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

Xiaoyang Zheng*[1] , Xin Tan[1] , **Jianming Guo**[1] , Lizhuang Ma[1]

[1] Shanghai Jiao Tong University

- Problem
- Method
  - Network Structure
  - Synthetic Video Clip Generation
  - Inference
- Experiments
  - Experiments Settings
  - Quantitative Evaluation
  - Qualitative Evaluation
  - Ablation Study
- Conclusion

- Problem
- Method
  - Network Structure
  - Synthetic Video Clip Generation
  - Inference
- Experiments
  - Experiments Settings
  - Quantitative Evaluation
  - Qualitative Evaluation
  - Ablation Study
- Conclusion

Problem overview

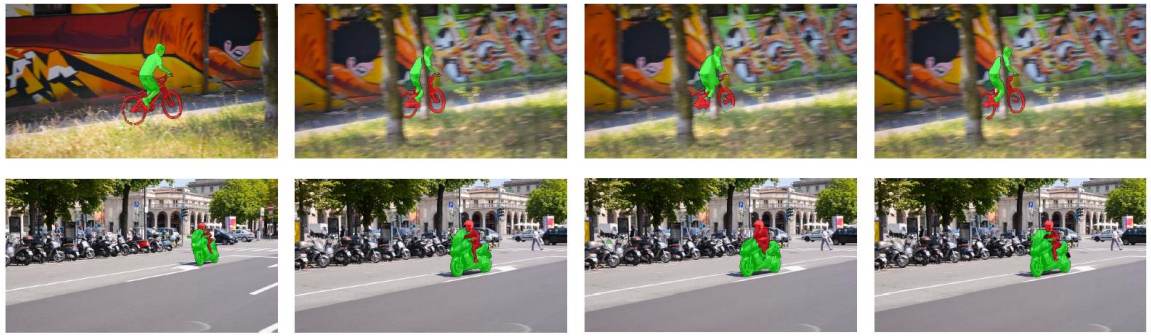

Fig. 1. The shape variance and the motion difference across different frames.

Targets:

- Track and segment one or multiple objects in a video sequence
- the mask annotation is only given at **the first frame** of the video sequence

Challenges:

- **object deformation and motions**
- Difficulty to describe the **diversity**
- Difficulty to adapt to the **shape variance** of target object
- Scarcity of training data and **annotations**

- Problem
- Method
  - Network Structure
  - Synthetic Video Clip Generation
  - Inference
- Experiments
  - Experiments Settings
  - Quantitative Evaluation
  - Qualitative Evaluation
  - Ablation Study
- Conclusion

Method

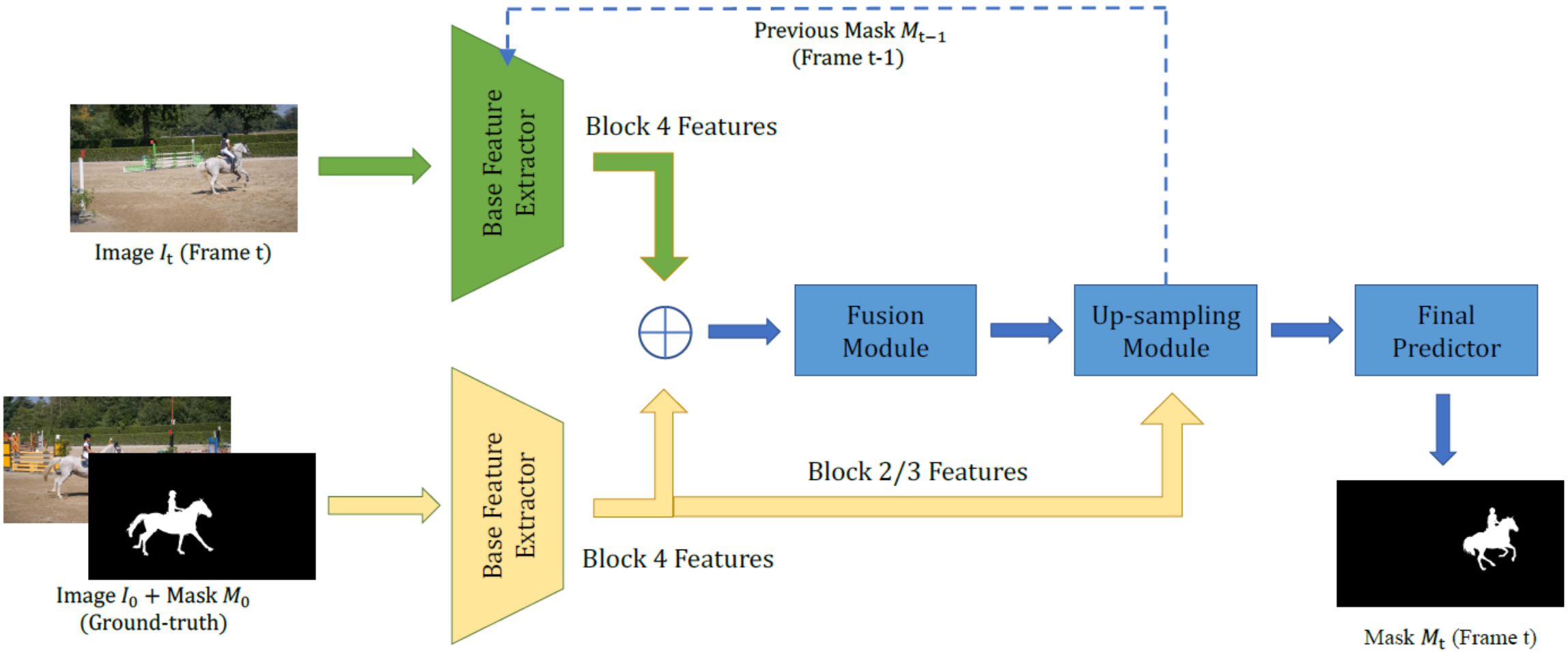

Fig. 2. The structure of our proposed network.

Method-*network structure*

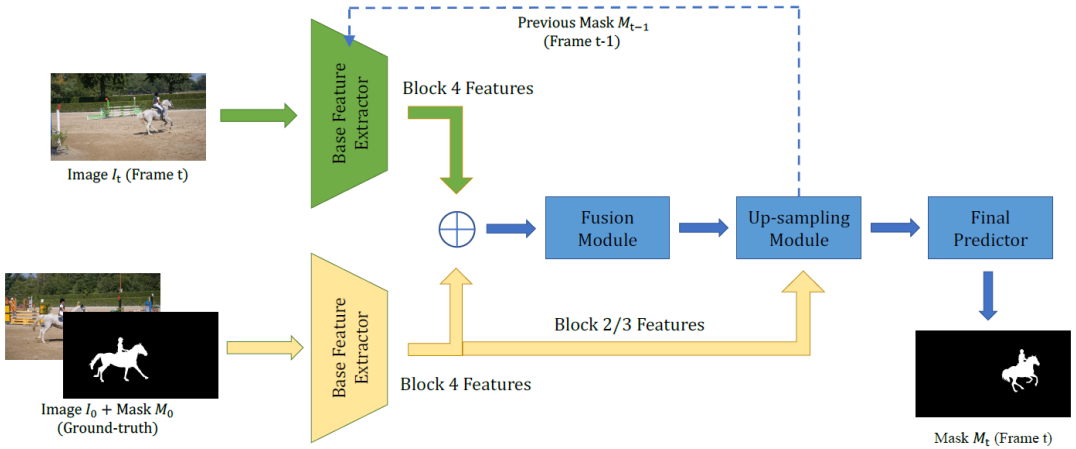

Fig. 2. The structure of our proposed network.

- **Backbone**
  - **ResNet50** as the backbone feature extractor
  - an additional channel for the **pixel-level mask**
  - Obtain the knowledge from past frames
  - maintains a **temporal coherence** explicitly
- **Fusion Module**
  - Input: the feature streams of the initial frame and the current frame
  - Learn the target **appearance**
  - enlarge the effective receptive field and support global feature matching
- **Upsampling Module**
  - Produce a **soft segmentation** $\widehat{y_p}$
  - localize target object
  - Mask propagation

Method-*synthetic video clip generation*

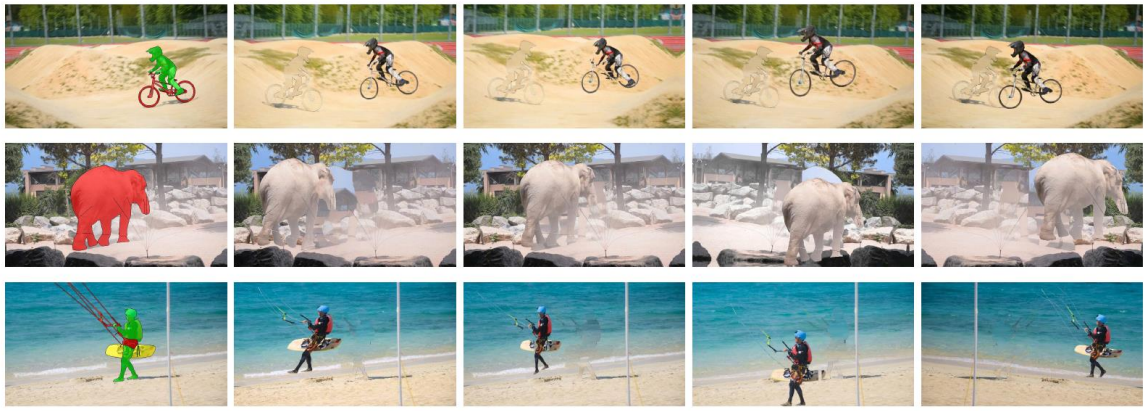

Fig. 3. Synthetic video snippets generated from DAVIS-2017 training set

- **Object Deformation Simulation**
  - adapt to **object deformation**
  - (M_t, L_t) to (M_{t+1}, L_{t+1})
  - Simulate **shape variance**
- **Motion Simulation**
  - **smooth** intermediate transformation
  - Natural development

Method-*Inference*

- Multiple object → several single-object segmentation problems
- Masks $\widehat{y_p}$ s→ aggregated mask
- Combine the output probability $\widehat{y_p}$ of the previous frame and current frame
- The way of **aggregation**

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

- Offline methods
  - End-to-end
  - Without extra appearance and motion cues

- Problem
- Method
  - Network Structure
  - Synthetic Video Clip Generation
  - Inference
- Experiments
  - Experiments Settings
  - Quantitative Evaluation
  - Qualitative Evaluation
  - Ablation Study
- Conclusion

Experiments-*Experiments Settings*

- **Datasets**
  - DAVIS-2016
  - DAVIS-2017
  - YouTube-VOS
- **Evaluation Metrics**
  - Region similarity: the region-based segmentation similarity
  - Contour accuracy: F-measure between the contour points of the predicted mask and the ground-truth segmentation
- **Baselines**
  - Online methods: CINM, OSVOS-S, OnAVOS, MSK, SFL, OSVOS
  - Offline methods: PML, CTN, VPN, RGMP, FAVOS, OSMN

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**

Tab. 1.COMPARISON WITH STATE-OF-THE-ART METHODS ON THE VALIDATION SET OF DAVIS-2016 , DAVIS-2017 AND YOUTUBE-VOS

Experiments-*Qualitative Evaluation*

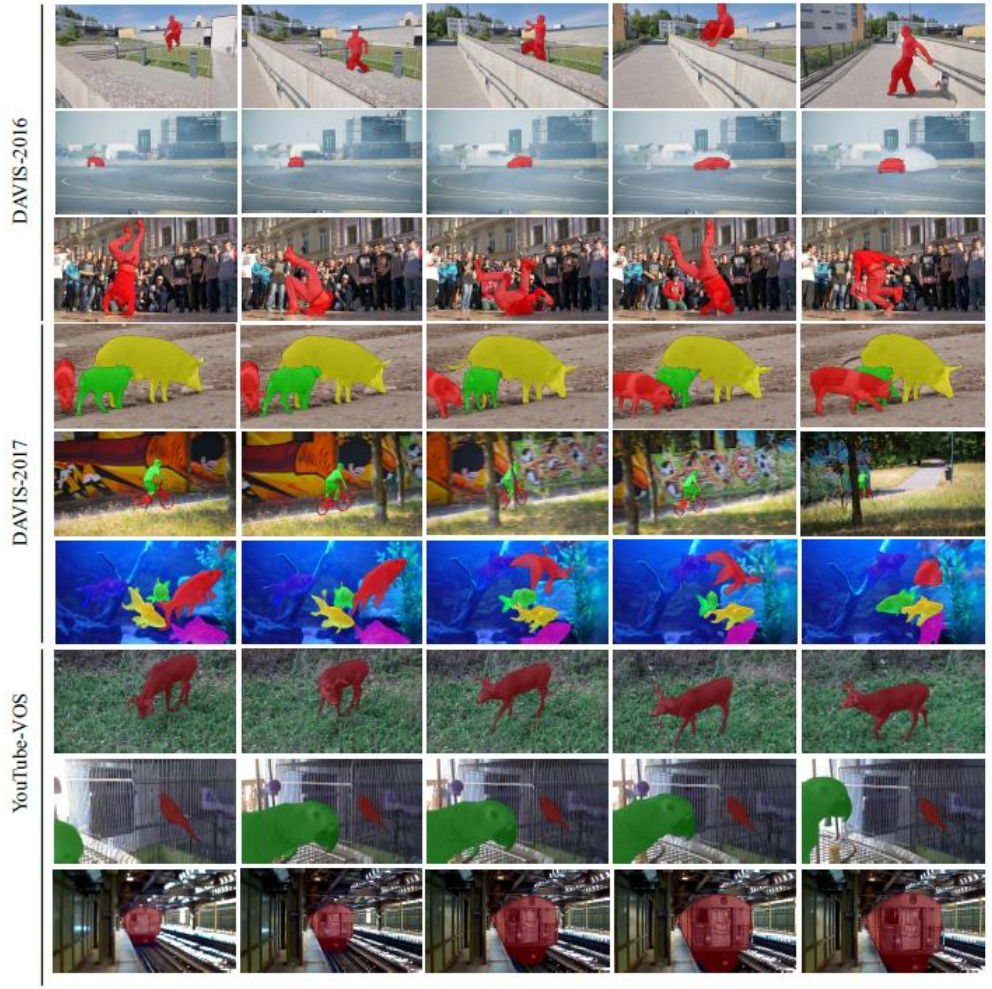

Fig. 4. Qualitative result of our proposed method.

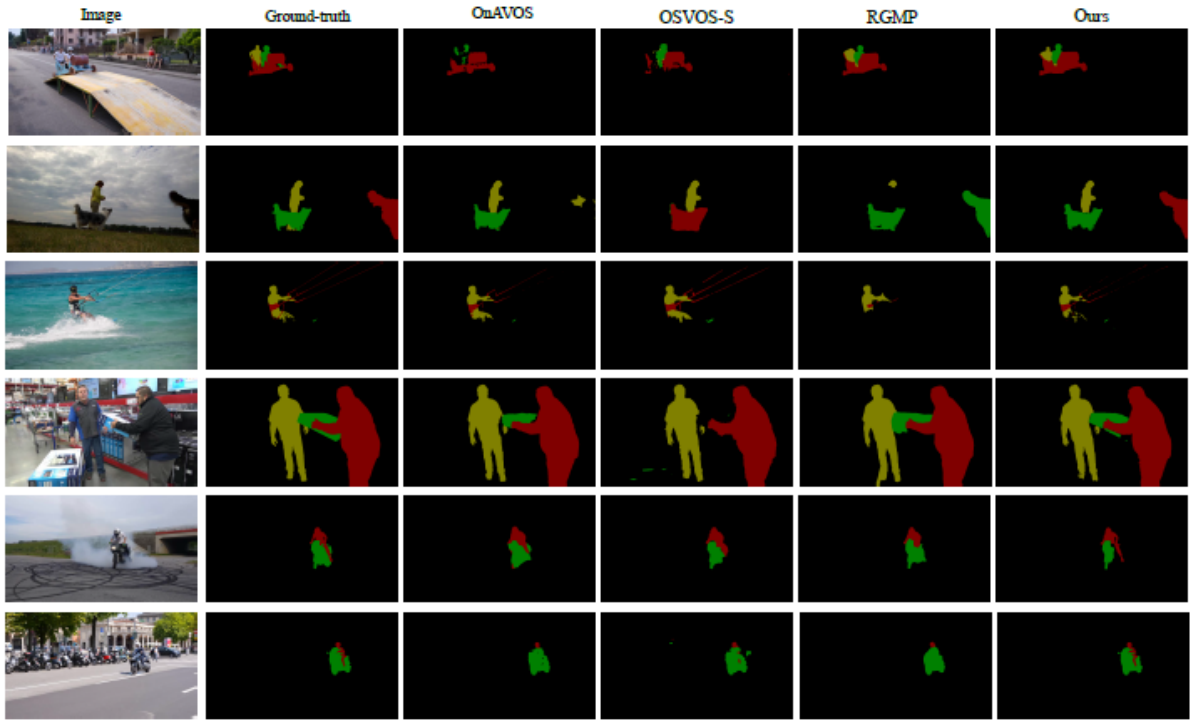

Fig. 5. Comparison with state-of-the-art methods.

Experiments-*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

Tab. 2.RESULT OF THE ABLATION STUDY ON DAVIS-2017 VALIDATION SET

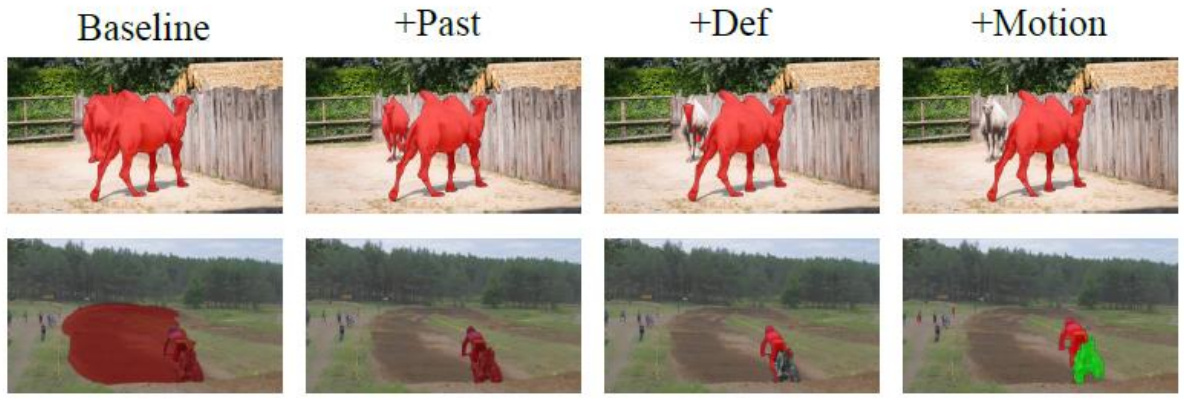

Fig. 6. Ablation study on synthetic video clip generation

- Problem
- Method
  - Network Structure
  - Synthetic Video Clip Generation
  - Inference
- Experiments
  - Experiments Settings
  - Quantitative Evaluation
  - Qualitative Evaluation
  - Ablation Study
- Conclusion

Conclusion

Contributions:

- Propose the mask-propagation-based model
- Adapt to the shape variance of target
- Adapts to object motions
- Avoid the use of extensive online fine-tuning

Future work:

- Expand the video object segmentation method into interactive scenarios

Thanks