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

Throught the years, there are so many excellent video segmentation datasets

3.1. Datasets

Youtube-Objects [?] TheYoutube-Objects is a database of videos collected from YouTube which contain objects from ten PASCAL VOC classes: aeroplane, bird, boat, car,cat, cow, dog, horse, motorbike, and train. That database does not contain pixel-wise annotations but Jain et al. [**?**] manually annotated a subset of 126 video sequences. They took every 10th frame from those sequences and generated semantic labesl. That totals 10167 annotated frames at 480×360 pixels resolution.

DAVIS [?, ?, ?] Densely-Annotated VIdeo Segmentation is a difficult challenge, which is mainly purposed for video object segmentation. Frame resolution varies across sequences but all of them were downsampled to 480 p for the challenge. Pixel-wise annotations are provided for each frame for four different categories: human, animal, vehicle, and object. DAVIS-2016 [**?**]is composed by 50 high-definition sequences which add up to 2079 and 1376 frames for training and validation respectively. This dataset mainly addresses problem on primary objects segmentaion, which means that there are at least one target foreground object in videos. DAVIS-2017 [**?**] have extended the number of sequences to 90, which is 60 sequences for training and 30 for validation, which contains 4209 and 1999 frames respectively. They segment the main moving objects in the scene and divide them by their semantic, even though they might have the same motion. So we can refer DAVIS-2017 as the video instance segmentations.

SegTrack-]v2 [?] The dataset is an updated version of the SegTrack dataset, which provide more additional annotations of objects for other individual objects. The total 14 videos contain 1066 frames with pixel- level annotations.

4. Methods

4.1. Semi-supervised VOS

test

4.2. Unsupervised VOS

test

5. Discussion

5.1. Evaluation Metrics

In a supervised evaluation framework, given a groundtruth mask G on a particular frame and an output segmentation M, any evaluation measure ultimately has to answerthe question how well M fits G. As justified in [**?**], for images one can use two complementary points of view, regionbased and contour-based measures. As videos extends the dimensionality of still images to time, the temporal stability of the results must also be considered.

5.1.1 Accuracy

Region Similarity $\mathcal{J}$ To measure the region-based segmentation similarity, i.e. the number of mislabeled pixels, one employ the Jaccard index $\mathcal{J}$ defined as the intersectionover-union of the estimated segmentation and the groundtruth mask. The Jaccard index has been widely adopted since its first appearance in PASCAL VOC2008 [**?**], as it provides intuitive, scale-invariant information on the number of mislabeled pixels. Given an output segmentation M and the corresponding ground-truth mask G it is defined as $\mathcal{J} = \frac{|M \cap G|}{|M \cup G|}$

Contour Accuracy $\mathcal{F}$ From a contour-based perspective, one can interpret M as a set of closed contours $c(M)$ delimiting the spatial extent of the mask. Therefore, one can compute the contour-based precision and recall Pc and Rc between the contour points of $c(M)$ and $c(G)$, via a bipartite graph matching in order to be robust to small inaccuracies, as proposed in [**?**]. So the F-measure F is a good trade-off between two, which is defined as $F = \frac{2PcRc}{Pc+Rc}$.

Temporal Stability $\mathcal{T}$ Temporal stability $\mathcal{T}$. Intuitively, $\mathcal{J}$ measures how well the pixels of the two masks match, while F measures the accuracy of the contours. However, temporal stability of the results is a relevant aspect in video object segmentationsince the evolution of object shapes is an important cue for recognition and jittery, unstable boundaries are unacceptable in video editing applications.

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

Dataset	Metrics	NLC [?]	CUT [?]	FST [?]	SFL [?]	LMP [?]	FSEG [?]	LVO [?]	ARP [?]	IET [?]
DAVIS	$\mathcal{J}$ 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
SegTrack-V2	$\mathcal{J}$ Mean	52.3	55.2	51.1	66.7	65.9	65.3	72.1	70.6	75.5
	$\mathcal{F}$ Mean	52.3	55.2	51.1	66.7	65.9	65.3	72.1	70.6	75.5
Youtube-Objects	$\mathcal{J}$ Mean	52.3	55.2	51.1	66.7	65.9	65.3	72.1	70.6	75.5
	$\mathcal{F}$ 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 Video Objects Segmentation datasets .