Joint Multi-object Detection and Segmentation from an Untrimmed Video

Xinling Liu[1], Le Wang[1](✉), Qilin Zhang[2], Nanning Zheng[1], and Gang Hua[3]

[1] Xi'an Jiaotong University, Xi'an 710049, Shannxi, People's Republic of China
lewang@xjtu.edu.cn
[2] HERE Technologies, Chicago, IL 60606, USA
[3] Wormpex AI Research, Beijing 100028, People's Republic of China

Abstract. In this paper, we present a novel method for jointly detecting and segmenting multiple objects from an untrimmed video. Unlike most existing video object segmentation methods that can only handle a trimmed video in which all video frames contain the target objects, we address a more practical and difficult problem, *i.e.*, joint multi-object detection and segmentation from an untrimmed video where the target objects do not always appear per frame. In particular, our method consists of two modules, *i.e.*, object decision module and object segmentation module. The object decision module is used to detect the objects and decide which target objects need to be separated out from video. As there are usually two or more target objects and they do not always appear in the whole video, we introduce the data association into object decision module to identify their correspondences among frames. The object segmentation module aims to separate the target objects identified by object decision module. In order to extensively evaluate the proposed method, we introduce a new dataset named UNVOSeg dataset, in which 7.2% of the video frames do not contain objects. Experimental results on four datasets demonstrate that our method outperforms most of the state-of-the-art approaches.

Keywords: Video object segmentation · Data association · Object detection

1 Introduction

Video object segmentation aims at segmenting the primary objects from the background across all frames. It is a fundamental yet important task in computer vision, which can be used in many applications, such as autonomous driving, video surveillance, action recognition. In this paper, our objective is to separate two or more target objects and link the object instances for each target object across the whole video, simultaneously.

Despite the success of image object segmentation with the development of convolutional neural networks in recent years, it is still challenging when it comes

Published by Springer Nature Switzerland AG 2020
I. Maglogiannis et al. (Eds.): AIAI 2020, IFIP AICT 583, pp. 317–329, 2020.
https://doi.org/10.1007/978-3-030-49161-1_27

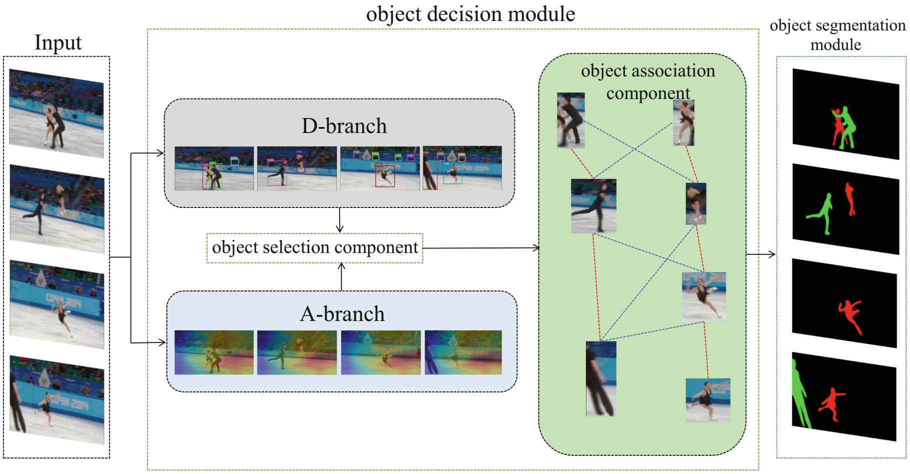
object decision module
Input
object segmentation module
D-branch
object association component
object selection component
A-branch

Fig. 1. Illustration of the proposed multiple video object detection and segmentation method.

to video object segmentation. Training a video object segmentation network requires a large amount of annotated video frames, which is very expensive and time-consuming. This invisibly increases the difficulty of video object segmentation. The video object segmentation methods can be roughly divided into two categories, *i.e.*, semi-supervised methods [1–7] and unsupervised methods [8–10]. Semi-supervised methods need some extra user annotations (*e.g.*, segmentation mask for the first frame) to indicate which object(s) need to be separated; while unsupervised methods automatically extract the primary object(s).

Although many methods have achieved significant performances on standard benchmarks [11–15], there are still several problems that need to be further addressed. For example, previous methods [16–20] are more suitable for trimmed videos that the target objects exist in (almost) all frames; while for untrimmed videos, the objects disappear intermittently throughout the whole video. Previous video object segmentation methods for trimmed videos cannot be directly leveraged to handle the untrimmed videos with noisy frames not containing the target objects. Taking the above issues into account, we propose a multiple video object detection and segmentation method to separate the target objects from an untrimmed video, which consists of two modules, *i.e.*, object decision module and object segmentation module. Figure 1 illustrates the proposed multiple video object detection and segmentation method.

The object decision module is designed to detect the objects and decide which target objects need to be separated out from the video, and then identify their correspondence across adjacent frames. The object segmentation module is built to segment out the target objects identified by the object decision module. Specifically, we first feed an untrimmed video into the object decision module, which contains two components, *i.e.*, object selection and object association. In object selection, we combine the object detector and an attention mechanism through a gating mechanism to make the network focus on the main objects denoted by bounding boxes in the video. In object association, we take the bounding boxes between adjacent frames as input to train a Siamese network, and then obtain the correspondences of these objects throughout the whole video. As a result, the object decision module can output the bounding boxes of multiple objects and their associations across the whole video. Then, we propose a weakly supervised segmentation method as our object segmentation module, and the objective is to segment the target objects from these bounding boxes generated by the object decision module. In detail, we first use multi-scale combinatorial grouping (MCG) [21] to generate a set of region proposals as pseudo-ground truth, as as to train a segmentation network which is employed to produce the final segmentation results.

We conduct extensive experiments on four datasets, including SegTrack [22], DAVIS2016 [23], DAVIS2017 [24], and our UNVOSeg. Experimental results show that our method significantly outperforms most of the state-of-the-art approaches, which validates that it can jointly detect and segment multiple objects from an untrimmed video.

The key contributions of this paper are summarized as follows:

- We present a joint multi-object detection and segmentation method, which is able to segment the primary objects in an untrimmed video in which multiple objects co-exist but disappear intermittently throughout the video.
- We propose a object decision module to identify the target objects and their correspondence across the untrimmed video.
- We establish a noisy video object segmentation dataset, named UNVOSeg dataset, including 63 untrimmed long videos and per-frame ground truth segmentation annotations, to verify the proposed method.

2 Related Work

Since our work addresses the problem of automatically segmenting multiple objects from an untrimmed video, we briefly review recent work on semi-supervised and unsupervised video object segmentation.

2.1 Semi-supervised Video Object Segmentation

Given some user annotations in the first frame or several frames, semi-supervised video object segmentation aims at segmenting out the target objects from a video. The research work in this line can benefit from the appearance model initialized by the segmentation mask annotations. Varun *et al.* [2] proposed a bilateral network followed by a standard spatial network to propagate information across frames. Sergi *et al.* [3] cast video object segmentation as a per-frame segmentation problem, given the object model from one or more manually segmented frames. Federico *et al.* [4] first trained a propagating network with static images, then took the same technology to fine-tune online learning strategies. Jingchun *et al.* [5] proposed a framework which is trained offline to learn a generic notion, and fine-tuned online for specific objects iteratively. The framework is capable of simultaneously predicting pixel-wise object segmentation and optical flow in videos. Hu *et al.* [6] classified each pixel in succeeding frames, using pixel-wise embeddings learned from supervision in the first frame. Xu *et al.* [7] modified traditional method Multiple Hypotheses Tracking (MHT) to semi-supervised video object segmentation, adapting a combination of mask propagation score and motion score to determine the affinity between hypotheses.

2.2 Unsupervised Video Object Segmentation

Unsupervised video object segmentation aims to discover and segment the most primary objects from the background with only the video as input. Some previous methods tackled this task based on clustering of point trajectories [9,10], motion characteristics [11], appearance [12,13], or saliency [14–16]. Li *et al.* [17] proposed a motion-based bilateral network to filter false positive region, and combined the background with instance embeddings. Liu *et al.* [18] coupled two dynamic Markov Networks to make video object discovery and video object segmentation tasks mutually beneficial. CDTS [19] developed a collaborative detection, tracking, and segmentation technique to extract multiple segment tracks accurately. RVOS [20] proposed a fully end-to-end trainable recurrent network and was the first one to report quantitative results for multiple object video object segmentation on DAVIS2017 [24] and YouTube-VOS [25] benchmarks.

3 Problem Formulation

In this section, we present the details of the proposed method, including two modules, *i.e.*, object decision module and object segmentation module. Given an input video $V = \{v_t\}_{t=1}^{T}$ of T frames, our objective is to identify the target objects $O = \{O_1, \cdots, O_N\}$ that need to be separated out from V, where N is the total number of video objects. Meanwhile produce their segmentation masks $S = \{s_O^t\}_{t=1}^{T}$ along with the correspondences throughout the whole video (*i.e.*, giving an object identity $n \in \{1, \cdots, N\}$ for each segmented object instance). The video frames are first put into the object decision module, which includes an object selection component and an object association component. The object selection component combines object detection, attention mechanism and a gating mechanism together to focus on the target objects that need to be segmented out. Specifically, the object detection is to detect the objects per frame, and then generate bounding boxes $B_t = \{b_k^t\}_{k=1}^{K}$ for each frame v_t, where K is the bounding box number in frame v_t. The attention mechanism is to make the location cues more prominent, and thus we can filter out irrelevant objects using a gating strategy. The bounding boxes of the target objects need to be segmented are then determined as $Q_t = \{q_n^t\}_{n=1}^{N}$. Subsequently, we feed them to the object association component, where we adopt the Siamese network [26] to identify

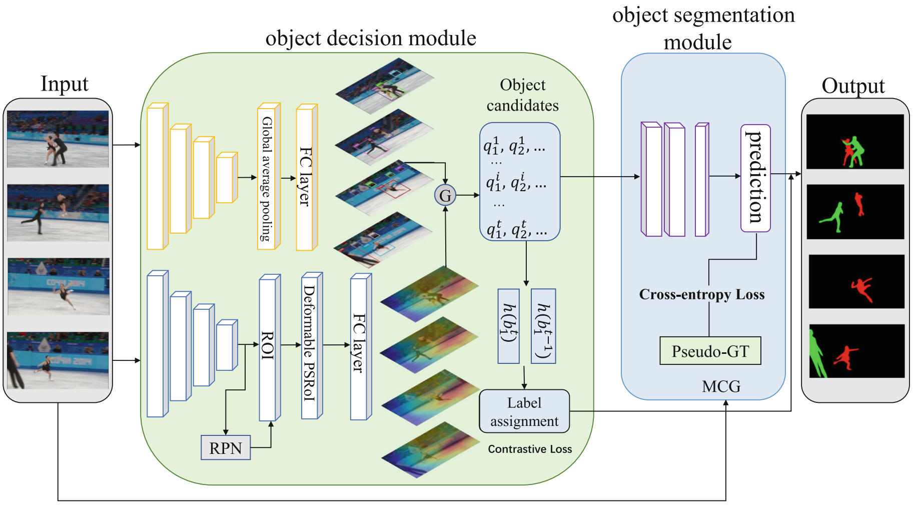

Fig. 2. The framework of our proposed method, which consists of two modules, *i.e.*, object decision module (shown in green) and object segmentation module (shown in blue). The object decision module is to detect the objects and decide which target objects need to be separated out in the video, in which an object association component is leveraged to identify the correspondences of object instances of a same object among frames. The object segmentation module is to segmented out the target objects identified by the object decision module. (Color figure online)

their correspondences between adjacent frames. The object segmentation module proceeds to perform a binary segmentation on bounding boxes of each frame v_t along with the object identities produced by object association component as the final segmentation result. Figure 2 presents the framework of our proposed method.

3.1 Object Decision Module

The purpose of object decision module is to generate reliable bounding boxes that will aid the object segmentation module to segment the primary objects in the video. Taking a video sequence $V = \{v_t\}_{t=1}^{T}$ of T frames as input, we first apply a backbone network ResNet-101 [27] to extract their features F.

Object Detection. Then, we feed F into two branches, *i.e.*, detection-branch (D-branch) and attention-branch (A-branch). In D-branch, we use a RPN network [28] to generate region proposals, and then use a position-sensitive ROI-Pooling [29] followed by a classification and a bounding box regression subnetwork to obtain the bounding boxes of object proposals $B_t = \{b_k^t\}_{k=1}^{K}$ and their detection sores $P_t = \{p_k^t\}_{k=1}^{K}$, where K is the total number of detected objects in frame t. Because video object segmentation is at instance level, we set the category number $C = 1$ in this paper. To enhance the detection capability, we utilize the deformable convolutional network [30] to estimated 2D offsets so that we can change the fixed geometric structure, and in the process of ROI-Pooling, the deformable convolutional network predicts 2D offsets for each filter.

Attention Mechanism and Gating. Since the object detector can detect all the objects in the video, including the noisy objects that do not need to be segmented. To filter out the noisy objects appeared in the video, we use an attention mechanism to find out the region where the target objects appear. Specifically, in A-branch, similar to CAM [31], we add a convolutional layer of size 3 × 3, stride 1, pad 1 with 1024 units, followed by a global average pooling layer and a softmax layer. We utilize a gating mechanism to determine which objects are the target objects that need to be segmented out. The weighted addition is used as the final object selection score $G(b_k^t)$, which can be formulated as follows:

$$G(b_k^t) = \alpha p_k^t + \beta d(b_k^t, R) \tag{1}$$

where $d(b_k^t, R)$ is the distance between the object O_i and region R. Finally, we obtain the bounding boxes of the target objects $Q_t = \{q_n^t\}_{n=1}^N$, where N is the total number of target objects.

Object Association. Since there are more than one object in a video, it is necessary to design an object association method to correlate the object instances of the same object among different frames, *i.e.*, the temporal consistency of each video object. We employ the Siamese network in object association module. In particular, we collect a dataset containing diverse object pairs to improve the robustness of network for discriminating different object pairs, which contains three cases: 1) positive pairs of the same object; 2) negative pairs in the same object category; and 3) negative pairs in different object categories. We take the contrastive loss [32] to supervise the training process. With the filtered objects in every frame, we first rescale them to the same size, then send the objects in adjacent frames into the trained Siamese network to get the similarity score $A_b(box_i^t, box_j^{t+1})$. When an object in the current frame did not find a corresponding object in the previous frame, we need to calculate another similarity score A_v as follows:

$$A_v = \frac{1}{N}\sum_{q=1}^{N} A_b(box_i^t, Z_q) \tag{2}$$

where Z_i denotes the bounding box collection of i-th object. Because A_v use more frame information than A_b, it can achieve better results in object association.

3.2 Object Segmentation Module

After getting the bounding boxes of the target objects need to be segmented out, we try to train a segmentation network by feeding the regions in these bounding boxes as input. In particular, we implement the whole training process in the following two steps:

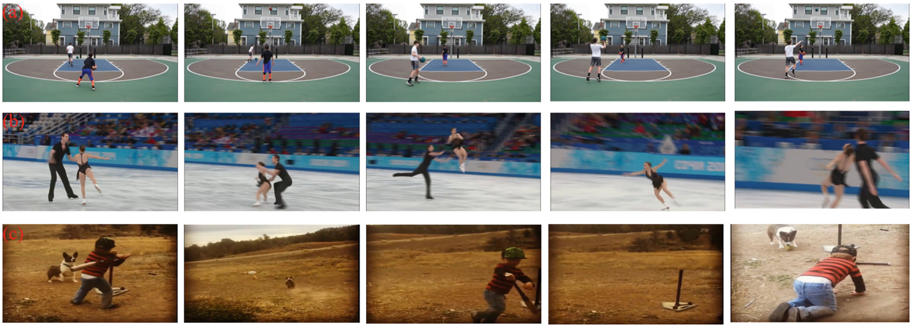

Fig. 3. Some examples of the proposed UNVOSeg dataset: (a) case1: all targets appear in each video frames; (b) case2: one of targets is absent in some video frames; (c) case3: all targets are absent from certain video frames.

Generation of Pseudo Ground Truth. In this step, we generate all the object segmentation candidates from bounding boxes. We adopt a proposal generator, named multi-scale combinatorial grouping (MCG) trained on BSD500 [33] to generate segmentation proposals. Firstly, we use a multiresolution image pyramid to generate edge probability map, which can independently generate the ultrametric contour map (UCM) at each scale, from which we can obtain the connected regions, then a hierarchical segmentation can be acquired by merging these connected regions. In summary, we can get four lists of proposals in each UCM, a total of 16 lists of proposals. After obtaining a complete proposal set, we extract 2D basic features, such as area, perimeter, and boundary strength of each proposal. Finally, we use these features to form a vector to represent these proposals, and then train a random forest regression to rank these proposals.

Object Segmentation. In the second step, we train a standard fully convolutional networks (FCN) [34] by using the segmentation candidates that have the highest overlap rate with bounding boxes as groundtruth. Specifically, for a frame with m bounding boxes, we train a segmentation network with m corresponding pseudo-groundtruth, and the pixels outside the bounding boxes are not calculated.

4 Experiments and Discussions

4.1 Evaluation Datasets

We conduct extensive experiments on four datasets, including SegTrack, DAVIS 2016, DAVIS2017 and our UNVOSeg, to evaluate our method. These datasets are introduced as follows:

SegTrack is a widely used dataset in video object segmentation task. It consists of 14 short videos of 1,080 frames, including 8 single-object videos and 6 multi-object videos with pixel-wise annotations.

Table 1. The statistics of SegTrack, DAVIS2016, DAVIS2017, and UNVOSeg datasets.

Video	SegTrack	DAVIS2016	DAVIS2017	UNVOSeg
Number of videos	14	50	150	63
Number of frames	1080	3455	10474	13129
Noise rate	0.0%	0.90%	0.93%	7.2%

DAVIS2016 is a frequently used dataset, including 50 videos with pixel-wise labels. The goal is to reconstruct real video scenes, such as camera shake, background blur, occlusion, and other complex conditions.

DAVIS2017 is a recently proposed dataset consists of 150 short videos of 10459 frames with multiple objects. The noisy frames only accounts for a small portion of the total video.

UNVOSeg is a newly introduced dataset which can handle the task of multi-object segmentation from an untrimmed video. Because the existing benchmark datasets do not fit our requirement, we establish a new video object segmentation dataset, named UNVOSeg dataset, which meets the following characteristics: 1) The background of the video is more complicated; 2) There are frames that not containing the objects, which is referred as noisy frames; 3) There are cases where the video objects disappear intermittently throughout the video. Figure 3 shows some examples of the proposed UNVOSeg dataset. We collect 63 videos of 13,129 frames from YouTube that satisfies the above criteria, and provide pixel-level annotations for each frame. Considering that most of the videos in existing benchmark datasets contain dozens of frames, the videos in our new dataset contain about 200 to 300 frames. Note that the noisy rate of our new dataset is significantly higher than SegTrack, DAVIS2016, and DAVIS2017 datasets. The statistics of the four datasets are summarized in Table 1.

4.2 Evaluation on Single-Object Datasets

Although the proposed method is designed to solve multi-object detection and segmentation, it can still be applied to single-object video segmentation task. Therefore, we compare our method with other unsupervised single video object segmentation approaches, including FST [11], KEY [12], FSEG [15], SFM [35], UOVOS [36], SAL [37], JOS [18], CVOS [38], TRC [10], and LMP [39]. On SegTrack and DAVIS2016 datasets. The results are evaluated by using J-measure, which are are presented in Table 2 and Table 3, respectively.

Table 2. The comparison results of our method with other competing methods on SegTrack by using J-measure.

Method	FST	KEY	FSEG	UOVOS	JOS	Ours
J-measure	53.5	57.3	61.4	64.3	80.6	51.1

Table 3. The comparison results of our method with other competing methods on DAVIS2016 by using J-measure.

Method	SAL	CVOS	TRC	SFM	LMP	FST	Ours
J-measure	42.6	51.4	50.1	53.2	69.7	57.5	48.3

From the results, we can see that our method does not obtain the best results, which is most likely because we use pseudo-ground truth in the object segmentation module. Nevertheless, our method still outperforms SAL by a margin of 5.7%, and the comparison with other methods is also competitive. Besides, we notice that JOS and LMP achieve best results on SegTrack and DAVIS2016, respectively. In JOS, two coupled dynamic Markov networks are utilized so that video object discovery and video object segmentation tasks can facilitate each other, while LMP exploit optical flow to assist in segmenting moving objects, yet these information are not used in this paper. It is worth mentioning that our method outperforms all other state-of-the-art methods on video Mallard-Fly with a value of 66.1%.

4.3 Evaluation on Multi-object Datasets

We compare our method with RVOS on both DAVIS2017 and UNVOSeg datasets, so as to evaluate its effectiveness in multi-object segmentation. Table 4 shows the results on DAVIS2017 and UNVOSeg dataset with J-measure and F-measure. The results show that the performance of our method is slightly lower than that of RVOS. The reason is that the video classification standards are not uniform. For example, people and backpacks are seen as one target in one video and two targets in another video. Therefore, this dataset is more suitable for semi-supervised video object segmentation. It should be noticed that our method still surpasses RVOS on several video sequences. In Table 5, we give some evaluation results of several videos on the DAVIS2017 and UNVOSeg datasets. Note that the J-measure and F-measure in video gold-fish is extremely low, because it is very challenging to distinguish so many similar objects. However, our method outperforms RVOS by a significant margin, from 11.2% to 35.8% with J-measure, in video people-dog. The reason is that our method is able to reconnect the same video object, even after a long time before re-entering the field of vision. Figure 4 show some visual example results of both our method and RVOS on DAVIS2017 and the UNVOSeg dataset, respectively.

Table 4. The results on the DAVIS2017 and UNVOSeg dataset with J-measure and F-measure.

Dataset	RVOS		Ours	
	J-Mean	F-Mean	J-Mean	F-Mean
DAVIS2017	39.0	48.3	27.9	29.9
UNVOSeg	16.4	18.2	31.7	34.3

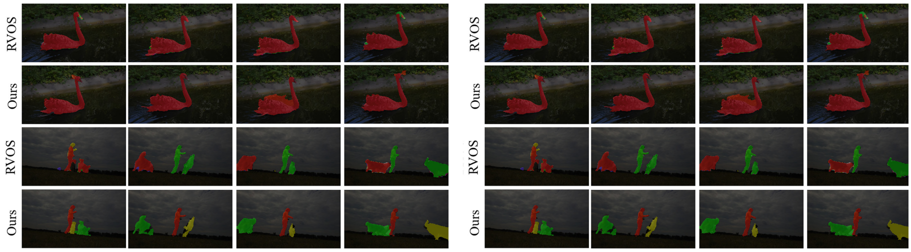

Fig. 4. Some visual example results of our method and RVOS on DAVIS2017 and UNVOSeg datasets.

Table 5. Some of the results on DAVIS2017 and UNVOSeg dataset with J-measure and F-measure.

DAVIS2017	RVOS		Ours		UNVOSeg	RVOS		Ours	
	J	F	J	F		J	F	J	F
Dogs-jump	40.7	51.3	**60.4**	**65.8**	Cats	**26.5**	**22.8**	21.6	19.4
India	22.8	25.5	**45.6**	**44.5**	People-dog	11.2	13.6	**35.8**	**37.8**
Lab-coat	10.0	**22.3**	**20.9**	19.1	Bullfight	22.8	25.5	**45.6**	**44.5**
Soapbox	27.0	**42.3**	**30.3**	39.9	Birds	15.7	16.9	**28.6**	**31.3**
Gold-fish	**25.9**	**38.9**	8.9	8.6	Monkey-dog	5.3	**10.2**	**11.7**	9.6
Kite-surf	**29.7**	**43.8**	26.2	32.3	Elephants	40.2	35.1	**61.4**	**60.4**
Judo	**25.1**	**52.3**	24.4	28.9	Bears	11.4	8.6	**34.8**	**38.6**

To summarize, the results on the above four datasets clearly reveal that: Although our method produces inferior results on single-object segmentation evaluation on video datasets without noisy frames (*i.e.*, SegTrack, DAVIS2016, and DAVIS2017 datasets), we are able to obtain better results on multi-object segmentation evaluation on video dataset with noisy frames (*i.e.*, UNVOSeg dataset), when compared with state-of-the-art methods. This strongly demonstrates that our method is capable of jointly detecting and segmenting the multiple objects from an untrimmed video, where object detection and object segmentation work in a joint way.

5 Conclusion

We proposed a multi-object detection and segmentation method to separate the target objects from an untrimmed video, which benefits from an object decision module and an object segmentation module. The object decision module can identify the target objects and their correspondences across the whole video, and the object segmentation module can segment the target objects out from the background. What's more, we present a new UNVOSeg dataset to fully evaluate the multi-object segmentation performance of our method. Extensive comparisons with the state-of-the-art methods on four datasets validated the efficacy of our method.

Acknowledgement. This work was supported partly by National Key R&D Program of China Grant 2018AAA0101400, NSFC Grants 61629301, 61773312, and 61976171, China Postdoctoral Science Foundation Grant 2019M653642, and Young Elite Scientists Sponsorship Program by CAST Grant 2018QNRC001.

References

1. Shin Yoon, J., Rameau, F., Kim, J., Lee, S., Shin, S., So Kweon, I.: Pixel-level matching for video object segmentation using convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2167–2176 (2017)
2. Jampani, V., Gadde, R., Gehler, P.V.: Video propagation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 451–461 (2017)
3. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 221–230 (2017)
4. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2663–2672 (2017)
5. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: joint learning for video object segmentation and optical flow. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 686–695 (2017)
6. Hu, Y.T., Huang, J.B., Schwing, A.G.: Videomatch: matching based video object segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 54–70 (2018)
7. Xu, S., Liu, D., Bao, L., Liu, W., Zhou, P.: MHP-VOS: multiple hypotheses propagation for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 314–323 (2019)
8. Haller, E., Leordeanu, M.: Unsupervised object segmentation in video by efficient selection of highly probable positive features. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5085–5093 (2017)
9. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 282–295. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15555-0_21
10. Fragkiadaki, K., Zhang, G., Shi, J.: Video segmentation by tracing discontinuities in a trajectory embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1846–1853 (2012)
11. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1777–1784 (2013)
12. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1995–2002 (2011)

13. Khoreva, A., Galasso, F., Hein, M., Schiele, B.: Classifier based graph construction for video segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 951–960 (2015)
14. Jang, W.D., Lee, C., Kim, C.S.: Primary object segmentation in videos via alternate convex optimization of foreground and background distributions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 696–704 (2016)
15. Jain, S.D., Xiong, B., Grauman, K.: Fusionseg: learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2126 (2017)
16. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convlstm for video salient object detection. In: Proceedings of the European Conference on Computer Vision, pp. 715–731 (2018)
17. Li, S., Seybold, B., Vorobyov, A., Lei, X., Jay Kuo, C.C.: Unsupervised video object segmentation with motion-based bilateral networks. In: Proceedings of the European Conference on Computer Vision, pp. 207–223 (2018)
18. Liu, Z., et al.: Joint video object discovery and segmentation by coupled dynamic markov networks. IEEE Trans. Image Process. **27**(12), 5840–5853 (2018)
19. Jun Koh, Y., Kim, C.S.: CDTS: collaborative detection, tracking, and segmentation for online multiple object segmentation in videos. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
20. Ventura, C., Bellver, M., Girbau, A., Salvador, A., Marques, F., Giro-i Nieto, X.: RVOS: end-to-end recurrent network for video object segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (2019)
21. Arbeláez, P., Pont-Tuset, J., Barron, J.T., Marques, F., Malik, J.: Multiscale combinatorial grouping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 328–335 (2014)
22. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2192–2199 (2013)
23. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 724–732 (2016)
24. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv:1704.00675 (2017)
25. Xu, N., et al.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
26. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a “siamese” time delay neural network. In: Advances in neural information processing systems, pp. 737–744 (1994)
27. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99 (2015)

29. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: object detection via region-based fully convolutional networks. In: Advances in neural information processing systems, pp. 379–387 (2016)
30. Dai, J., et al.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
31. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)
32. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1735–1742 (2006)
33. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 898–916 (2010)
34. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440 (2015)
35. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166–3173 (2013)
36. Zhuo, T., Cheng, Z., Zhang, P., Wong, Y., Kankanhalli, M.: Unsupervised online video object segmentation with motion property understanding. IEEE Trans. Image Process. **29**, 237–249 (2019)
37. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3395–3402 (2015)
38. Taylor, B., Karasev, V., Soatto, S.: Causal video object segmentation from persistence of occlusions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4268–4276 (2015)
39. Tokmakov, P., Alahari, K., Schmid, C.: Learning motion patterns in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3386–3394 (2017)