The 2018 DAVIS Challenge on Video Object Segmentation - CVPR 2018 Workshops

Video Object Segmentation with Joint Re-identification and Attention-Aware Mask Propagation

Xiaoxiao Li Chen Change Loy
Department of Information Engineering, The Chinese University of Hong Kong
{lx015,ccloy}@ie.cuhk.edu.hk

Abstract

The problem of semi-supervised video object segmentation, in the context of the 2018 DAVIS Challenge [2], can become extremely challenging when multiple instances coexist. While each instance may exhibit large scale and pose variations, the problem is compounded when instances occlude each other causing failures in tracking. In this study, we formulate a deep recurrent network that combines temporal propagation and re-identification functionalities into an end-to-end framework. In particular, we present a re-identification module with template expansion to retrieve missing objects despite their large appearance changes. Besides, we contribute an attention-based recurrent mask propagation approach that is robust to distractors not belonging to the target segment. In order to increase the output resolution, we further modify the input of DyeNet from full-image to bounding box and use the feature pyramid technique. Finally, we achieve a competitive global mean (Region Jaccard and Boundary F measure) of 73.8 in the 2018 DAVIS Challenge[1].

1. Introduction

Semi-supervised video object segmentation [10] aims at tracking the foreground objects from the background region in the video sequence, given their ground-truth masks in the first frame. A notable and challenging benchmark for this task is 2018 DAVIS Challenge [2]. An example of a sequence is shown in Fig. 1. The DAVIS dataset presents real-world challenges that need to be solved from two key aspects. First, since there are multiple instances in a video, it is likely that they will occlude each other. Second, instances typically experience substantial variations in both scale and pose across frames.

To address the occlusion problem, notable studies such as [1, 13] adapt generic semantic segmentation deep model to the task of specific object segmentation. Often, a fixed set of templates such as the masks of target objects in the first frame are used for matching targets. This paradigm fails in some challenging cases (see Fig. 1(a)), as using a fixed set of templates cannot sufficiently cover large scale and pose variations. To mitigate the variations in both scale and pose across frames, existing studies [9, 8] exploit temporal information to maintain continuity of individual segmented regions across frames. On unconstrained videos with severe occlusions (see Fig. 1(b)), approaches based on temporal continuity are prone to errors.

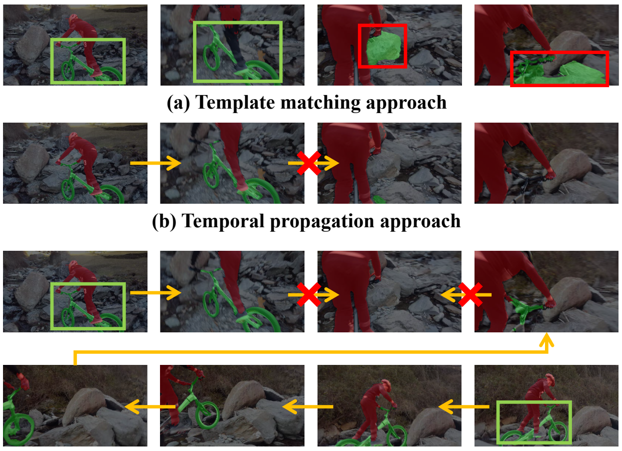
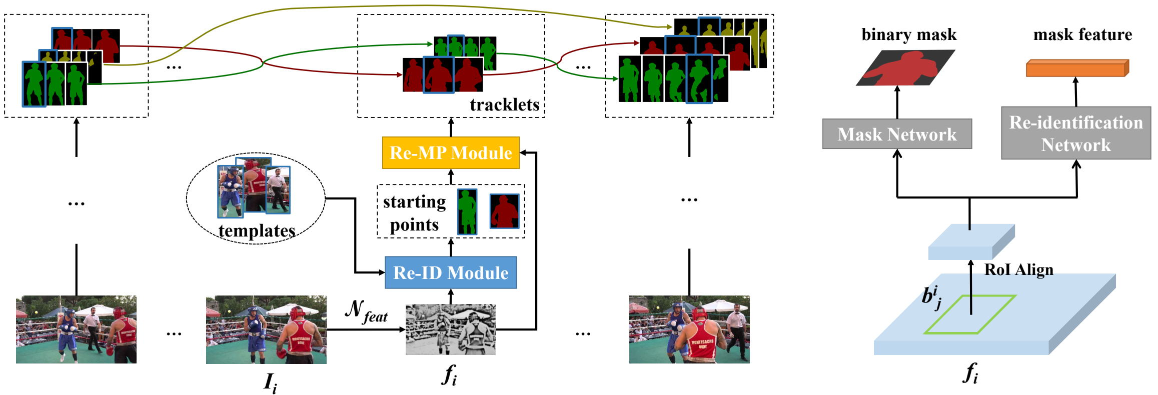

(a) Template matching approach
(b) Temporal propagation approach
(c) DyeNet

Figure 1: In this example, we focus on the bicycle. (a) shows the result of template matching approach which is affected by large scale and pose variations. As shown in (b), the temporal propagation approach is incapable of handling occlusion. The proposed DyeNet joints them into a unified framework, first retrieves high confidence starting points and then propagates their masks bidirectionally to address those issues. The result of DyeNet is visualized in (c). **Best viewed in color.**

In this study, we bring template matching and temporal propagation approaches into a single unified network. Our network hinges on two main modules, namely a re-identification (Re-ID) module and a recurrent mask propagation (Re-MP) module. The Re-ID module helps to es-

[1] An earlier and more complete report of this study is available at https://arxiv.org/abs/1803.04242



(a) DyeNet

(b) Re-ID Module

Figure 2: (a) The pipeline of DyeNet. The network hinges on two main modules, namely a re-identification (Re-ID) module and a recurrent mask propagation (Re-MP) module. (b) The network architecture of the re-identification (Re-ID) module. **Best viewed in color.**

tablish confident starting points in non-successive frames and retrieve missing segments caused by occlusions. Based on the segments provided by the Re-ID module, the Re-MP module propagates their masks bidirectionally by a recurrent neural network to the entire video. Besides, a new attention mechanism makes the Re-MP module more resilient to distractors. In addition, the Re-ID and Re-MP steps are conducted in an iterative manner, which allows us to identify confidently predicted mask in each iteration and expand the template set in the Re-ID module. The process of conducting Re-ID followed by Re-MP may be imagined as dyeing a fabric with multiple color dots (*i.e.*, choosing starting points with re-identification) and the color disperses from these dots (*i.e.*, propagation). Drawing from this analogy, we name our network as *DyeNet*. As shown in Fig. 1(c), DyeNet is capable of segmenting multiple instances across a video with high accuracy through Re-ID and Re-MP.

In the 2018 DAVIS Challenge, we modify the architecture of original DyeNet (*e.g.*, two-stream network, feature pyramid) to further improve the performance. Finally, we achieve a competitive global mean (Region Jaccard and Boundary F measure) of 0.738.

2. Methodology

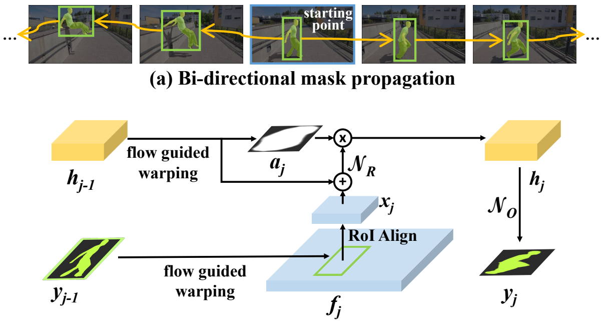
We provide an overview of the proposed approach. Figure 2(a) depicts the architecture of DyeNet. It consists of two modules, namely the re-identification (Re-ID) module and the recurrent mask propagation (Re-MP) module. The network first performs feature extraction.

Feature extraction. Given a video sequence with N frames $\{I_1, ..., I_N\}$, for each frame I_i, we first extract a feature f_i by a convolutional feature network $\mathcal{N}_{feat}$, *i.e.*, $f_i = \mathcal{N}_{feat}(I_i)$. Both Re-ID and Re-MP modules employ the same set of features in order to save computation. Considering model capacity and speed, we use 'conv1' to 'conv4_x' of ResNet-101 [6] as the backbone of $\mathcal{N}_{feat}$. To increase the resolution of features, we decrease the convolutional strides and employ the dilated convolutions similar to [3].

Iterative inference with template expansion. After feature extraction, DyeNet runs Re-ID and Re-MP in an iterative manner to obtain segmentation masks of all instances across the whole video sequence. In the first iteration, the Re-ID module generates a set of masks from object proposals and compares them with ground-truth templates given in the first frame. Masks with a high similarity to templates are chosen as the starting points for Re-MP. Subsequently, Re-MP propagates each selected mask (*i.e.*, starting point) bidirectionally, and generates a sequence of segmentation masks, which we call tracklet. In subsequent iterations, DyeNet chooses confidently predicted masks to expand the template set and reapplies Re-ID and Re-MP. Template expansion avoids heavy reliance on the masks provided by the first frame, which may not capture sufficient pose variations of targets. In this work, DyeNet stops the iterative process when no more high-confident masks can be found by the Re-ID module. Next, we present the Re-ID and Re-MP modules.

2.1. Re-identification

Figure 2(b) illustrates the Re-ID module to search for targets in the video sequences. For the i-th frame, besides the feature f_i, the Re-ID module also requires the object proposals [12] $\{b^i_1, ..., b^i_M\}$ as input where M indicates the number of proposal bounding boxes on this frame. For each candidate bounding box b^i_j, we first extract its feature from f_i, and resize the feature into a fixed size $m \times m$ (*e.g.*, 28×28) by RoIAlign [5]. The extracted features are fed into two shallow sub-networks. The first sub-network is a mask network that predicts a $m \times m$ binary mask that represents the segmentation mask of the main instance in candidate bounding box b^i_j. The second sub-network is a re-identification network that projects the extracted features into an L2-normalized 256-dimensional subspace to obtain



(a) Bi-directional mask propagation

(b) Re-MP Module

Figure 3: (a) Illustration of bi-direction mask propagation. (b) The network architecture of the recurrent mask propagation (Re-MP) module. **Best viewed in color.**

the mask features. The templates are also projected onto the same subspace for feature extraction. By computing the cosine similarities between the mask and template features, we can measure the similarity between candidate bounding boxes and templates. If a candidate bounding box is sufficiently similar to any template, we will keep its mask as a starting point for mask propagation.

We employ 'conv5_x' block of ResNet-101 as the backbone of the sub-networks. However, some modifications are necessary to adapt them to the respective tasks. In particular, we decrease the convolutional strides in the mask network to capture more details of prediction. For the re-identification network, we keep the original strides and append a global average pooling layer and a fully connected layer to project the features into the target subspace.

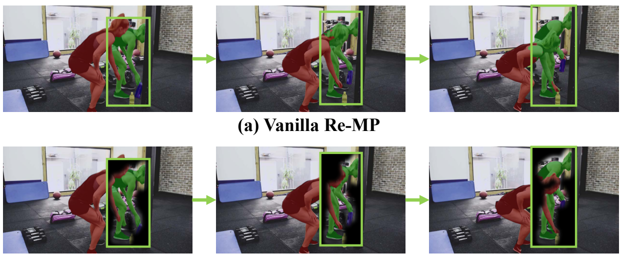
2.2. Recurrent Mask Propagation

As shown in Fig. 3(a), we bi-directionally extend the retrieved masks (*i.e.*, starting points) to form tracklets by using the Re-MP module. We formulate the Re-MP module as a Recurrent Neural Network (RNN). Figure 3(b) illustrates the mask propagation process between adjacent frames.

Suppose $\hat{y}$ is a retrieved segmentation mask for instance k in the i-th frame, and we have propagated $\hat{y}$ from i-th frame to $(j-1)$-th frame, $\{y_{i+1}, y_{i+2}, ..., y_{j-1}\}$ is the sequence of binary masks that we obtain. We now aim to predict y_j, *i.e.*, the mask for instance k in the j-th frame. In a RNN framework, the prediction of y_j can be solved as

$$
\begin{aligned}
h_j &= \mathcal{N}_R(h_{(j-1)\rightarrow j}, x_j), & (1)\\
y_j &= \mathcal{N}_O(h_j), & (2)
\end{aligned}
$$

where $\mathcal{N}_R$ and $\mathcal{N}_O$ are the recurrent function and output function, respectively. We first explain Eq. (1). We begin with estimating the location, *i.e.*, the bounding box, of instance k in the j-th frame from y_{j-1} by flow guided warping. FlowNet 2.0 [7] is used to extract the optical flow $F_{(j-1)\rightarrow j}$ between $(j-1)$-th and j-th frames. The binary mask y_{j-1} is warped to $y_{(j-1)\rightarrow j}$ according to $F_{(j-1)\rightarrow j}$ by a bilinear warping function. After that, we obtain the bounding box of $y_{(j-1)\rightarrow j}$ as the location of instance k in the j-th frame. We extract the feature map according to this bounding box from f_j by RoIAlign operation. The feature of this bounding box is denoted as x_j. The historical information of instance k from i-th frame to $(j-1)$-th frame is expressed by a hidden state or memory $h_{j-1} \in \mathbb{R}^{m\times m\times d}$, where $m \times m$ denotes the feature size and d represents the number of channels. We warp h_{j-1} to $h_{(j-1)\rightarrow j}$ by optical flow for spatial consistency. With both x_j and $h_{(j-1)\rightarrow j}$ we can estimate h_j by Eq. (1). Similar to the mask network described in Sec. 2.1, we employ 'conv5_x' block of ResNet-101 as our recurrent function $\mathcal{N}_R$. The mask for the instance k in the j-th frame, y_j, can then be obtained by using the output function in Eq. (2). The output function $\mathcal{N}_O$ is modeled by three convolutional layers.



(a) Vanilla Re-MP

(b) Re-MP with Attention Mechanism

Figure 4: Region attention in mask propagation.

Region attention. The quality of propagation to obtain y_j relies on how accurate the model in capturing the shape of target instance. As shown in Fig 4(a), if we directly generate y_j from h_j, a model is likely to be confused by distractors that appear in the bounding box. To overcome this issue, we leverage the attention mechanism to filter out potentially noisy regions. Specifically, given the warped hidden state $h_{(j-1)\rightarrow j}$, we first feed it into a single convolutional layer and then a softmax function, to generate the attention distribution $a_j \in \mathbb{R}^{m\times m\times 1}$ over the bounding box. Figure 4(b) shows the attention distributions we learned. Then we multiply the current hidden state h_j by a_j across all channels to focus on the regions we interested. As shown in Fig. 4, the Re-MP module concentrates on the tracked object thanks to the attention mechanism. After the forward and backward propagation, $\hat{y}$ is finally extended to a tracklet $\{y_{k1}, ..., y_{i+1}, \hat{y}, y_{i+1}, ..., y_{k2}\}$. This process is applied to all the starting points to generate a set of tracklets.

Linking the tracklets. We introduce a greedy approach to link the potentially segmented tracklets into consistent mask tubes. It sorts all tracklets descendingly by cosine similarities between their respective starting point and templates. Given the sorted order, the method then examines all tracklets in turn. A tracklet is merged with a tracklet of higher order if there is no contradiction between them. In practice, this simple mechanism works well.

Table 1: Ablation study of each module in DyeNet with DAVIS$_{17}$ *test-dev*.

	Variant	$\mathcal{J}$-mean	$\mathcal{F}$-mean	$\mathcal{G}$-mean	$\Delta\mathcal{G}$-mean
MSK[9]	ResNet-101	50.9	52.6	51.7	-
Re-MP	no attention	55.4	60.5	58.0	+ 6.2
	full	59.1	62.8	61.0	+ 9.2
+ Re-ID		65.8	70.5	68.2	+ 7.2
Advanced DyeNet	bbox input two-streams	**67.5**	**71.0**	**69.3**	+ 1.1

3. DAVIS 2018 Challenge

In the 2018 DAVIS Challenge, we apply several essential modifications to the original DyeNet to further improve the performance. Rather than full-resolution images, the advanced DyeNet accepts size-normalized patches that enclose objects of interest as input, which allows our model to better cope with objects of different scales. We employ the two-stream architecture in our advanced DyeNet to better capture the temporal information in the video sequences. In addition, we also train another DyeNet with a deeper backbone (Xception65 [4]) and ensemble two variants together to obtain the final results.

4. Experiments

We evaluate our DyeNet on DAVIS 2017 [11] (DAVIS$_{17}$) dataset, which contains 150 high-quality video sequences with all frames annotated with pixel-wise object masks. The *train set* and *val set* of DAVIS$_{17}$ are used for training. Followed [11], we adopt region ($\mathcal{J}$), boundary ($\mathcal{F}$) and their average ($\mathcal{G}$) measures for evaluation.

Effectiveness of each component in DyeNet. Table 1 summarizes how performance gets improved by adding each component step-by-step into our DyeNet on the *test-dev set* of DAVIS$_{17}$. The state-of-the-art mask propagation method, MSK [9], is chosen as the baseline. To ensure a fair comparison, we re-implement MSK to have the same backbone ResNet-101 as DyeNet. All models in this experiment are first offline trained on the *train* and *val set*, and then online trained on the *test-dev set*.

Compared with MSK, our vanilla Re-MP module significantly improves $\mathcal{G}$-mean by 6.2. The attention mechanism allows Re-MP module to focus on foreground regions, which further improves $\mathcal{G}$-mean by 3.0. The full DyeNet that contains both Re-ID and Re-MP modules achieves 68.2 by using the greedy algorithm to link the tracklets. After modified the input from full-image to bounding box and employed the two-stream architecture, our advanced DyeNet finally achieves 69.3 on the *test-dev set*.

Benchmark. As shown in Table 2, in 2018 DAVIS Challenge, ensemble DyeNet achieves a competitive $\mathcal{G}$-mean of 73.8 on *test-challenge* set.

Table 2: Results on DAVIS$_{17}$ *test-challenge*.

	$\mathcal{J}$-mean	$\mathcal{F}$-mean	$\mathcal{G}$-mean
Apata	65.1	70.6	67.8
TeamILC_RIL	67.5	71.5	69.5
Dawnsix	66.9	72.5	69.7
Jono	71.0	**78.4**	**74.7**
Ours	**71.9**	75.8	73.8

5. Conclusion

We present DyeNet, which joints re-identification and attention-based recurrent temporal propagation into a unified framework to address challenging video object segmentation with multiple instances. After applied several essential modifications, our DyeNet achieves a competitive global mean of 73.8 in the 2018 DAVIS Challenge.

References

[1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017. 1

[2] S. Caelles, A. Montes, K.-K. Maninis, Y. Chen, L. Van Gool, F. Perazzi, and J. Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv:1803.00557*, 2018. 1

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 2

[4] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017. 4

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017. 2

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[7] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 3

[8] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. In *CVPRW*, 2017. 1

[9] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, 2017. 1, 4

[10] 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 *CVPR*, 2016. 1

[11] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 4

[12] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2

[13] J. S. Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, and I. S. Kweon. Pixel-level matching for video object segmentation using convolutional neural networks. In *CVPR*, 2017. 1