A Review of Deep Learning Based Methods for Unsupervised Video Object Segmentation

1 INTRODUCTION

The task of video object segmentation aims to seperate the pixels corresponding to objects of interest from the rest in consecutive frames of a video, which is a prerequisite for various high-level applications, including autonomous driving[REF_], video surveillance[REF_], video editing[REF_], action recognition[REF_] and robot vision[REF_]. While supervised segmentation methods show great effectiveness[REF_], most of them require massive pixel-wise human annotations on video frames, which is expensive and time-consuming, in contrast, self-supervised methods [REF_] recently has started attracting attention by learning from unlabeled video sequences. Considering the selection of the objects of interest, the segmentation task is divided to three categories: semi-supervised video object segmentation (SVOS, also called one-shot VOS), interactive video object segmentation, and unsupervised video object segmentation (UVOS, also called zero-shot VOS). While semi-supervised video object segmentation requires the explicit mask of the interest object for a video frame at test time, interactive video object segmentation segments the objects of interest with the iterative refinement inputs on the target objects from the user interaction, and unsupervised video object segmentation(UVOS) attempts to automatically identify the primary object pixel-wise in the video consequencet. Due to the lack of any prior knowledge about the target objects, in addition to the challenges for SVOS and interactive VOS, for instance, object deformation, occlusion, and background clutters, UVOS suffers from how to correctly distinguish the foreground objects from the complex background.

In this article, I mainly focus on the state-of-the-art deep learning based methods for UVOS task to summarize their mechanism and architecture, advantage and deficiency, and compare their performance.

2 MAJOR METHODS

Since PSPNet[1] and DeepLabv3[2] achieved state-of-the-art performance on various datasets of semantic segmentation task in late 2016, and the publish of DAVIS 2016 dataset[3], video object segmentation has become a breakout.

2.1 Pyramid Dilated Deeper ConvLSTM

In 2018, Song et al.[REF_PDB] suggested a modified ConvLSTM for fast video salient object detection. The model consists of four components, a ResNet-50 like backbone, a Pyramid Dilated Convolution (PDC) module, a PDB-ConvLSTM module, and a segmentation module. The PDC module constitutes four dilated convolutions with different dilation rates for explicitly extracting spatial saliency features in the same size but with different receptive field sizes. The multi-scale spatial features are then concatenated together with the original input feature to feed the two parallel DB-ConvLSTMs that are both implemented with 3x3x32 kernels for interpreting the temporal features of video frames and fusing spatial and temporal features automatically. The output of the DB-ConvLSTM branches in PDB-ConvLSTM are further concatenated as a multi-scale spatiotemporal saliency feature to feed a 1x1x1 convolution layer with a sigmoid activation for generating a binary mask which is at last upsampled to the size of the video frame via bilinear interpolation.

Compared to the previous shallow, parallel bi-directional feature extraction strategy [REF_2015 Convolutional LSTM network Shi, X., Chen, Z., Wang, H.,], the PDB-ConvLSTM improves it with a deeper and cascaded learning process to bidirectionally capture information of both the forward and backward frames in a video.

The training procedure has three steps, the PDC module and the ResNet are pre-trained using two image saliency datasets: MSRA10K and DUT-OMRON, and the training set of DAVIS16. And then use the static and video data to train the entire model. After that, the parameters of PDC are fixed and only the PDB-ConvLSTMs are learned with the training set of DAVIS16. The loss function is inspired by [REF_2018 CVPR Salient object detection driven by fixation prediction] and denoted as:

$$\mathcal{L}(\mathbf{S},\mathbf{G}) = \mathcal{L}_{cross\_entropy}(\mathbf{S},\mathbf{G}) + \mathcal{L}_{MAE}(\mathbf{S},\mathbf{G})$$

Where $S \in [0,1]^{473x473}$ is the prediction, and $G \in [0,1]^{473x473}$ is the groundtruth.

This method with a CRF post-process achieved state-of-the-art results on DAVIS16 and FBMS in 2018.

2.2 Co-Attention Siamese Network

Lu et al.[REF_COSNET] proposed an end-to-end co-attention method to learn the global correlations and scene context in a video in 2019. This method is implemented with a siamese network which consists of three cascaded parts and a CRF post-process. The three parts are a DeepLabv3 based feature embedding module, a co-attention module, and a segmentation module. The DeepLabv3 backbone is pre-trained with MSRA10K and DUT saliency datasets.

In the training phase, the DeepLabv3 based siamese network takes two streams (a query frame and a random sampled reference frame from the same video) as input to build the feature representations. The output are then refined by the co-attention module which first learns the normalized similarities between the feature representations, second computes the attention summaries with the feature representations and the normalized similarities, and lastly feeds the co-attention enhanced features to a 1x1 convolutional layer to weight information from different input frames. The vanilla co-attention module was implemented using a fully connected layer with 512x512 parameters, while the channel-wise co-attention is built on a Squeeze-and-Excitation-like module. Following the co-attention module, the features are passed into a segmentation network that consists of four convolutional layers and a sigmoid activation to generate the binary mask for the query frame. After all, the binary mask is post-processed by CRF. The loss function is a weighted binary cross entropy.

$$\mathcal{L}_{\mathcal{C}}(\mathbf{Y},\mathbf{O}) = -\sum\nolimits_{x}(1-\eta)o_x\log(y_x)+\eta(1-o_x)\log(1-y_x)$$

Where Y is the prediction, O is the groundtruth.

In the ablation study from the paper, without the co-attention module, the mean region similarity drops 9.2% on DAVIS16, 5.5% on FBMS, and 7.6% on Youtube-Objects.

2.3 Focus on Foreground Network

Liu et al.[REF_F2Net] proposed a Focus on Foreground Network(F2Net) method of considering the center point as the spatial prior guidance of the primary object and then segment the mask from such point to its surroundings. The implementation network consists of three main parts: Siamese Encoder Module for extracting features of the reference frame and the query frame, Center Guiding Appearance Diffusion Module for capturing the inter-frame and intra-frame feature regarding the prediction of the center location of the foreground object to lead the next stage work to focus on the foreground, and Dynamic Information Fusion Module with Attention to select relatively important features providing more optimal representations for final segmentation.

The siamese encoder takes a pair of RGB images as inputs, including a query frame and a reference frame (the first frame of the video is taken by assuming the first frame always contains the foreground objects).

The center point prediction problem is considered as a heatmap estimation task with two steps: Feature Upsampling and Heatmap Generation. Feature Upsampling merges features in different scales to enhance the information for high-resolution features. Heatmap Generation estimate the heatmap by learning the affine transformation of the gauss map from the previous frame and accumulating it with the semantic information from the query frame. The center point is chosen among top score points by comparing the distances to the center of the previous frame.

Spatial-Prior Guided Appearance Diffusion aims to determine the foreground object with two considerations, 1) distinguishable in an individual frame (locally saliency), and 2) frequently appearing throughout the video sequence (globally consistent). To achieve the first goal, a non-local operation on the current feature is applied. To achieve the second goal, another non-local operation on both current and the reference features is utilized for alleviating appearance drift. Compared to other appearance matching methods, the key difference of this method is that the encoded feature representation is weighted according to their similarity with the foreground determined by the center point.

Since the original feature only contains a coarse clue for inferring the foreground object. Intra-frame feature ontains more accurate salient object information in current frame, but not the appearance changes in a video sequence. Inter-frame feature contains more contexts about appearance changes of the foreground objects, Dynamic Information Fusion module selectively aggregate them with Channel-wise attention implemented with four FC layers and followed by Spatial-wise attention implemented with a 1x1 convolutional layer.

After the channel-wise and spatial-wise attention modules, the information flows from three-level input features are aggregated into one feature map. And then a decoder which consists of two convolutional layers is employed to generate the final binary mask result.

Similar to COSNet, there are two alternated steps to in training process. In the static-image iteration, The image saliency dataset MSRA10K is employed to fine-tune the DeepLabV3 and the Center Prediction Branch. In the dynamic-video iteration, the whole model is trained with the training set in DAVIS16. The loss function consists of two parts, the loss of the centre point heatmap estimation, and the loss of the mask prediction.

Method	Pub	Backbone	Extra Training Dataset	Components	Optical Flow	Post-process
PDB	ECCV2018	ResNet-50	MSRA10K, DUT-OMRON	Backbone, PDC, PDB-ConvLSTM, Segmentation	No	No
PDB+	ECCV2018	ResNet-50	MSRA10K, DUT-OMRON	Backbone, PDC, PDB-ConvLSTM, Segmentation	No	CRF
COSNet	CVPR2019	DeepLabv3	MSRA10K, DUTS	Backbone, Co-attention, Segmentation	No	CRF
MuG						
AnDiff	ICCV2019					
DFNet	ECCV2020					
GMNet						
MANet						
MATNet	AAAI2020					
F2Net	AAAI2021	DeepLabv3	MSRA10K	Backbone, Center Prediction, Center-Guiding Appearance Diffusion, Attention, Segmentation	No	No

Dataset	Measures	PDB	PDB+	COSNet	AnDiff		DFNet		MATNet	F2Net

DAVIS16	J	Mean ↑	74.3	77.2	80.5	81.7		83.4		82.4	83.1
		Recall ↑		90.1	94.0*	90.9				94.5	95.7
		Decay ↓		0.9	0.0*	2.2				5.5	0.0
	F	Mean ↑	72.8	74.5	79.4*	80.5		81.8		80.7	84.4
		Recall ↑		84.4	90.4*	85.1				90.2	92.3
		Decay ↓		-0.2	0.0*	0.6				4.5	0.8
	T	Mean↓		29.1	31.9*	21.4		15.9		21.6	20.9
DAVIS17	J	Mean ↑									
		Recall ↑									
		Decay ↓									
	F	Mean ↑									
		Recall ↑									
		Decay ↓									
	T	Mean↓									
FBMS	J	Mean ↑	72.3	74.0	75.6					76.1	77.5
	F	Mean ↑		81.5		81.2		82.3			
	T	Mean↓									
Youtube-Objects	J	Mean ↑		65.4	70.5					69.0	75.6
		Recall ↑									
		Decay ↓									
	F	Mean ↑									
		Recall ↑									
		Decay ↓									
#Parameter (M)											
Time (s/image)			0.05	0.5-1	0.45			0.28			

Overall results of Region Similarity (J), Contour Accuracy (F) and Temporal Stability (T) for methods. The higher numbers mean the better performance for rows having an upward arrow in the headers (e.g, recall), and vice versa.
PDB+ means PDB with CRF post-process

3 DATASETS AND METRICS

4 CONCLUSION

REFERENCES

Xiankai Lu, Wenguan Wang, Jianbing Shen, Yu-Wing Tai, David Crandall, Steven C. H. Hoi in CVPR2020, Learning Video Object Segmentation from Unlabeled Videos

Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, Weidi Xie in 2021, Self-supervised Video Object Segmentation by Motion Grouping

Zihang Lai, Weidi Xie in BMVC 2019, Self-supervised Learning for Video Correspondence Flow