The 2019 DAVIS Challenge on Video Object Segmentation - CVPR 2019 Workshops

Discriminative Learning and Target Attention for the 2019 DAVIS Challenge on Video Object Segmentation

Andreas Robinson* Linköping University
Felix Järemo Lawin* Linköping University
Martin Danelljan ETH Zürich
Michael Felsberg Linköping University

Abstract

In this work, we address the problem of semi-supervised video object segmentation, where the task is to segment a target object in every image of the video sequence, given a ground truth only in the first frame. To be successful it is crucial to robustly handle unpredictable target appearance changes and distracting objects in the background. In this work we obtain a robust and efficient representation of the target by integrating a fast and light-weight discriminative target model into a deep segmentation network. Trained during inference, the target model learns to discriminate between the local appearances of target and background image regions. Its predictions are enhanced to accurate segmentation masks in a subsequent refinement stage.

To further improve the segmentation performance, we add a new module trained to generate global target attention vectors, given the input mask and image feature maps. The attention vectors add semantic information about the target from a previous frame to the refinement stage, complementing the predictions provided by the target appearance model. Our method is fast and requires no network fine-tuning. We achieve a combined J and F-score of 70.6 on the DAVIS 2019 test-challenge data.

1. Introduction

Video object segmentation is the problem of segmenting the image regions occupied by specific target objects, in every frame of a given video sequence. With numerous applications in robotics, surveillance, autonomous driving and action recognition it is currently seeing an increasing commercial interest. In this work we focus on the semi-supervised setting, where the ground-truth mask of the object are given only in the first frame. Videos generally contain many challenging aspects such as changes of appearance in moving targets, occlusions and distracting objects in the background.

To successfully handle these challenges, a robust and adaptive target representation is required. Earlier methods such as [1, 8] achieve target adaption by fine-tuning a deep neural network on the first frame. Although having the potential to be very accurate, these approaches struggle with appearance changes in the more challenging scenarios and suffer from very low frame-rates.

In contrast, the recently proposed method in [7] performs target adaptation through a dedicated light-weight target appearance model, carefully integrated into a deep neural network for object segmentation. The appearance model is efficiently trained during inference (online) to discriminate between the target and the background, whereas the deep network is trained offline to refine the predictions from the target appearance model. Consequently, the refinement network remains target agnostic and retains powerful and generic segmentation functionalities.

The target appearance model aggregates local information of both target and the background. During inference however, only the spatial prediction from the current frame is used in target segmentation. In this work, we modify the approach in [7] by adding a target attention module. The module generates global target attention vectors given an input target mask prediction and corresponding image feature maps. These attention vectors provide a global semantic representation to the model given the previous target appearance, complementing the local representation in the online target model. As a result, the robustness to false predictions is increased without significantly affecting the inference time. In addition we improve the performance by adding a new upsampling layer.

Our method is efficient and operates without any additional post-processing, optical flow, or other external components. We evaluate our method on the DAVIS 2019 test-challenge dataset and include an ablative study on DAVIS 2019 test-dev. Our proposed method achieve a combined J and F-score of 70.6, improving over the baseline in [7] by over 5 percentage points.

2. Method

In this work, we represent the target in a discriminative target appearance model and a global target attention mod-

[1] Authors contributed equally

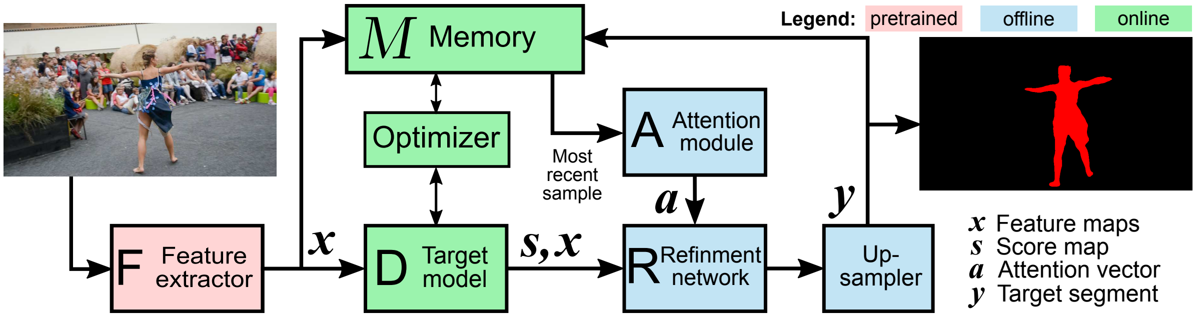

Figure 1. Overview of our video segmentation architecture, consisting of a feature extractor F, a target appearance model D, a target attention module A and a refinement network R. First, the pre-trained feature extractor F outputs feature maps $\mathbf{x}$ from the incoming image, at four different depths. Next, the second deepest feature map is processed by the target appearance model D, generating a coarse segmentation score map $\mathbf{s}$. The score map $\mathbf{s}$, the deep features $\mathbf{x}_i$ and attention vectors $\mathbf{a}$ from A are processed in R, producing the output segmentation mask $\mathbf{y}$. Finally, $\mathbf{y}$ and $\mathbf{x}$ are stored in the memory M, which is used to update the target model D. We periodically update D using the memory samples and feed the most recent feature map and segmentation mask through A, to produce new target attention vectors $\mathbf{a}$.

ule, integrating both in a deep object segmentation network. The target appearance model is efficiently trained in real-time during inference with Gauss-Newton conjugate gradient optimization. Its low-resolution predictions are combined with attention vectors from the attention module in a deep refinement network to produce high-resolution segmentation masks. Both the target attention module and refinement network are trained offline.

Our architecture is shown in figure 1. During inference, we are first given an initial video frame and the corresponding ground truth mask. We extract features $\mathbf{x}$ with the feature extractor F and train the target appearance model D. In addition, global attention vectors $\mathbf{a}$ are predicted with the target attention module A. In subsequent frames, the target model first predicts a coarse score map $\mathbf{s}$ from incoming frames. The score map and the attention vectors are then passed to the refinement network, which outputs the final segmentation $\mathbf{y}$. We store $\mathbf{y}$ and the corresponding features $\mathbf{x}$ in the memory M and process them with A to produce new attention vectors. Finally, we use M to update D before the next incoming frame.

The individual components are described in more detail in the following sections below.

2.1. Discriminative target appearance model

The aim in [7] is to provide an powerful discriminative representation of the target. To allow for fast adaptation to appearance changes, the target model $D(\mathbf{x};\mathbf{w})$ is designed to be as light-weight as possible. Specifically, it is implemented as $D(\mathbf{x};\mathbf{w}) = \mathbf{w}_2 * (\mathbf{w}_1 * \mathbf{x})$, where the parameters $\mathbf{w}_1$ is a 1x1 channel-reducing filter and $\mathbf{w}_2$ is a 3x3 kernel with one output channel. Following [7], the weights $\mathbf{w}$ are trained by minimizing the loss

$$\mathcal{L}_D(\mathbf{w};M) = \sum_k \gamma_k \|\mathbf{v}_k(\mathbf{y}_k - U(D(\mathbf{x}_k))\|^2 + \sum_j \lambda_j \|\mathbf{w}_j\|^2, \tag{1}$$

where the parameters λ_j control the regularization term and $\mathbf{v}_k$ are weight masks balancing the impact of target and background pixels. U is a bilinear up-sampling operator applied to the output from the target model for recovering the spatial resolution of the labels $\mathbf{y}_k$. The loss is minimized wrt the dataset $M = \{(\mathbf{x}_k, \mathbf{y}_k, \gamma_k)\}_{k=1}^{K}$ consisting of sample feature maps $\mathbf{x}_k$, target labels $\mathbf{y}_k$ and sample weights γ_k. For this purpose, we use the fast and robust Gauss-Newton conjugate gradient (GNCG) optimization strategy proposed in [3]. During inference, new samples from the video sequence are added to the M and the model is updated by minimizing (1). The influence of new samples are controlled by setting parameter γ_k to an appropriate value.

2.2. Refinement network

The discriminative target model outputs coarse segmentation score maps. To recover a label map of higher resolution, a deep refinement network R is employed. The network has two types of building blocks: a target segmentation encoder (TSE) and a refinement module. These are used to construct a U-Net based architecture for object segmentation as in [10].

The refinement network takes features maps $\mathbf{x}^d$ from multiple 'layers' d in the backbone network F as input. Each layer has a corresponding refinement block consisting of a TSE and a refinement module. Refinement proceeds in a bottom-up fashion, operating on backbone features, one layer at a time. First, a TSE merges the current layer's feature map and the target appearance model score map. It projects the backbone features to 64 channels to reduce the subsequent computational complexity. After projection, the features and the segmentation score $\mathbf{s}$ are concatenated and encoded by three convolutional layers. The encoded features are then processed by a refinement module with two residual blocks (RRB) and a channel attention block (CAB). The first RRB block allows for further transformations of the encoded features from the TSE. The CAB integrates spatially global information into the stream, by pooling features from the current and next deeper layers. With these sources of attention, a channel-wise modulation is performed on the stream of features from the TSE and RRB. The second RRB transforms the features further before passing them up to the next refinement block.

Finally, a representation of the segmentation emerges from the last RRB and is upsampled to the resolution of the input video. For this purpose we use a bilinear upsampler followed by a learned post-processing layer to correct for deficiencies in the bilinear function.

2.3. Target attention module

The target appearance model D is trained to track the target regions based on local appearance. We complement this with global attention vectors generated from a previous frame by the target attention module A. In order to include both deep and shallow information in the target representation we apply our target attention module at each layer d in the backbone feature extractor F. For each refinement block in R, we use a TSE block to merge the past feature map and its corresponding segmentation. A target appearance vector $\mathbf{a}^d$ is then formed through global average pooling and transformation in a subsequent a pair of convolutional layers. Finally, we integrate this attention vector in the refinement network by concatenating it with the other pooled features in the CAB.

2.4. Inference

In the inference stage, the discriminative target model is trained by minimizing the loss in eq. (1). We generate a training dataset M_0, from the the first frame in the video sequence given the ground truth mask. By using data augmentation we include 19 additional frames. We also generate initial target attention vectors $\mathbf{a}_0$ from the initial frame and $\mathbf{y}_0$ with the target attention module A.

In the subsequent video frames, D first predicts a coarse segmentation mask $\mathbf{s}$. These are upsampled with the refinement network R, supported by the target descriptors $\mathbf{a}_0$ as described in section 2.3. The refinement network outputs a high resolution segmentation mask $\mathbf{y}_i$, which along with the features $\mathbf{x}_i$ are used to update the dataset M_i. Following [7], the new sample $(\mathbf{x}_i, \hat{\mathbf{y}}_i, \gamma_i)$ is provided an update weight. Every four or eight frames, D is updated by re-optimization with the current dataset M_i. In addition, new attention vectors $\mathbf{a}_i$ are generated from $\mathbf{x}_i$ and $\mathbf{y}_i$.

The framework generalizes to multi-object segmentation with one instance of M, D and $\mathbf{a}$ for each target. The individual segmentations are then merged according to [5].

2.5. Network training

We train the network formed by R and A on annotated video sequences, while keeping the backbone feature extractor. Each training sample is assembled from three random frames (feature maps and a target segmentation mask) from the same sequence. One frame is used to train the target appearance model D, while the other two are passed through D, A and R as in the inference mode. Unlike inference however, neither the target appearance model dataset nor the attention vectors are updated during training.

3. Experiments

Training is performed as in [7] on the training splits from DAVIS 2017 [6] and YouTube-VOS [9], although we let all experiments except the baseline run for 240 epochs. However, training seems to converged at 180 epochs.

We evaluate our method on the DAVIS 2019 test-dev and DAVIS 2019 [2] challenge datsets. The impact of the various components presented in this paper are tested on the DAVIS 2019 test-dev dataset, results are shown in table 1.

Our baseline is the method from [7]. It is configured with one target appearance model, working on features from ResNet-101 [4] ”conv4_x” (stride 16, 1024 channels), with 96 output channels in $\mathbf{w}_1$. The refinement network takes features from the max pooling output in ”conv2_x” and the outputs of blocks ”conv2_x” through ”conv5_x”. Further, it lacks the new target attention module and employs basic bilinear interpolation in the final upsampling step.

In subsequent experiments, we increase the number of output channels in $\mathbf{w}_1$ to 128 (’128ch’ column in table 1), add the improved upsampler from section 2.2 (the ’up’ column) and include the target attention module (A). Further, we remove the max pooling output in ”conv2_x” in the refinement. Our DAVIS 2019 challenge method have all these features enabled.

Table 1 shows the result of the above mentioned modifications on the DAVIS test-dev dataset. By increasing the output channels in $\mathbf{w}_1$ to 128 channels the $\mathcal{J}\&\mathcal{F}$ score is improved by 0.2 points. Further improvements are acheived by adding the new upsampler, increasing the score by 1.0 points compared to the baseline. Finally, our challenge-variant generates the highest $\mathcal{J}\&\mathcal{F}$ score at 63.1, owing to our proposed attention module.

Detailed DAVIS 2019 challenge results are reported in table 2. The challenge-variant has a global mean of 69.5%, scoring 2.4 percentage points higher than the baseline, thus motivating the use of the additional components. In addition, by reducing the number of frames between model updates (see section 2.4) from eight to four we achieve the highest score at 70.6.

Variant	ch	up	A	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
baseline	128			61.3	58.4	64.2
baseline	96			61.3	58.1	64.4
baseline	128			61.5	58.1	64.8
upsampler	128	✓		62.3	59.0	65.6
challenge	128	✓	✓	**63.1**	**60.0**	**66.2**

Table 1. Ablation study on the DAVIS 2019 test-dev dataset [2]. Adding more channels and the new upsampler sightly increase the accuracy of the baseline method in [7]. Including the proposed attention module achieves the best result with a $\mathcal{J}\&\mathcal{F}$ at 62.3, improving by 2 points over the baseline.

	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$			$\mathcal{F}$		
Variant	mean	mean	recall	decay	mean	recall	decay
baseline	67.1	65.2	74.9	**20.0**	69.1	79.6	**22.9**
challenge	69.5	67.3	76.7	20.1	71.6	82.4	23.4
challenge-4	**70.6**	**68.5**	**78.1**	20.3	**72.8**	**84.2**	24.0

Table 2. Results from the DAVIS 2019 challenge [2]. We compare the baseline from [7] to our challenge version including the attention module in section 2.3. The challenge method improves both the $\mathcal{J}$ and the $\mathcal{F}$ scores compared to the baseline. Additionally we include the challenge-4 method where the model update rate is lowered from eight to four frames, reaching a $\mathcal{J}\&\mathcal{F}$ score of 70.6.

3.1. Qualitative results

We provide qualitative results are in figure 2. Reflecting the challenge scores, the differences between baseline and challenge frames are subtle but noticeable; The challenge variant (rightmost column) appears to produce more defined segmentations as shown in bike-packing, where the wheel axles are more defined. However, it incorrectly removes part of the real wheel. It also seems to handle small details better, correctly segmenting the green goldfish's tail and the yellow material falling into large carriage. In addition, it correctly segments the boy's head while the baseline (center image) marks it as part of the girl.

4. Conclusion

We present a method for video object segmentation, tracking the target with an attention module and a lightweight discriminative target model, integrating both into a deep object segmentation network. Guided by attention vectors and low-resolution score maps, the segmentation network creates high resolution target masks. Our approach achieves an overall score of 70.6 on the DAVIS 2019 challenge.

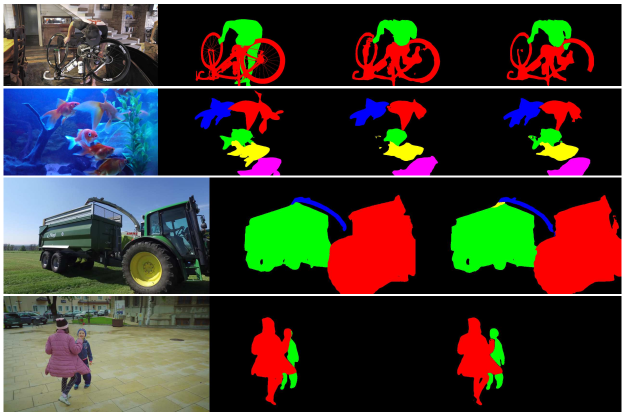

Figure 2. Qualiative examples from DAVIS 2017 validation set (first two rows) - left-to-right: image, ground truth, baseline, challenge-4 variant outputs. DAVIS 2019 test-challenge (last two rows) left-to-right: image, baseline, challenge-4 variant outputs

References

[1] Sergi Caelles, K-K Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR 2017*, pages 5320–5329. IEEE, 2017.

[2] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019.

[3] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. ATOM: Accurate tracking by overlap maximization. In *CVPR*, 2019.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[5] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR 2018*, pages 7376–7385. IEEE, 2018.

[6] 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. In *Computer Vision and Pattern Recognition*, 2016.

[7] Andreas Robinson, Felix Järemo-Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Discriminative online learning for fast video object segmentation, 2019.

[8] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017.

[9] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *European Conference on Computer Vision*, pages 603–619. Springer, 2018.

[10] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. *arXiv preprint arXiv:1804.09337*, 2018.