YouTube-VOS: A Large-Scale Video Object Segmentation Benchmark

Ning Xu[1], Linjie Yang[2], Yuchen Fan[3], Dingcheng Yue[3], Yuchen Liang[3], Jianchao Yang[2], and Thomas Huang[3]

[1] Adobe Research, USA
nxu@adobe.com
[2] Snapchat Research, USA
{linjie.yang,jianchao.yang}@snap.com
[3] University of Illinois at Urbana-Champaign, USA
{yuchenf4,dyue2,yliang35,t-huang1}@illinois.edu

Abstract. Learning long-term spatial-temporal features are critical for many video analysis tasks. However, existing video segmentation methods predominantly rely on static image segmentation techniques, and methods capturing temporal dependency for segmentation have to depend on pretrained optical flow models, leading to suboptimal solutions for the problem. End-to-end sequential learning to explore spatial-temporal features for video segmentation is largely limited by the scale of available video segmentation datasets, i.e., even the largest video segmentation dataset only contains 90 short video clips. To solve this problem, we build a new large-scale video object segmentation dataset called YouTube Video Object Segmentation dataset (YouTube-VOS). Our dataset contains 4,453 YouTube video clips and 94 object categories. This is by far the largest video object segmentation dataset to our knowledge and has been released at http://youtube-vos.org. We further evaluate several existing state-of-the-art video object segmentation algorithms on this dataset which aims to establish baselines for the development of new algorithms in the future.

Keywords: Video object segmentation, Large-scale dataset, Benchmark.

1 Introduction

Learning effective spatial-temporal features has been demonstrated to be very important for many video analysis tasks. For example, Donahue *et al.* [1] propose long-term recurrent convolution network for activity recognition and video captioning. Srivastava *et al.* [2] propose unsupervised learning of video representation with a LSTM autoencoder. Tran *et al.* [3] develop a 3D convolutional network to extract spatial and temporal information jointly from a video. Other works include learning spatial-temporal information for precipitation prediction [4], physical interaction [5], and autonomous driving [6].

Video segmentation plays an important role in video understanding, which fosters many applications, such as accurate object segmentation and tracking, interactive video editing and augmented reality. Video object segmentation, which

targets at segmenting a particular object instance throughout the entire video sequence given only the object mask on the first frame, has attracted much attention from the vision community recently [7,8,9,10,11,12,13,14]. However, existing state-of-the-art video object segmentation approaches primarily rely on single image segmentation frameworks [7,8,9]. For example, Caelles *et al.* [7] propose to train an object segmentation network on static images and then fine-tune the model on the first frame of a test video over hundreds of iterations, so that it remembers the object appearance. The fine-tuned model is then applied to all following individual frames to segment the object without using any temporal information. Even though simple, such an online learning or one-shot learning scheme achieves top performance on video object segmentation benchmarks [15,16]. Although some recent approaches [11,10,13] have been proposed to leverage temporal consistency, they depend on models pretrained on other tasks such as optical flow [17,18] or motion segmentation [19], to extract temporal information. These pretrained models are learned from separate tasks, and therefore are suboptimal for the video segmentation problem.

Learning long-term spatial-temporal features directly for video object segmentation task is, however, largely limited by the scale of existing video object segmentation datasets. For example, the popular benchmark dataset DAVIS [20] has only 90 short video clips, which is barely sufficient to learn a sequence-to-sequence network from scratch like other video analysis tasks. Even if we combine all the videos from available datasets [16,21,22,23,24,25], its scale is still far smaller than many other video analysis datasets such as YouTube-8M [26] and ActivityNet [27]. To solve this problem, we present the first large-scale video object segmentation dataset called YouTube-VOS (YouTube Video Object Segmentation dataset) in this work. Our dataset contains 4,453 YouTube video clips featuring 94 categories covering humans, common animals, vehicles, and accessories. Each video clip is about 3∼6 seconds long and often contains multiple objects, which are manually segmented by professional annotators. Compared to existing datasets, our dataset contains a lot more videos, object categories, object instances and annotations, and a much longer duration of total annotated videos. Table 1 provides quantitative scale comparisons of our new dataset against existing datasets. The dataset has been released at https://youtube-vos.org. We elaborate the collection process of our dataset in Section 3.

In this report, we also retrain state-of-the-art video object segmentation algorithms on YouTube-VOS and benchmark their performance on the validation set which contains 474 videos. In addition, our validation set contains 26 unique categories that do not exist in the training set and are used to evaluate the generalization ability of existing approaches on unseen categories. We provide the detailed results in Section 4.

2   Related work

In the past decades, several datasets [16,21,22,23,24,25] have been created for video object segmentation. All of them are in small scales which usually contain only dozens of videos. In addition, their video content is relatively simple

Table 1: Scale comparison between YouTube-VOS and existing datasets. "Annotations" denotes the total number of object annotations. "Duration" denotes the total duration (in minutes) of the annotated videos.

Scale	JC [21]	ST [22]	YTO [16]	FBMS [24]	DAVIS [15]	[20]	YouTube-VOS (Ours)
Videos	22	14	96	59	50	90	**4,453**
Categories	14	11	10	16	-	-	**94**
Objects	22	24	96	139	50	205	**7,755**
Annotations	6,331	1,475	1,692	1,465	3,440	13,543	**197,272**
Duration	3.52	0.59	9.01	7.70	2.88	5.17	**334.81**

(*e.g.* no heavy occlusion, camera motion or illumination change) and sometimes the video resolution is low. Recently, a new dataset called DAVIS [15,20] was published and has become the benchmark dataset in this area. Its 2016 version contains 50 videos with a single foreground object per video while the 2017 version has 90 videos with multiple objects per video. In comparison to previous datasets [16,21,22,23,24,25], DAVIS has both higher-quality of video resolutions and annotations. In addition, their video content is more complicated with multi-object interactions, camera motion, and occlusions.

Early methods [16,28,29,30,31] for video object segmentation often solve some spatial-temporal graph structures with hand-crafted energy terms, which are usually associated with features including appearance, boundary, motion and optical flows. Recently, deep-learning based methods were proposed due to its great success in image segmentation tasks [32,33]. Most of these methods [7,8,10,11,9] build their models based on an image segmentation network and do not involve sequential modeling. Online learning [7] is commonly used to improve their performance. To make the model temporally consistent, the predicted mask of the previous frame is used as a guidance in [8,9,14]. Other methods have been proposed to leverage spatial-temporal information. Jampani *et al.* [12] use spatial-temporal consistency to propagate object masks over time. Tokmakov *et al.* [13] use a two-stream network to model objects' appearance and motion and use a recurrent layer to capture the evolution. However, due to the lack of training videos, they use a pretrained motion segmentation model [19] and optical-flow model [17], which leads to suboptimal results since the model is not trained end-to-end to best capture spatial-temporal features. Recently, Xu *et al.* [34] propose a sequence-to-sequence learning algorithm to learn long-term spatial-temporal information for segmentation. Their models are trained on a preliminary version of YouTube-VOS and do not depend on existing optical flow or motion segmentation models.

3   YouTube-VOS

To create our dataset, we first carefully select a set of video categories including animals (*e.g. ant, eagle, goldfish, person*), vehicles (*e.g. airplane, bicycle, boat,*

Table 2: A complete list of object categories and number of instances in YouTube-VOS. Objects are sorted from most frequent to least frequent.

person	1702	cat	115	train	77	hedgehog	49	squirrel	24	table	10
ape	239	snake	114	owl	74	eagle	45	rope	24	camera	10
parrot	222	zebra	111	plant	73	snail	44	chameleon	22	watch	9
giant_panda	222	giraffe	110	airplane	73	toilet	43	box	20	stuffed_toy	9
sedan	221	bear	97	bus	70	camel	40	tissue	18	guitar	8
lizard	189	fox	90	shark	66	frisbee	39	kangaroo	18	microphone	7
duck	186	leopard	88	tiger	66	whale	38	cloth	18	cup	6
dog	177	elephant	87	surfboard	64	knife	38	bottle	17	shovel	6
skateboard	173	horse	87	earless_seal	63	tennis_racket	38	small_panda	16	flag	6
monkey	164	others	86	frog	63	crocodile	37	spider	14	mirror	5
sheep	155	deer	86	mouse	63	umbrella	36	ball	14	ring	5
fish	138	motorbike	85	boat	61	paddle	33	jellyfish	13	necklace	4
rabbit	135	turtle	84	snowboard	59	raccoon	29	eyeglasses	11	ant	3
hat	131	bird	81	penguin	53	parachute	28	backpack	11		
cow	128	truck	81	lion	52	bucket	28	butterfly	11		
hand	121	dolphin	80	sign	50	bike	28	handbag	11		

sedan), accessories (*e.g. eyeglass, hat, bag*), common objects (*e.g. potted plant, knife, sign, umbrella*), and humans in various activities (*e.g. tennis, skateboarding, motorcycling, surfing*). The videos containing human activities have diversified appearance and motion, so we collect human-related videos using a list of activity tags to increase the diversity of human motion and behaviors. Most of these videos contain interactions between a person and a corresponding object, such as tennis racket, skateboard, motorcycle, etc. The entire category set includes 78 categories that covers diverse objects and motions, and should be representative for everyday scenarios.

We then collect many high-resolution videos with the selected category labels from the large-scale video classification dataset YouTube-8M [26]. This dataset consists of millions of YouTube videos associated with more than 4,700 visual entities. We utilize its category annotations to retrieve candidate videos that we are interested in. Specifically, up to 100 videos are retrieved for each category in our segmentation category set. There are several advantages to using YouTube videos to create our segmentation dataset. First, YouTube videos have very diverse object appearances and motions. Challenging cases for video object segmentation, such as occlusions, fast object motions and change of appearances, commonly exist in YouTube videos. Second, YouTube videos are taken by both professionals and amateurs and thus different levels of camera motions are shown in the crawled videos. Algorithms trained on such data could potentially handle camera motion better and thus are more practical. Last but not the least, many YouTube videos are taken by today's smart phone devices and there are demanding needs to segment objects in those videos for applications such as video editing and augmented reality.

Fig. 1: The ground truth annotations of sample video clips in our dataset. Different objects are highlighted with different colors.

Since the retrieved videos are usually long (several minutes) and have shot transitions, we use an off-the-shelf video shot detection algorithm [4] to automatically partition each video into multiple video clips. We first remove the clips from the first and last 10% of the video, since these clips have a high chance of containing introductory subtitles and credits lists. We then sample up to five clips with appropriate lengths (3∼6 seconds) per video and manually verify that these clips contain the correct object categories and are useful for our task (*e.g.* no scene transition, not too dark, shaky, or blurry). After the video clips are collected, we ask human annotators to select up to five objects of proper sizes and categories per video clip and carefully annotate them (by tracing their boundaries instead of rough polygons) every five frames in a 30fps frame rate, which results in a 6fps sampling rate. Given a video and its category, annotators are first required to annotate objects belonging to that category. If the video contains other salient objects, we ask the annotators to label them as well, so that each video has multiple objects annotated, and the object categories are not limited to our initial 78 categories. In human activity videos, both the human subject and the object he/she interacts with are labeled, *e.g.*, both the person and the skateboard are required to be labeled in a "skateboarding" video. Further, the instance-level categories are labeled for each annotated object, including not only the video-level categories, but also additional categories that the labelers has labeled, resulting in a total of 94 object categories. The activity categories are removed since they do not represent single objects. Note in an earlier version of the dataset [34], only video-level categories are available. Some annotation examples are shown in Figure 1. Unlike dense per-frame annotation in previous datasets [21,15,20], we believe that the temporal correlation between five consecutive frames is sufficiently strong that annotations can be omitted for intermediate frames to reduce the annotation efforts. Such a skip-frame annotation strategy allows us to scale up the number of videos and objects under the same annotation budget, which are important factors for better performance. We find empirically that our dataset is effective in training different segmentation algorithm.

As a result, our collected dataset YouTube-VOS consists of 4,453 YouTube video clips which is about 50 times larger than YouTubeObjects [16], the exist-

---

[4] http://johmathe.name/shotdetect.html

ing video object segmentation dataset with the most videos. Our dataset also has a total of 197,272 object annotations which is 15 times larger than those of DAVIS 2017 [20]. There are 94 different object categories including person, animals, vehicles, furnitures, and other common objects. A complete list of object categories can be seen in Table 2. Therefore, YouTube-VOS is the largest and most comprehensive dataset for video object segmentation to date.

4 Experiments

In this section, we retrain state-of-the-art video object segmentation methods on YouTube-VOS training set and evaluate their performance on YouTube-VOS validation set. All the algorithms are trained and tested under the same setting. We hope the experiment results could setup baselines for the development of new algorithms in the future.

4.1 Settings

The whole dataset which consists of 4,453 videos is split into training (3,471), validation (474) and test (508) sets. Since the dataset has been used for a workshop competition (*i.e.* The 1st Large-scale Video Object Segmentation Challenge) [5], the test set will only be available during the competition period while the validation set will be always publicly available. Therefore we only use the validation set for evaluation. In the training set, there are 65 unique object categories which are regarded as seen categories. In the validation set, there are 91 unique object categories which include all the seen categories and 26 unseen categories. As stated, the unseen categories are used to evaluate the generalization ability of different algorithms. For training the state-of-the-arts algorithms, we first resize the training frames to a fixed size (*i.e.* 256×448) and then use their publicly released codes to train their models. We also evaluate the algorithms on other image resolutions such as 480p but the difference is negligible. All the models are trained sufficiently until convergence. For evaluation, we follow the evaluation method used by the workshop, which computes the region similarity $\mathcal{J}$ and the contour accuracy $\mathcal{F}$ as in [15]. The final result is the average of four metrics: $\mathcal{J}$ for seen categories, $\mathcal{F}$ for seen categories, $\mathcal{J}$ for unseen categories, and $\mathcal{F}$ for unseen categories.

4.2 Methods

We compare several recently proposed algorithms which achieved state-of-the-art results on previous small-scale benchmarks. These algorithms are OSVOS [7], MaskTrack [8], OSMN [9], OnAVOS [35] and S2S [34]. For more details of these algorithms, please refer to their papers.

---

[5] https://youtube-vos.org/challenge/challenge2018

4.3   Results

The results are presented in Table 3. The first four methods use static image segmentation models and three of them (*i.e.* OSVOS, MaskTrack and OnAVOS) require online learning. S2S leverages long-term spatial-temporal coherence by recurrent neural networks (RNN) and its model without online learning (the second last row in Table 3) achieves comparable performance compared to the best results of online-learning methods, which effectively demonstrates the importance of long-term spatial-temporal information for video object segmentation. With online learning, S2S is further improved and achieves around 6% absolute improvement over the best online-learning method OSVOS on overall accuracy. Surprisingly, OnAVOS which is the best performing method on DAVIS does not achieve good results on our dataset. We believe the drastic appearance changes and complex motion patterns in our dataset makes the online adaptation fail in many cases.

Next we compare the generalization ability of existing methods on unseen categories in Table 3. All the methods have obviously better results on seen categories than unseen categories. Among them, OSVOS has the least discrepancy, possibly due to the pre-training on large-scale image segmentation dataset. It is also worth noting that methods with online learning also suffer from this problem, which suggests that although online learning is helpful to improve the accuracy on unseen categories, pre-training on some large-scale object segmentation dataset is still important to learn general object feature representation. In general, the results shows a much larger performance gap between seen and unseen categories compared to [34]. We believe it is because instance categories are used to split the seen and unseen subset in the current setting, comparing to [34] in which subsets are split using video-level categories. The current setting leads to a clearer separation between seen and unseen categories and is more challenging.

Lastly we compare the inference speed of all the methods averaged per frame. OSMN and S2S (w/o OL) do not use online learning and thus have very fast inference speed, which can be applied in real time. This is a big advantage over those online learning methods especially for mobile applications. While the performance is still inferior to online learning ones.

5   Conclusion

In this report, we introduce the largest video object segmentation dataset to date. The new dataset called YouTube-VOS, much larger than existing datasets in terms of number of videos and annotations, allows us to evaluate existing state-of-the-art video object segmentation methods more comprehensively. We believe the new dataset will foster research on video-based computer vision in general.

Table 3: Comparisons of state-of-the-art methods on YouTube-VOS validation set. "$\mathcal{J}$" and "$\mathcal{F}$" denote the region similarity and the contour accuracy. "seen" and "unseen" denote the results averaged over the seen categories and unseen categories. "Overall" denote the results averaged over the four metrics. "OL" denotes online learning. The best results are highlighted in bold.

Method	$\mathcal{J}$ seen	$\mathcal{J}$ unseen	$\mathcal{F}$ seen	$\mathcal{F}$ unseen	Overall	Speed (s/frame)
OSVOS [7]	59.8%	54.2%	60.5%	60.7%	58.8%	10
MaskTrack [8]	59.9%	45.0%	59.5%	47.9%	53.1%	12
OSMN [9]	60.0%	40.6%	60.1%	44.0%	51.2%	**0.14**
OnAVOS [35]	60.1%	46.6%	62.7%	51.4%	55.2%	13
S2S (w/o OL) [34]	66.7%	48.2%	65.5%	50.3%	57.6%	0.16
S2S (with OL) [34]	**71.0%**	**55.5%**	**70.0%**	**61.2%**	**64.4%**	9

References

1. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: CVPR. (2015)
2. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: International conference on machine learning. (2015)
3. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. (2015)
4. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems. (2015) 802–810
5. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. In: Advances in neural information processing systems. (2016)
6. Xu, H., Gao, Y., Yu, F., Darrell, T.: End-to-end learning of driving models from large-scale video datasets. In: CVPR. (2017)
7. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: CVPR. (2017)
8. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., A.Sorkine-Hornung: Learning video object segmentation from static images. In: CVPR. (2017)
9. Yang, L., Xiong, X., Wang, Y., Yang, J., Katsaggelos, A.K.: Efficient video object segmentation via network modulation. In: CVPR. (2018)
10. Cheng, J., Tsai, Y.H., Wang, S., Yang, M.H.: Segflow: Joint learning for video object segmentation and optical flow. In: IEEE International Conference on Computer Vision (ICCV). (2017)
11. Dutt Jain, S., Xiong, B., Grauman, K.: Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
12. Jampani, V., Gadde, R., Gehler, P.V.: Video propagation networks. In: CVPR. (2017)

13. Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. In: ICCV. (2017)
14. Hu, Y.T., Huang, J.B., Schwing, A.: Maskrnn: Instance level video object segmentation. In: NIPS. (2017)
15. 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: CVPR. (2016)
16. Jain, S.D., Grauman, K.: Supervoxel-consistent foreground propagation in video. In: ECCV. (2014)
17. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: Evolution of optical flow estimation with deep networks. CVPR (2017)
18. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Epicflow: Edge-preserving interpolation of correspondences for optical flow. In: CVPR. (2015)
19. Tokmakov, P., Alahari, K., Schmid, C.: Learning motion patterns in videos. In: CVPR. (2017)
20. 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)
21. Fan, Q., Zhong, F., Lischinski, D., Cohen-Or, D., Chen, B.: Jumpcut:non-successive mask transfer and interpolation for video cutout. In: ACM Trans. Graph., 34(6). (2015)
22. Li, F., Kim, T., Humayun, A., Tsai, D., Rehg, J.M.: Video segmentation by tracking many figure-ground segments. In: ICCV. (2013)
23. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: ECCV. (2010) 282–295
24. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE transactions on pattern analysis and machine intelligence **36**(6) (2014) 1187–1200
25. Galasso, F., Nagaraja, N.S., Cárdenas, T.J., Brox, T., Schiele, B.: A unified video segmentation benchmark: Annotation, metrics and analysis. In: ICCV, IEEE (2013)
26. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
27. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE (2015) 961–970
28. Nagaraja, N.S., Schmidt, F.R., Brox, T.: Video segmentation with just a few strokes. In: ICCV. (2015) 3235–3243
29. Faktor, A., Irani, M.: Video segmentation by non-local consensus voting. In: BMVC. (2014)
30. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 1777–1784
31. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. In: European conference on computer vision, Springer (2010) 282–295
32. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(4) (2017) 640–651

33. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In: IEEE T-PAMI. Volume 40. (2018) 834–848
34. Xu, N., Yang, L., Yue, D., Yang, J., Price, B., Yang, J., Cohen, S., Fan, Y., Liang, Y., Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 585–601
35. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. arXiv preprint arXiv:1706.09364 (2017)