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

No. 1 in DAVIS 2017 Challenge

Presentator: Jia Zheng

SIST, ShanghaiTech

May 23, 2018

Outline

Outline

Multiple objects semi-supervised video object segmentation

Task

Segment foreground multiple objects from the background region in a video sequence when given each mask of the first frame.

Challenge

1. Scale and pose variations
2. Occlusion

Recap: Two main approach to DAVIS 2016

- OSVOS [1]: segment the frames independently, no use of temporal information in the video.
- MaskTrack [6]: take temporal information into account.

OSVOS

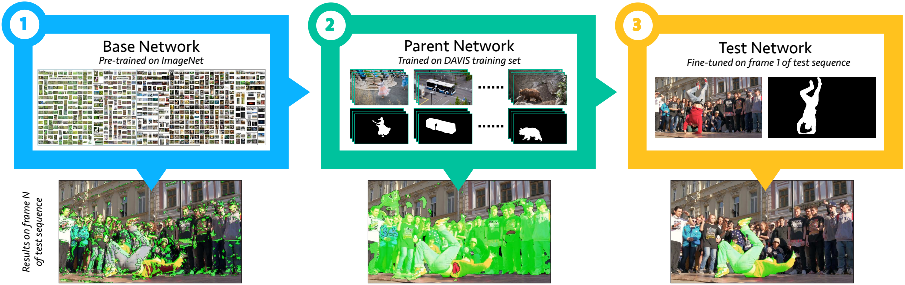


Figure: OSVOS pipeline. Figure extracted from OSVOS [1].

MaskTrack

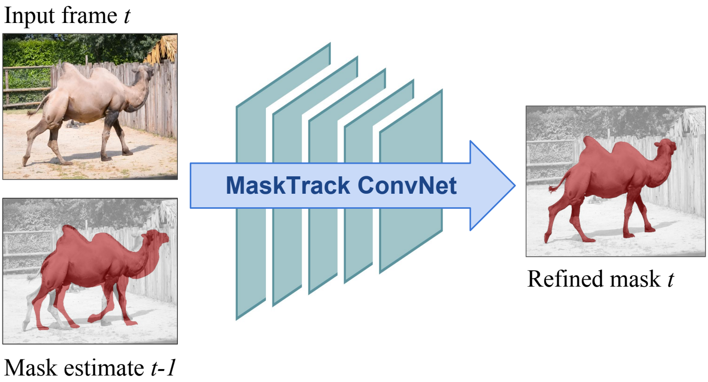


Figure: Network architecture (DeepLabv2-VGG). Expand input from RGB to RGB + mask channels. Figure extracted from MaskTrack [6].

Outline

Overview

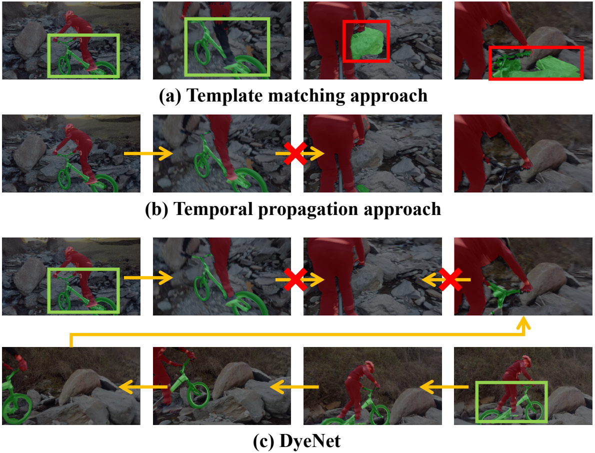
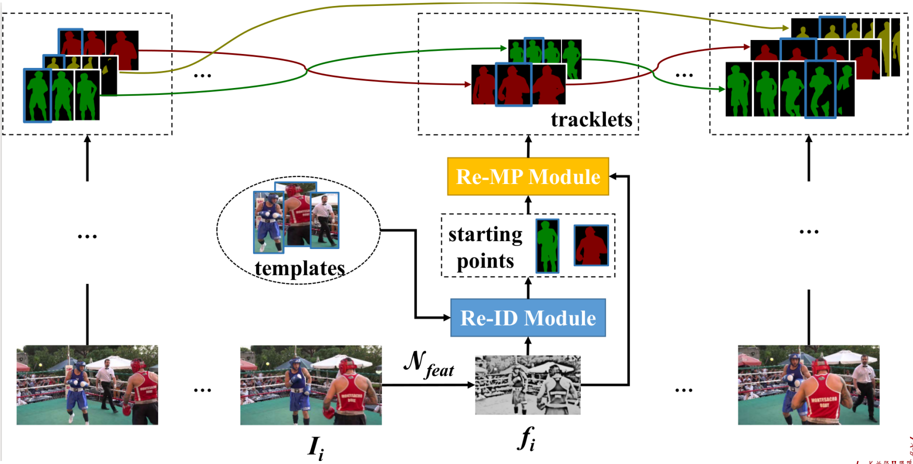


Figure: DyeNet joints template matching and temporal propagation into a unified framework.

Pipeline

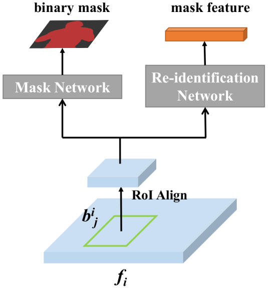


...
tracklets
Re-MP Module
starting
points
templates
Re-ID Module
N_{feat}
I_i
f_i

Inference

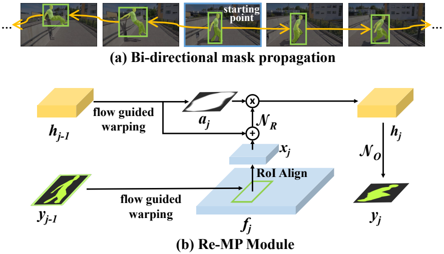
1. Extract feature by ResNet-101 [2]
2. Re-ID generates a set of masks from object proposals and compare them with templates. Masks with a high similarity to templates are chosen as a starting points for Re-MP module.
3. Re-MP propagates each selected mask bidirectionally, and generates a sequence of segmentation masks (tracklets).
4. Post-processing to link tracklets.

Re-ID Module



1. Use RPN[7] to propose candidate object bounding boxes on every frame.
2. Extract its feature, resize by RoIAlign[3], feed into two sub-networks to get binary mask and mask feature.
3. Use cosine distance to measure similarities between candidate bounding box and templates.
4. If a candidate is sufficiently similar to any template, keep its mask as a starting point for Re-MP.

Re-MP Module

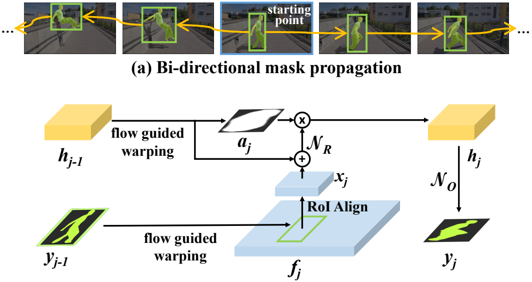


$$h_j = \mathcal{N}_R(h_{(j-1)\rightarrow j}, x_j) \qquad (1)$$

$$y_j = \mathcal{N}_O(h_j) \qquad (2)$$

1. warp previous mask y_{j-1} and hidden state h_{j-1} by optical flow
2. obtain bounding box according to the warped mask, extract its feature x_j by RoIAlign
3. propagate mask by RNN by Equ. 1 and Equ. 2

Attention Mechanism

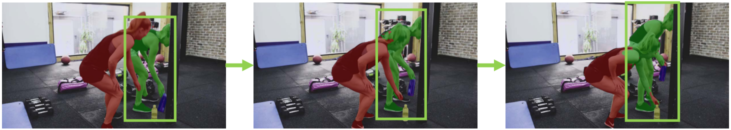


(a) Bi-directional mask propagation

(b) Re-MP Module

Feed warped hidden state $h_{(j-1)\rightarrow j}$ into a single convolutional layer, and then a softmax function.

Region Attention



(a) Vanilla Re-MP

(b) Re-MP with Attention Mechanism

Linking the tracklets

Greedy approach

1. Sort all tracklets descendingly by cosine similarities between their respective starting point and templates. Extend the starting points according to the sorted order.
2. Skip the starting point whose mask highly overlaps with a mask in existing tracklets.
3. Tracklet with the highest similarities are assigned to the respective templates.
4. A tracklet is merged with a tracklet of higher order if there is no contradiction between them.

Outline

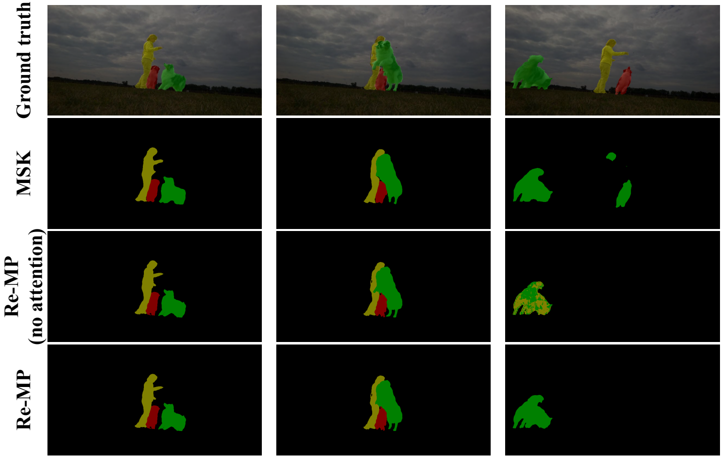
1. Introduction
2. Approach
3. Experiments
    - Ablation Study
    - Benchmark

Re-MP module

Table: Ablation study on Re-MP with DAVIS$_{17}$ *val*.

	Variant	$\mathcal{J}$-mean	$\mathcal{F}$-mean	$\mathcal{G}$-mean
MaskTrack [6]	ResNet-101	63.3	67.2	65.3
Re-MP	no attention	65.3	69.7	67.5
	full	**67.3**	**71.0**	**69.1**

Re-MP module



Re-ID Module

Table: Ablation study on Re-ID with DAVIS$_{17}$ *val*. The improvement of $\mathcal{G}$-mean between rows is because of template expansion.

ρ_{reid}	0.9	0.8	0.7	0.6
	$\mathcal{G}$-mean	$\mathcal{G}$-mean	$\mathcal{G}$-mean	$\mathcal{G}$-mean
Iter. 1	72.3	73.2	73.2	73.4
Iter. 2	73.3	73.7	**74.1**	74.0
Iter. 3^+	73.6	73.7	**74.1**	73.9

Each component in DyeNet

Table: 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
MaskTrack [6]	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
offline	offline only	60.2	64.8	62.5	- 5.6

DAVIS 2017 Benchmark

Table: Results on DAVIS_{17} *test-dev*

	online training		$\mathcal{J}$-mean	$\mathcal{F}$-mean	$\mathcal{G}$-mean
	dataset	video			
OnAVOS [8]†	$\surd$	$\surd$	53.4	59.6	56.5
LucidTracker [4]	$\surd$	$\surd$	60.1	68.3	64.2
VS-ReID [5]	$\surd$	$\times$	64.4	67.8	66.1
LucidTracker [4]†	$\surd$	$\surd$	63.4	69.9	66.6
DyeNet (offline)	$\times$	$\times$	60.2	64.8	62.5
DyeNet	$\surd$	$\times$	**65.8**	**70.5**	**68.2**

Approaches with ensemble are marked with †.

Visualization

Figure: Visualization of DyeNet's prediction.

Reference

- Sergi Caelles et al. "One-shot video object segmentation". In: *CVPR*. 2017.
- Kaiming He et al. "Deep residual learning for image recognition". In: *CVPR*. 2016.
- Kaiming He et al. "Mask r-cnn". In: *ICCV*. 2017.
- A. Khoreva et al. "Lucid Data Dreaming for Multiple Object Tracking". In: *arXiv preprint arXiv: 1703.09554*. 2017.
- X. Li et al. "Video Object Segmentation with Re-identification". In: *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops* (2017).
- Federico Perazzi et al. "Learning video object segmentation from static images". In: *CVPR*. 2017.
- Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *NIPS*. 2015.
- Paul Voigtlaender and Bastian Leibe. "Online Adaptation of Convolutional Neural Networks for Video Object Segmentation". In: *BMVC*. 2017.

Thanks

Thanks for Attention!