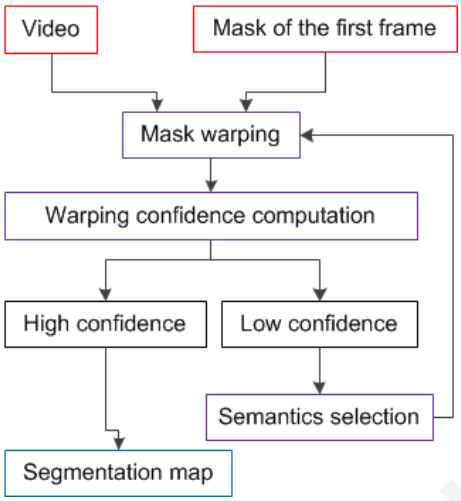


Fig. 1. The proposed block-diagram SWVOS.

proposal of the previous frame to the optical flow vector. The warping function is defined as:

$$f_j = \omega(f_i, V_{i \to i+1}) \tag{1}$$

where f_j denotes the warped map of the frame j, ω is the bilinear warping function, f_i denotes the warped map of the previous frame i (for the first frame, the proposal is the provided mask), $V_{i \to j}$ is the optical flow vectors between pairs of successive frames i and j.

3.2 Warping confidence computation

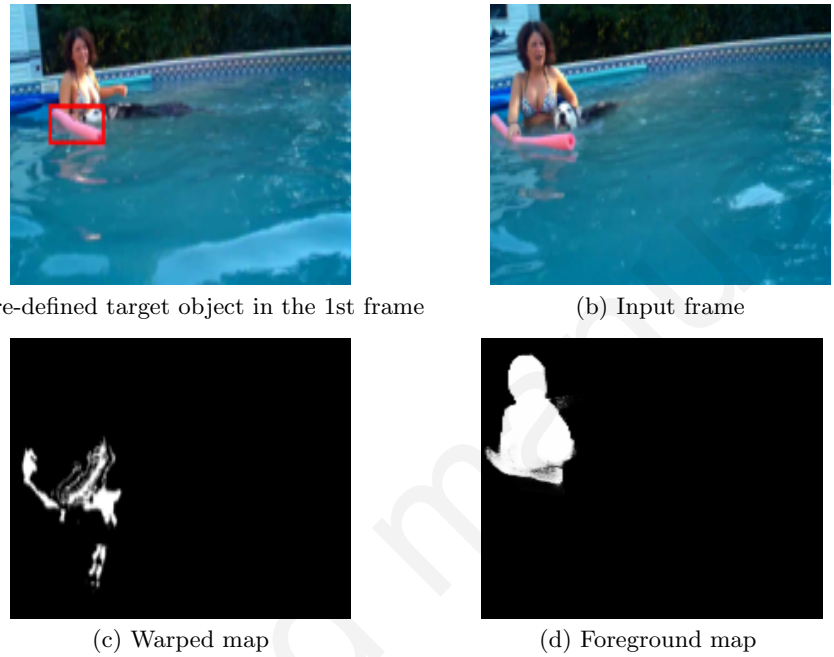
For the generated warped map, overlap ratio and contiguous groups number are used for warping confidence computation (WCC). Overlap ratio (OR) is the ratio of the object that belongs to the warped map (WM) and the foreground map (FM), the larger is better.

$$\mathrm{OR} = \frac{|\mathrm{WM} \bigcap \mathrm{FM}|}{|\mathrm{FM}|} \tag{2}$$

Contiguous groups number (CGN) is the number of contiguous regions in the warped map, the smaller is better. The warped map with a low OR value or a high CGN is regarded as low-confidence in the WCC.

The foreground map (FM) is obtained with a fully covolutional network (FCN), which is a modified NLDF (Non-Local Deep Features) network [10]. Our FCN differs from the NLDF [10] in that (1) the NLDF resizes the input image to a fixed size while our FCN uses it with its original size; (2) the NLDF adopts the VGG [14] as the baseline and uses the output of the 5-th block in the VGG as

the global feature, while our FCN removes this global feature which may bring noises for complex scenes; (3) the NLDF uses the cross entropy loss and the boundary IOU loss for training while our FCN only uses the cross entropy loss since our pre-experiment showed that the boundary IOU loss does not influence a lot our method's performances.



(a) Pre-defined target object in the 1st frame (b) Input frame

(c) Warped map (d) Foreground map

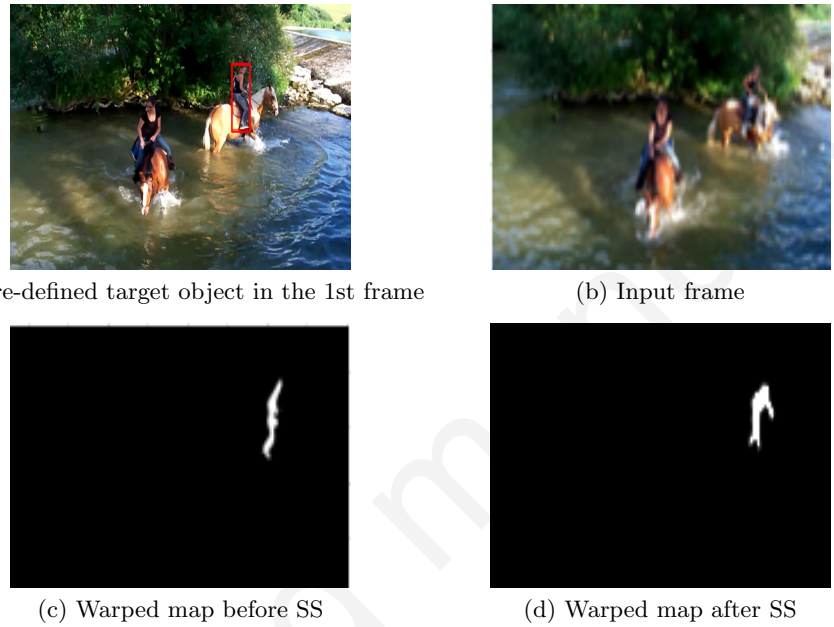
Fig. 2. One example of the warping confidence computation. The target object is denoted in red box in (a).

One example of the WCC is given in Fig.2. In this example, we can see that the warped map not only contains many contiguous groups, but also has low overlap region with the foreground map. Thus, it is judged to be a warped map with low-confidence. In this paper, the threshold for the OR is just set to be a small number 0.001. The threshold for the CGN is set to be 10, i.e. about five times of the average number of objects in each frame in the video sequence.

3.3 Semantics selection

The warped map with low-confidence is optimized using semantic selection (SS) as following. Firstly, the semantic label of the target object in the first frame is detected using the MASK R-CNN [3]. Secondly, for the frame with low-confidence warped map, semantics of all objects are detected using the MASK

R-CNN. Thirdly, the object in the frame that satisfies two conditions is segmented to generate the optimized warped map: (1) the object has the same semantic label as the target object, (2) the object is the closest one to the center of gravity of the low-confidence warped map. Here the MASK R-CNN is fine-tuned with the YouTube-VOS-train dataset [18] in order to recognize categories in this dataset. An example is given in Fig.3.



(a) Pre-defined target object in the 1st frame (b) Input frame

(c) Warped map before SS (d) Warped map after SS

Fig. 3. Example of semantics selection (SS). The target object is denoted in red box in (a).

For a video sequence with multiple pre-defined objects, target objects are detected separately, and then merged together to generate the final segmentation map. If the pixel is detected belonging to multiple target objects, it is set to the one that has the smallest size in the provided manual labels in the first frame.

4 Experiments and analysis

This section firstly introduces the used dataset and metrics, and then shows the performance of our approach.

4.1 Datasets

The YouTube-VOS dataset [18] is a recently published and the largest dataset with high resolution for semi-supervised video object segmentation. It is the most

challenging dataset, and it contains three sets: Train, Validation and Test. It has the total number 197,272 of object annotations. For the Test set, it contains 508 video sequences with the first-frame ground truth provided. 65 categories of objects in the Test set appear in Train set, which are called as "seen objects"; and 29 categories of objects in the Test set do not appear in Train set, which are called as "unseen objects".

4.2 Evaluation metrics

For semi-supervised video object segmentation, Region Similarity J and Contour Accuracy F [13] are used to measure the similarity between the generated segmentation map (M) and the ground truth (GT). Region Similarity J is defined as the intersection-over-union of M and GT. Contour Accuracy F is computed by the contour-based precision Pc and recall Rc.

$$J = \frac{|M \bigcap \mathrm{GT}|}{|M \bigcup \mathrm{GT}|} \quad F = \frac{2P_cR_c}{P_c + R_c} \tag{3}$$

A larger J value and a larger F value mean a better performance. For the overall evaluation, the final measure is the average of four scores: J for seen categories, J for unseen categories, F for seen categories and F for unseen categories.

4.3 Results and discussions

Table 1 compares our proposed method with the state-of-the-art methods. We can see that the proposed method achieves better overall performance than Segflow [2] on the YouTube-VOS-test dataset. We must note that the compared methods OSVOS, OnAVOS and MaskTrack perform better than our proposed method. However they all use the time-consuming online learning step, which is not suitable for real-world applications. Our proposed method has not this limitation.

Table 1. Performance comparison between the proposed method (SWVOS) and existing models over the YouTube-VOS-test dataset. The best score is in **bold**.

Methods	J_seen↑	J_unseen↑	F_seen↑	F_unseen↑	Overall↑
OnAVOS	0.557	0.568	0.613	0.623	0.590
MaskTrack	0.569	**0.607**	0.593	0.637	0.602
OSVOS	**0.591**	0.588	**0.637**	**0.639**	**0.614**
SWVOS	0.513	0.367	0.494	0.419	0.448
Segflow	0.404	0.385	0.350	0.327	0.367

Fig.4 shows some segmentation maps generated by the proposed approach.

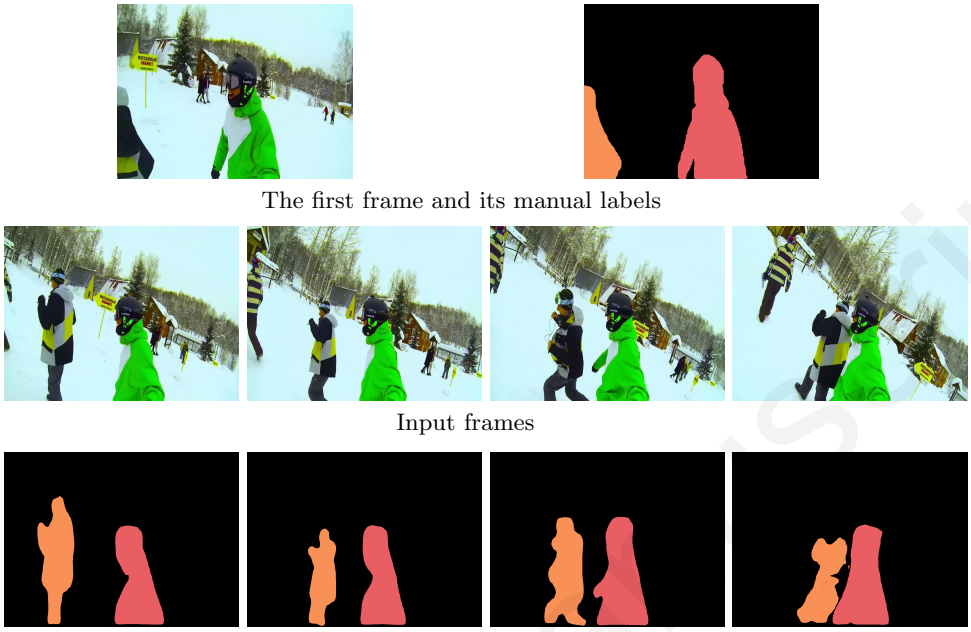


Fig. 4. Some examples of segmentation maps generated by proposed SWVOS.

For the semi-supervised video object segmentation task, the YouTube-VOS Challenge on video object segmentation 2018 uses YouTube-VOS-test dataset for competition. Our method achieves the 8th result in YouTube-VOS Challenge on video object segmentation 2018. In Table 2, we show the performance of our proposed models (named "SnowFlower") in the benchmarking table. Note that only 8 models are selected and listed.

Table 2. Performance benchmarking in the YouTube-VOS Challenge.

Team Name	Overall	J_seen	J_unseen	F_seen	F_unseen	Rank
Jono	0.722(1)	0.737(1)	0.648(2)	0.778(1)	0.725(2)	1st
speeding_zZ	0.720(2)	0.725(3)	0.663(1)	0.752(3)	0.741(1)	2nd
mikirui	0.699(3)	0.736(2)	0.621(4)	0.755(2)	0.684(4)	3rd
hi.nine	0.684(4)	0.706(5)	0.623(3)	0.728(5)	0.677(5)	4th
sunpeng	0.672(5)	0.707(4)	0.598(6)	0.736(4)	0.648(6)	5th
random_name	0.672(6)	0.672(6)	0.609(5)	0.709(6)	0.697(3)	6th
kduarte	0.539(7)	0.594(7)	0.483(7)	0.578(7)	0.502(7)	7th
SnowFlower	0.448(8)	0.513(8)	0.367(8)	0.494(8)	0.419(8)	8th

5 Conlusion

In this study, we have proposed a novel semi-supervised video object instance segmentation method that extracts each target object from each frame. This goal is achieved by using the mask warping technique. By employing the warping confidence computation, the method can firstly detect the warped map in low-level confidence. Then the optimized warped flow map is achieved through re-identifying the target object with semantics selection. The target object is extracted with better performance.

For the evaluation of video object segmentation, one recently published large-scale dataset: Youtube-VOS is used. Experimental results demonstrate that the proposed method achieves high J value and F value. Our method has not the time-consuming limitation caused by online learning step. Since, the proposed method is a combination of traditional method and deep-learning method, we will further investigate to improve its performance by training a network in an end-to-end way.

6 Acknowledgment

This work was supported in part by the China Scholarship Council (CSC) under Grants 201504490048, in part by National Key Research and Development Program of China (No. 2018YFE0126100).

References

1. Caelles, S., Maninis, K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Gool, L.V.: One-shot video object segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 5320–5329 (2017)
2. Cheng, J., Tsai, Y., Wang, S., Yang, M.: Segflow: Joint learning for video object segmentation and optical flow. In: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. pp. 686–695 (2017)
3. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell. **42**(2), 386–397 (2020)
4. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 1647–1655 (2017)
5. Khoreva, A., Benenson, R., Ilg, E., Brox, T., Schiele, B.: Lucid data dreaming for video object segmentation. Int. J. Comput. Vis. **127**(9), 1175–1197 (2019)
6. Le, T.N., Nguyen, K.T., Nguyen-Phan, M.H., Ton, T.V., Nguyen, T.A., Trinh, X.S., Dinh, Q.H., Nguyen, V.T., Duong, A.D., Sugimoto, A., Nguyen, T.V., Tran, M.T.: Instance re-identification flow for video object segmentation. The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2017)
7. Li, X., Loy, C.C.: Video object segmentation with joint re-identification and attention-aware mask propagation. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III. pp. 93–110 (2018)

8. Li, X., Qi, Y., Wang, Z., Chen, K., Liu, Z., Shi, J., Luo, P., Tang, X., Loy, C.C.: Video object segmentation with re-identification. The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2017)
9. Luiten, J., Voigtlaender, P., Leibe, B.: Premvos: Proposal-generation, refinement and merging for video object segmentation. In: Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part IV. pp. 565–580 (2018)
10. Luo, Z., Mishra, A.K., Achkar, A., Eichel, J.A., Li, S., Jodoin, P.: Non-local deep features for salient object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 6593–6601 (2017)
11. Maninis, K., Caelles, S., Chen, Y., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Gool, L.V.: Video object segmentation without temporal information. IEEE Trans. Pattern Anal. Mach. Intell. **41**(6), 1515–1530 (2019). https://doi.org/10.1109/TPAMI.2018.2838670, https://doi.org/10.1109/TPAMI.2018.2838670
12. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 3491–3500 (2017)
13. Perazzi, F., Pont-Tuset, J., McWilliams, B., Gool, L.V., Gross, M.H., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 724–732 (2016)
14. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
15. Sun, J., Yu, D., Li, Y., Wang, C.: Mask propagation network for video object segmentation. The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2018)
16. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. In: British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017 (2017)
17. Xiao, H., Feng, J., Lin, G., Liu, Y., Zhang, M.: Monet: Deep motion exploitation for video object segmentation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 1140–1148 (2018)
18. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.S.: Youtube-vos: A large-scale video object segmentation benchmark. CoRR **abs/1809.03327** (2018)
19. Xu, S., Bao, L., Zhou, P.: Class-agnostic video object segmentation without semantic re-identification. The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops (2018)