Video Object Segmentation: A Survey

Zhixin Piao   Yongfei Liu   Qiong Huang

School of Information Science and Technology, ShanghaiTech University

{piaozhx,liuyf,huangqiong}@shanghaitech.edu.cn

Abstract

Image semantic segmentation is more and more being of interest for computer vision and machine learning researchers. Many applications on the rise need accurate and efcient segmentation mechanisms: autonomous driving, indoor navigation, and even virtual or augmented reality systems to name a few. This demand coincides with the rise of deep learning approaches in almost every eld or application target related to computer vision, including semantic segmentation or scene understanding. This paper provides a review on deep learning methods for semantic segmentation applied to various application areas. Firstly, we describe the terminology of this eld as well as mandatory background concepts. Next, the main datasets and challenges are exposed to help researchers decide which are the ones that best suit their needs and their targets. Then, existing methods are reviewed, highlighting their contributions and their signicance in the eld. Finally, quantitative results are given for the described methods and the datasets in which they were evaluated, following up with a discussion of the results. At last, we point out a set of promising future works and draw our own conclusions about the state of the art of semantic segmentation using deep learning techniques.

1. Introduction

Nowadays, semantic segmentation applied to still 2D images, video, and even 3D or volumetric data

 is one of the key problems in the eld of computer vision. Looking at the big picture, semantic segmentation is one of the high-level task that paves the way towards complete scene understanding. The importance of scene understanding as a core computer vision problem is highlighted by the fact that an increasing number of applications nourish from inferring knowledge from imagery. Some of those applications include autonomous driving [1] [2] [3], human-machine interaction [4], computational photography

[5], image search engines [6], and augmented reality to name a few. Such problem has been addressed in the past using various traditional computer vision and machine learning techniques. Despite the popularity of those kind of methods, the deep learning revolution has turned the tables so that many computer vision problems semantic segmentation among them are being tackled using deep architectures, usually Convolutional Neural Networks (CNNs) [7] [8] [9]

[10] [11], which are surpassing other approaches by a large margin in terms of accuracy and sometimes even efciency. However, deep learning is far from the maturity achieved by other old-established branches of computer vision and machine learning. Because of that, there is a lack of unifying works and state of the art reviews. The ever-changing state of the eld makes initiation difcult and keeping up with its evolution pace is an incredibly time-consuming task due to the sheer amount of new literature being produced. This makes it hard to keep track of the works dealing with semantic segmentation and properly interpret their proposals, prune subpar approaches, and validate results.

To the best of our knowledge, this is the rst review to focus explicitly on deep learning for semantic segmentation. Various semantic segmentation surveys already exist such as the works by Zhu et al. [12] and Thoma [13], which do a great work summarizing and classifying existing methods, discussing datasets and metrics, and providing design choices for future research directions. However, they lack some of the most recent datasets, they do not analyze frameworks, and none of them provide details about deep learning techniques. Because of that, we consider our work to be novel and helpful thus making it a signicant contribution for the research community.

The key contributions of our work are as follows:

(1) We provide a broad survey of existing datasets that might be useful for segmentation projects with deep learning techniques.

(2) An in-depth and organized review of the most signicant methods that use deep learning for semantic segmentation, their origins, and their contributions.

(3) A thorough performance evaluation which gathers quantitative metrics such as accuracy, execution time, and memory footprint.

(4) A discussion about the aforementioned results, as well as a list of possible future works that might set the course of upcoming advances, and a conclusion summarizing the state of the art of the eld.

The remainder of this paper is organized as follows. Firstly, Section 2 introduces the semantic segmentation problem as well as notation and conventions commonly used in the literature. Other background concepts such as common deep neural networks are also reviewed. Next, Section 3 describes existing datasets, challenges, and benchmarks. Section 4 reviews existing methods following a bottom-up complexity order based on their contributions. This section focuses on describing the theory and highlights of those methods rather than performing a quantitative evaluation. Finally, Section 5 presents a brief discussion on the presented methods based on their quantitative results on the aforementioned datasets. In addition, future research directions are also laid out. At last, Section 6 summarizes the paper and draws conclusions about this work and the state of the art of the eld.

2. Backgrounding

2.1. Common Deep Network BackBone

2.1.1 VGG

VGG

2.1.2 ResNet

ResNet

2.2. Common Deep Network Architecture

2.2.1 FCN

FCN

2.2.2 SegNet

SegNet

2.2.3 U-Net

U-Net

2.2.4 DeepLab

DeepLab

2.3. Common Method for Video

2.3.1 Optical Flow

Optical Flow

2.4. Common Method for One-Shot

2.4.1 Fine-Tune

Fine-Tune

2.4.2 Mask Warp

Fine-Tune

3. Datasets

3.1. Datasets

Youtube-Objects [?]

DAVIS [?, ?, ?]

SegTrack [6]

4. Methods

4.1. Semi-supervised VOS

test

4.2. Unsupervised VOS

test

5. Discussion

5.1. Evaluation Metrics

5.1.1 Accuracy

Region Similarity $\mathcal{J}$

Contour Accuracy $\mathcal{F}$

Temporal Stability $\mathcal{T}$

5.1.2 Execution Time

5.2. Results

5.2.1 Single Object Semi-supervised VOS

5.2.2 Multiple Objects Semi-supervised VOS

5.2.3 Unsupervised VOS

5.3. Summary

5.4. Future Research Directions

6. Conclusion

References

[1] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang. Segflow: Joint learning for video object segmentation and optical flow.

	NLC [2]	CUT [4]	FST [8]	SFL [1]	LMP [9]	FSEG [3]	LVO [10]	ARP [5]	IET [7]
$\mathcal{J}$ Mean	55.1	55.2	55.8	67.4	70.0	70.7	75.9	76.2	78.5
$\mathcal{F}$ Mean	52.3	55.2	51.1	66.7	65.9	65.3	72.1	70.6	75.5

Table 1. The result of unsupervised methods on the DAVIS 2016 validation set.

In *ICCV*, 2017.

[2] A. Faktor and M. Irani. Video segmentation by non-local consensus voting. In *BMVC*, 2014.

[3] S. Jain, B. Xiong, and K. Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, 2017.

[4] M. Keuper, B. Andres, and T. Brox. Motion trajectory segmentation via minimum cost multicuts. In *ICCV*, 2015.

[5] Y. J. Koh and C.-S. Kim. Primary object segmentation in videos based on region augmentation and reduction. In *CVPR*, 2017.

[6] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013.

[7] S. Li, B. Seybold, A. Vorobyov, A. Fathi, Q. Huang, and C.-C. J. Kuo. Instance embedding transfer to unsupervised video object segmentation. In *CVPR*, 2018.

[8] A. Papazoglou and V. Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013.

[9] P. Tokmakov, K. Alahari, and C. Schmid. Learning motion patterns in videos. In *CVPR*, 2017.

[10] P. Tokmakov, K. Alahari, and C. Schmid. Learning Video Object Segmentation with Visual Memory. In *ICCV*, 2017.