Survey on Weakly Supervised Video Segmentation

Wenrao Pang

School of Mathematics, China University of Mining and Technology, China

Keywords: Weakly supervised learning, video segmentation, semi-supervised learning.

Abstract: Weakly supervised learning are challenging tasks in video segmentation. A survey on deep learning technique for video segmentation from all angles has been published in [1], including 3 major categories and 7 specific small problem directions. However, the literature only describe the main algorithms from 2017 - July 2021. This article mainly reviews the task of video segmentation from the perspective of weak supervision for video segmentation. Besides, we show the main data sets usually used in this assignment and analysis the main algorithms and the latest ones that not mentioned in above survey. Additionally, we state a famous challenge in this field and the evaluation methods to demonstrate the performance of algorithms.

1. Introduction

Machine learning has achieved great success in many tasks. Machine learning tasks can be roughly divided into two categories: one is supervised learning and the other is unsupervised learning. Nevertheless, they all need to learn a prognostic model from a training data set containing a large number of training samples, and each training sample links to an event or object.

Video segmentation [1] is a branch of the segmentation problem. It contains supervised, unsupervised and semi/weakly-supervised issues. Recent years have witnessed the compelling success of supervised video segmentation. For unsupervised tasks, the algorithm autonomously divides the real object. For the semi/weakly-supervised tasks, only the correct segmentation mask of the first frame of the video is given, and then the target is segmented at the pixel level in each subsequent frame. In essence, the task of Semi-supervised video segmentation is to perform target tracking at the pixel level.

Traditional definition of weakly supervised method [2] divide the weakly learning into three parts: incomplete supervision, inexact supervision and inaccurate supervision. Incomplete supervision combine a small amount of labeled data with abundant unlabeled data to obtain a better trainer. Both active learning and semi-supervised learning are currently considered to be the main solution to incomplete supervision. The former is human-computer interaction and the latter relies on the machine itself. Inexact supervision only has coarsegrained label information, which can be resolve by muti-instance learning. Inaccurate ones represent the data including wrong tags, this can be solved by learning with label noise. In Figure 1, bars represent feature vectors, red/blue are labels and"?" means the label may be inaccurate.

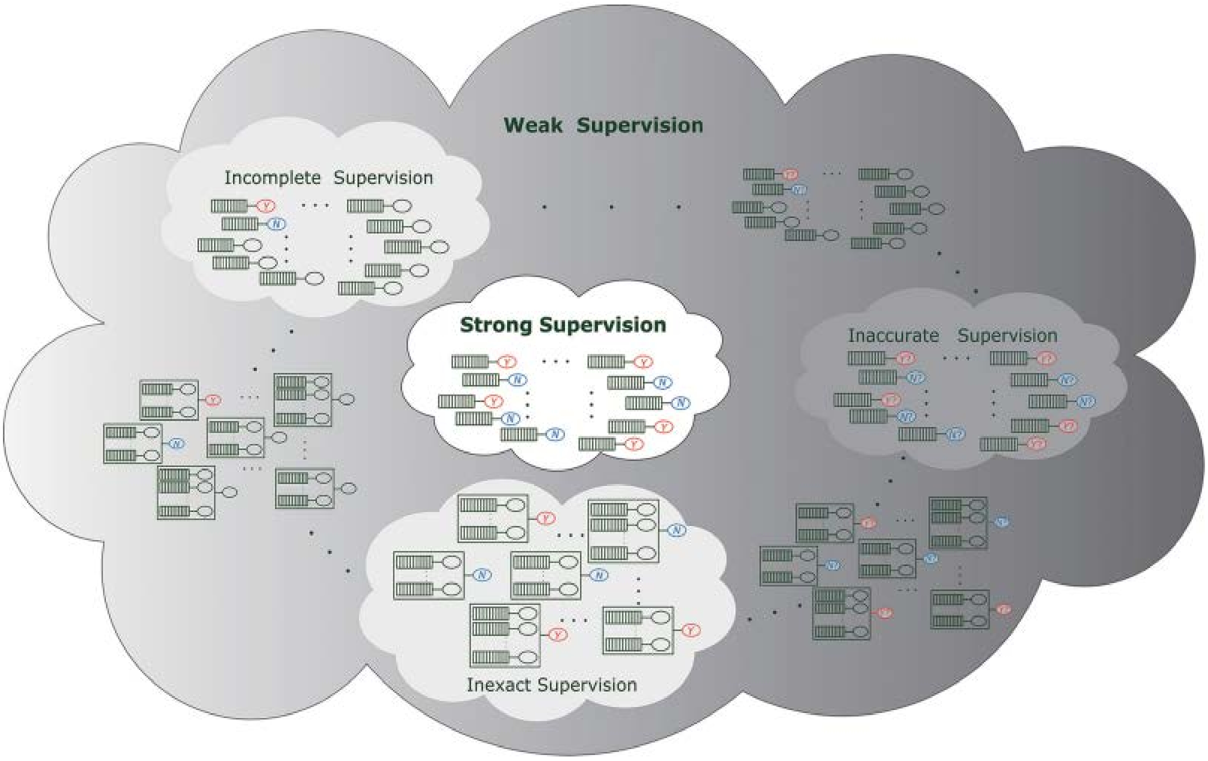

Figure 1. Illustration of three typical types of weak supervision [2].

My initial idea about the video segmentation is that the combination of semi-supervised and weak-supervised in video segmentation, after the first frame is marked, the center point is taken as the weakly marked label of the next frame.

2. Related work

Segmentation tasks currently include image segmentation [6] and video segmentation [1], which are very important tasks in computer vision. They are all pixel-level classification for an image or a certain frame of video. Classical image-based semantic segmentation algorithms are worthy of attention.

Conditional Random Field (CRF) [10] post processing are proposed to improve the segmentation. CRFs are graphical models which ‘smooth’ segmentation based on the underlying image intensities. Global scene categories matter because it provides clues on the distribution of the segmentation classes. Pyramid pooling module [11] captures this information by applying large kernel pooling layers. Unlike pre-training, self-training is always helpful when using stronger data augmentation, in both low-data and high-data regimes. EfficientNet-L2+NAS-FPN+Noisy Student [12] used the combination of PASCAL aug set as the source of unlabeled data, and NAS-FPN with EfficientNet-L2 as the segmentation model to get the highest scoring on the PASCAL VOC 2012 test set. Global Convolutional Network (Large Kernel Matters) [13] are proposed as an encoder-decoder architecture with very large kernels convolutions.

Different from the semantic segmentation of images, the video target segmentation adds a timing module, which is to find the corresponding pixels of the target in each continuous frame of the video. Therefore, it is difficult to achieve the performance of video processing by directly using classic semantic segmentation algorithms. This is why the MaskTrack [14] algorithm based on timing is better than the OSVOS algorithm [5] based on independent processing of video independent frames.

Wang *et al.* [15] reviewed the latest papers which are published before July 2, 2021, and summarized them as 3 major categories (supervised, unsupervised, and semi/weakly supervised learning based) and 7 specific small problem directions. To be more specific, see [15]. These 7 specific tasks with the three categories as follows:

Unsupervised video segmentation or zero-shot video segmentation with few papers about weakly:
(a) Object-level automatic video object segmentation (object-level AVOS);
(b) Instance-level automatic video object segmentation (instance-level AVOS);
Semi-supervised video segmentation or one-shot video segmentation:
(c) Semi-automatic video object segmentation (SVOS);
(d) Interactive video object segmentation (IVOS);
(e) language-guided video object segmentation (LVOS);
Supervised video segmentation:
(f) Video semantic segmentation (VSS);
(g) Video instance segmentation (VIS);
(h) Video panoptic segmentation (VPS).

3. Dateset

There are 13 data set about video segmentation tasks in website [3]. In this section, we will listed three famous ones whose citations are over 100: DAVIS with 324 papers, DAVIS2017 with 122 papers and CamVid with 159 papers. And the others are Virtual KITTI, SegTrack-v2, SUN3D, BDD100K, YouTubeVIS, Apollos cape, Kvasir, IKEA ASM and DukeMTMC-attribute, for specific introduce, see [3].

The Densely Annotation Video Segmentation data set (DAVIS) is a high quality and high resolution densely annotated video segmentation data set, which is under two resolutions: 480p and 1080p. It includes 50 video sequences with 3455 densely annotated frames in pixel level. DAVIS only divided into two parts: 30 videos with 2079 frames are for training and 20 videos with 1376 frames are for validation.

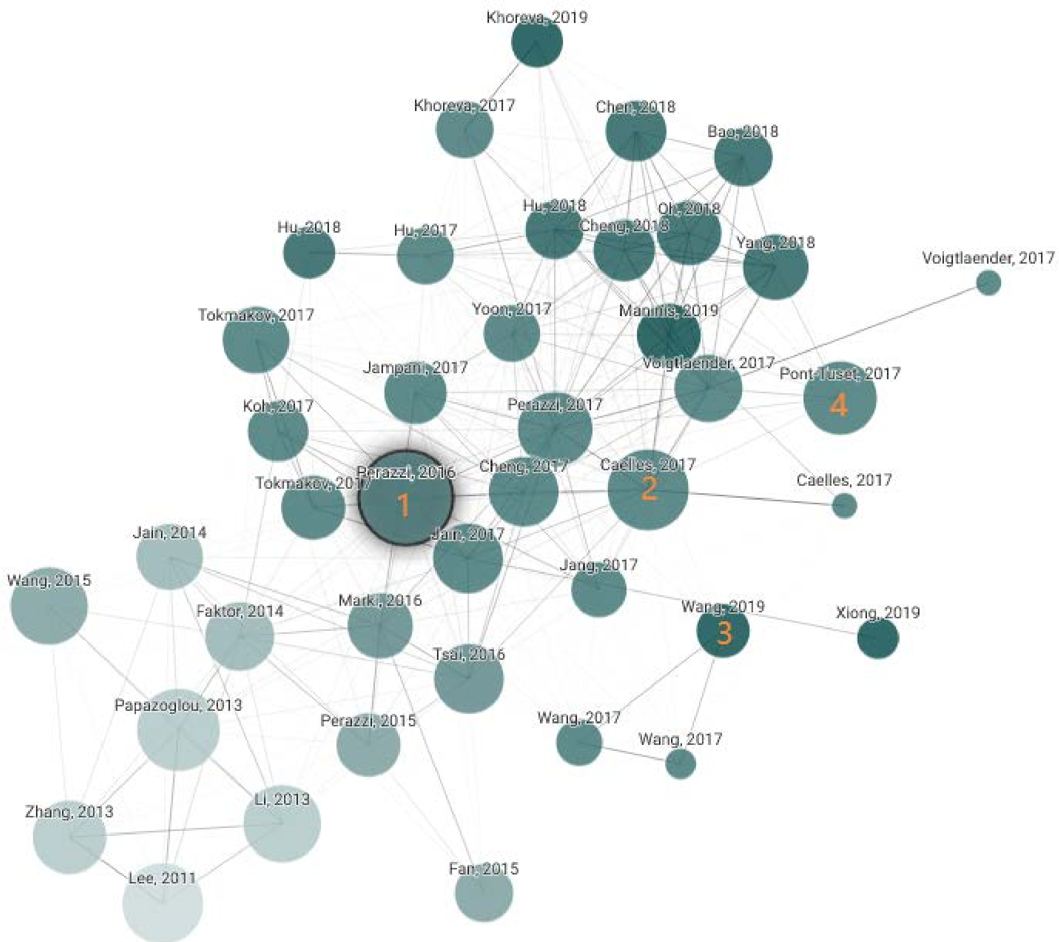

Figure 2. Literature analysis Network Diagram related to the DAVIS.

Figure 2 reflects the relationship of the published papers before 2019 which we can observe the history of video segmentation clearly. the orange number 1 represents the DAVIS in 2016, number 2 denote an influential work named One-Shot Video Object Segmentation (OSVOS) [5] orange number 4 is the most famous competition named the 2017 DAVIS Challenge on Video Object Segmentation [7]. In order to solve this difficulty, research has been carried out from various sources. Until 2019, there were a large number of semi/weakly-supervised papers represented by Paper 3.

DAVIS17 is a dataset for video object segmentation, which is an. It contains a total of 150 videos - 60 for training, 30 for validation, and 60 for testing. DAVIS17 is a competition data set proposed for the DAVIS. The details are explained in section 4.

CamVid (Cambridge-driving Labeled Video Database) [8] is a road/driving scene understanding database which is a high-definition ground truth database. Semantic object classes are clearly divided in the video. It was originally captured as five video sequences with a 960×720 resolution camera mounted on the dashboard of a car. Those sequences were sampled (four of them at 1 fps and one at 15 fps) adding up to 701 frames. Those stills were manually annotated with 32 classes.

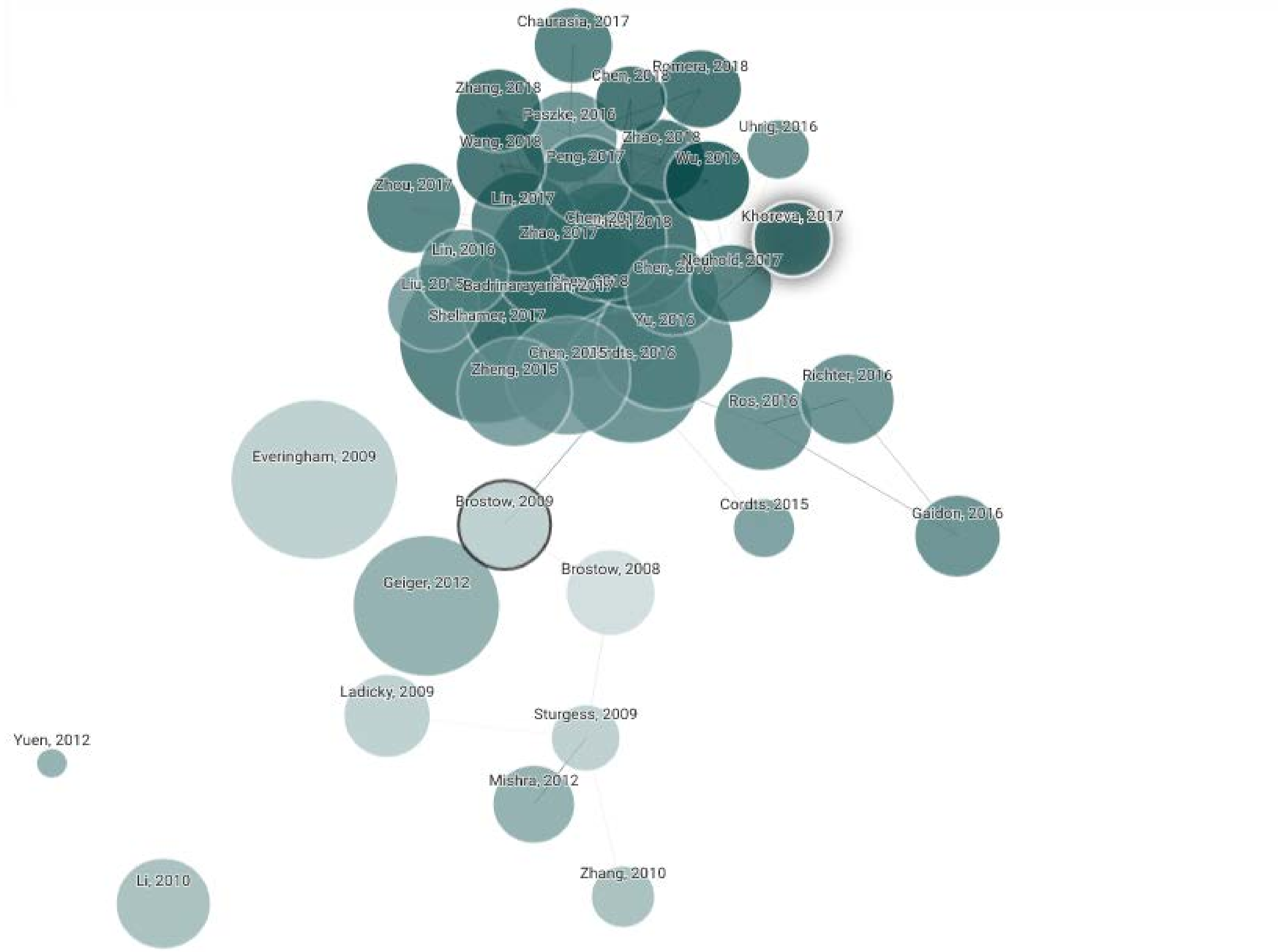

Figure 3. Literature analysis Network Diagram related to the CamVid.

Although CamVid has been proposed since 2009, it was not until 2017 that a method of semantic segmentation appeared to map this data set to the segmentation problem to solve it. And from the industrial world, the timing and recognition accuracy are relatively acceptable. As can be seen from Figure 3, the ones with black border is the published paper of the data set, and the ones with white border is the papers related to weakly supervised semantic segmentation.

4. Challenge and evaluations

DAVIS Challenge on Video Object Segmentation has held for 4 years from 2017 to 2020, which is organized in CVPR [9].The per-object measures are those described in the original DAVIS CVPR

2016 paper: Region Smilarity/Jaccard ($\mathcal{J}$) and Boundary F measure ($\mathcal{F}$). The overall ranking measures are computed as the mean between $\mathcal{J}$ and $\mathcal{F}$, both averaged over all objects.

Region Smilarity/Jaccard ($\mathcal{J}$) is answer the question that how well the output segmentation M fits the given ground-truth mask G. It defied as below which calculates the intersection-overunion (IoU) between M and G:

$$\mathcal{J} = \frac{\mathrm{M} \cap \mathrm{G}}{\mathrm{M} \cup \mathrm{G}} \tag{1}$$

Contour Accuracy/Boundary $\mathcal{F}$ measure ($\mathcal{F}$) can compute the contour-based precision as well as the recall $\mathrm{P_c}$ and $\mathrm{R_c}$ which are from the contour points of $\mathrm{c(M)}$ and $\mathrm{c(G)}$, and defined as

$$\mathcal{F} = \frac{2\mathrm{P_c}\mathrm{R_c}}{\mathrm{P_c}+\mathrm{R_c}} \tag{2}$$

The mean performance metric $\mathrm{m}(\mathcal{M}, \mathcal{S})$ which is computed as the mean between $\mathcal{J}$ and$\mathcal{F}$, both averaged over all objects. Given an objecto $\in \mathcal{O}_\mathrm{s}$, $\mathrm{s(o)} \in \mathrm{S}$ is the sequence time where object o has appeared?

$$\mathrm{m}(\mathcal{M}, \mathcal{S}) = \frac{1}{|\mathcal{O}_\mathrm{s}|} \sum_{\mathrm{o} \in \mathcal{O}_\mathrm{s}} \frac{1}{|\mathcal{F}_{\mathrm{s(o)}}|} \mathcal{M}(\mathrm{m}_\mathrm{o}^\mathrm{f}, \mathrm{g}_\mathrm{o}^\mathrm{f}) \tag{3}$$

Where $\mathcal{F}_{\mathrm{s(o)}}$the set of frames in sequence for object o is, $\mathrm{m}_\mathrm{o}^\mathrm{f}$ is the binary masks of the object o in framef, $\mathrm{g}_\mathrm{o}^\mathrm{f}$ is the ground truth for it.

Overall performance metric is the average of the mean region and contour accuracies.

$$\mathrm{M(S)} = \frac{1}{2}[\mathrm{m}(\mathcal{J}, \mathcal{S}) + \mathrm{m}(\mathcal{F}, \mathcal{S})] \tag{4}$$

5. Conclusion

A great deal of attention has focused on the video segmentation in recent years. In this paper, we are interested in the weakly supervised learning methods for its latest algorithms, common data sets and open challenges. Therefore, we have added some latest algorithms about weakly supervised video segmentation and introduced them in detail.

We also show the difference definition of weakly supervised learning in 2018 [2] and in 2021 [1]. The former covers three aspects from the perspective of data categories, namely incomplete supervision, inexact supervision as well as inaccurate supervision, and comprehensively summarizes and generalizes other situations that may be weak labels. The latter is aimed at the image and segmentation fields, and subdivides weak supervision into several aspects. In other words, the weak supervision in the segmentation field fills the gap in the paper’s weak supervision classification.

References

[1] Wenguan Wang, Tianfei Zhou, Fatih Porikli, David J. Crandall, Luc Van Gool: A Survey on Deep Learning Technique for Video Segmentation. CoRR abs/2107.01153 (2021)

[2] Zhi-Hua Zhou, A brief introduction to weakly supervised learning, National Science Review, Volume 5, Issue 1, January 2018, Pages 44–53, https://doi.org/10.1093/nsr/nwx106.

[3] https://paperswithcode.com/datasets?mod=videos&task=semantic-segmentation&page=1

[4] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 724-732, doi: 10.1109/CVPR.2016.85.

[5] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, Luc Van Gool:One-Shot Video Object Segmentation. CVPR 2017: 5320-5329

[6] Shijie Hao, Yuan Zhou, Yanrong Guo: A Brief Survey on Semantic Segmentation with Deep Learning. Neurocomputing 406: 302-321 (2020)

[7] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbelaez, Alexander Sorkine-Hornung, Luc Van Gool: The 2017 DAVIS Challenge on Video Object Segmentation. CoRR abs/1704.00675 (2017)

[8] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla: Semantic object classes in video: A high-definition ground truth database. Pattern Recognit. Lett. 30(2): 88-97 (2009)

[9] https://davischallenge.org/challenge2020/semisupervised.html

[10] Philipp Krähenbühl, Vladlen Koltun: Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. CoRR abs/1210.5644 (2012)

[11] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia: Pyramid Scene Parsing Network. CVPR 2017: 6230-6239

[12] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D. Cubuk, Quoc V. Le: Rethinking Pre-training and Self-training. CoRR abs/2006.06882 (2020)

[13] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun: Large Kernel Matters - Improve Semantic Segmentation by Global Convolutional Network. CVPR 2017: 1743-1751

[14] Anna Khoreva, Federico Perazzi, Rodrigo Benenson, Bernt Schiele, Alexander Sorkine-Hornung: Learning Video Object Segmentation from Static Images. CoRR abs/1612.02646 (2016)

[15] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, José García Rodríguez: A survey on deep learning techniques for image and video semantic segmentation. Appl. Soft Comput. 70: 41-65 (2018)