Learning Object Deformation and Motion Adaption for Semi-supervised Video Object Segmentation

Xiaoyang Zheng, Xin Tan, Jianming Guo, Lizhuang Ma
Digital Media & Computer Vision Lab, Shanghai Jiao Tong University

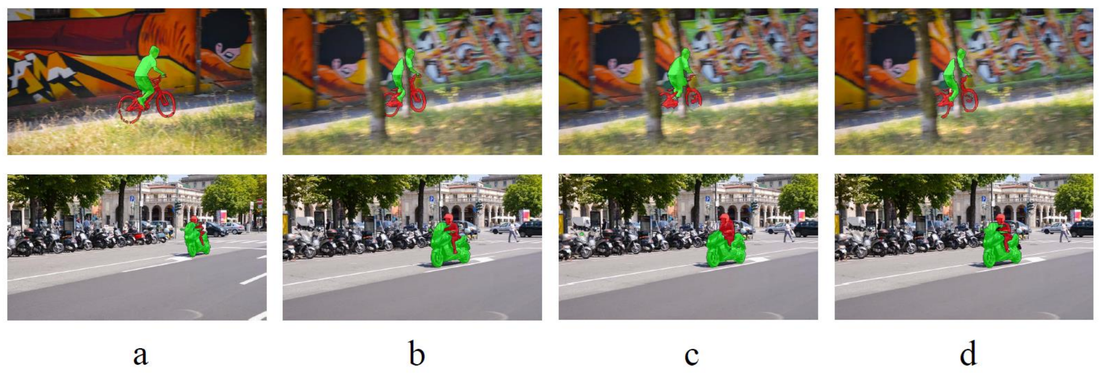
SHANGHAI JIAO TONG UNIVERSITY

Contributions

- Propose to employ the mask-propagation-based model to learn the past and current information without any online fine-tuning
- Adapts to the shape variance of target objects well with generated image/mask pairs from training videos
- Adapts to object motions by training on inserted frames generated based on two adjacent frames

Problem

- **Object deformation and motions**



a b c d

  - Difficult to adapt to the shape variance of target object in long sequence
  - Difficult to describe the diversity from annotated frames
- **Scarcity of training data and annotations**
  - Expensive and time-consuming to annotate densely on real video sequences
  - Unsatisfying performance on the semi-supervised task

Network Structure

- **Backbone**
  - Resnet50 as the backbone feature extractor
  - An additional channel for the pixel-level mask besides RGB channels
  - Obtains the knowledge from past frames and maintains a temporal coherence explicitly
- **Fusion Module**
  - Taking the feature streams of the initial frame and the current frame as inputs and build a connection between these two frames
  - Learn the target appearance information
  - Adopt a GCN to enlarge the effective receptive field and support global feature matching
- **Upsampling Module**
  - Produce a soft segmentation $\widehat{y_p}$
  - Use the coarsely predicted mask for segmentation of the following frame, in order to localize target object

Inference

- Treat video sequences with multiple objects as several single-object segmentation problems
- Other targets are viewed as the background.

$$p_{i,m} = \mathrm{softmax}(\mathrm{logit}(\hat{p}_{i,m})) = \frac{\hat{p}_{i,m}/(1-\hat{p}_{i,m})}{\sum_{j=0}^{M}\hat{p}_{i,j}/(1-\hat{p}_{i,j})}$$

Synthetic video clip generation

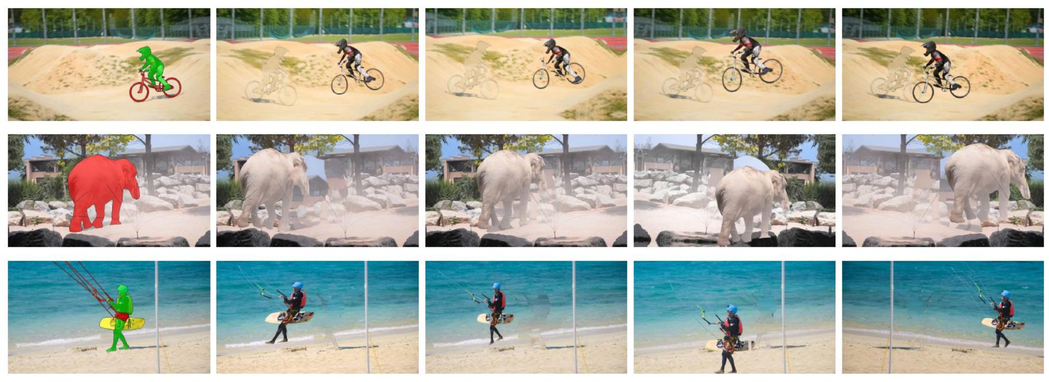


Image + Mask (Initial Frame) | Frame t | Frame t_1 | Frame t_2 | Frame t+1

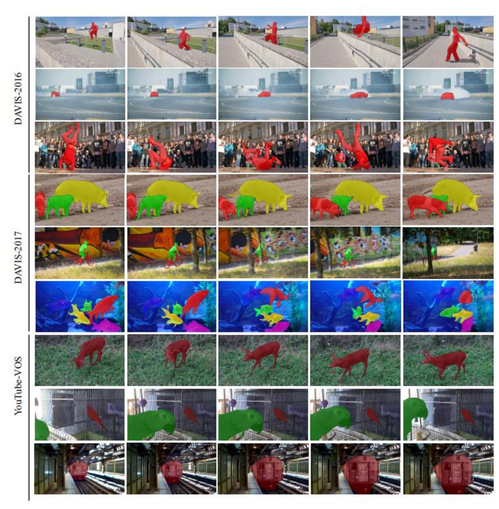
- **Object Deformation Simulation**
  - Conduct translation and deformation on the foreground objects
- **Motion Simulation**
  - Introduce smooth intermediate transformation among two key frames and model the natural development between consecutive frames

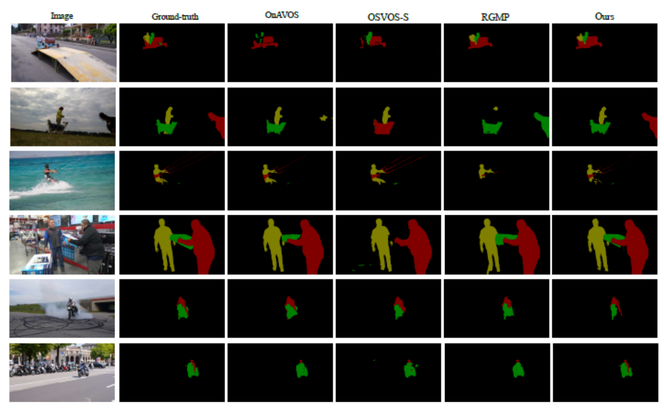
Experiments

- **Quantitative Evaluation**

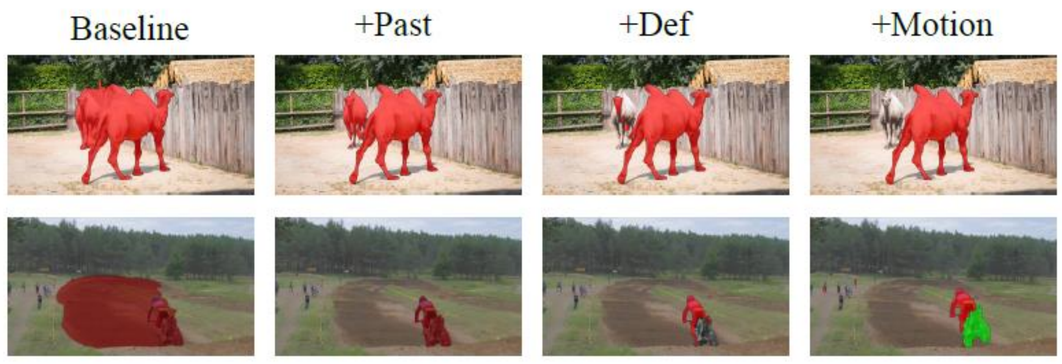
Method	Online	DAVIS-2016 [8]		DAVIS-2017 [24]		YouTube VOS [37]	
		Mean $\mathcal{J}\uparrow$	Mean $\mathcal{F}\uparrow$	Mean $\mathcal{J}\uparrow$	Mean $\mathcal{F}\uparrow$	Mean $\mathcal{J}\uparrow$	Mean $\mathcal{F}\uparrow$
CINM [17]	✓	83.4	85.0	**67.2**	**74.0**	-	-
OSVOS-S [20]	✓	85.6	**87.5**	64.7	71.3	-	-
OnAVOS [18]	✓	**86.1**	84.9	61.6	69.1	**59.3**	54.2
MSK [34]	✓	79.7	75.4	63.3	67.2	-	-
SFL [16]	✓	76.1	76.0	-	-	-	-
OSVOS [34]	✓	79.8	80.6	56.6	63.9	57.0	**56.8**
PML [38]	×	75.5	79.3	-	-	-	-
CTN [10]	×	73.5	69.3	-	-	-	-
VPN [27]	×	70.2	65.5	-	-	-	-
RGMP [39]	×	81.5	**82.0**	64.8	68.8	52.4	56.0
FAVOS [16]	×	**82.4**	79.5	54.6	61.8	-	-
OSMN [40]	×	74.0	72.9	52.5	57.1	52.4	50.8
Ours	×	82.0	79.7	**67.5**	**73.5**	**62.9**	**67.0**

- **Qualitative Evaluation**





- **Ablation Study**



Metric	Baseline	+*Past*	+*Past*+*Def*	+*Past*+*Def*+*Motion*
Mean $\mathcal{J}\uparrow$	55.6	60.5	66.1	67.5
Mean $\mathcal{F}\uparrow$	67.2	69.1	70.8	73.5

Framework