Video Object Segmentation with Re-identification

Xiaoxiao Li, Yuankai Qi, Zhe Wang, Kai Chen, Ziwei Liu, Jianping Shi
Ping Luo, Chen Change Loy, Xiaoou Tang

The Chinese University of Hong Kong, SenseTime Group Limited
Semi-supervised Segmentation

• Input: Video sequence, ground-truth label of the first frame

• Output: Masks of all instances
Challenge

- Instance Segmentation
  - Small objects and fine structures
  - Scale & pose-variations

- Tracking
  - Frequent occlusions
Challenge

• Instance Segmentation
  • Small objects and fine structures
  • Scale & pose-variations

• Tracking
  • Frequent occlusions

Mask Propagation Module

Re-identification Module

Short Term

Long Term
Proposed Framework

Video Object Segmentation with Re-identification (VS-ReID)
Video Object Segmentation with Re-identification (VS-ReID)

Mask Propagation Module

- Input Video Sequence
- Mask Propagation Module
- Mask Initialization
- Re-identification Module
- Iterative Inference
- Mask Propagation Module

Video Object Segmentation with Re-identification (VS-ReID)
Mask Propagation Module

- Inspired by MSK[1] and LucidTracker[2]

- Use the **temporal continuity** property of the video sequence

- Propagate the mask from **the previous frame** to **the current frame**

Mask Propagation Module

Image

Guided Probability Map

Optical Flow

RGB Branch

Flow Branch

Prediction
Mask Propagation Module

Image

Guided Probability Map

Optical Flow

RGB Branch

Flow Branch

Prediction
Mask Propagation Module

Previous Frame

Current Frame

Previous Mask
Mask Propagation Module

- Previous Frame
- Current Frame
- FlowNet
- Optical Flow
- Warping
- Previous Mask
- Guided Probability Map
Mask Propagation Module

Previous Frame \(\xrightarrow{\text{FlowNet}}\) Optical Flow \(\xrightarrow{\text{Warping}}\) Guided Probability Map

Current Frame
Mask Propagation Module

Image → Guided Probability Map → Optical Flow → RGB Branch → Flow Branch → Prediction
Mask Propagation Module

Image

Guided Probability Map

Optical Flow

RGB Branch

Flow Branch

Prediction
Mask Propagation Module

- Deeper Backbone Network
  - ResNet101

- RGB-branch
  - Pre-trained on the MS-COCO and PASCAL VOC dataset
    - Augmented ground-truth label as the guided probability map
    - Fine-tuned on the DAVIS dataset

- Flow-branch
  - Initialized with RGB-Branch’s weights
  - Trained on the DAVIS dataset

- Multi-instance
  - Inference on each instance individually
Mask Propagation Module
Proposed Framework

Video Object Segmentation with Re-identification (VS-ReID)
Re-identification Module

• Detection and re-identification

First Frame

Rest Frames

Template

Candidate Bounding Boxes

Most Confident Candidate
Re-identification Module

- Recover the mask from a bounding box
Re-identification Module

• Detection Model
  • Faster RCNN
  • Trained on the ImageNet

• Re-identification Model
  • ‘Identification Net’ in Person Search[1]
  • For the person category, we directly use the ‘Identification Net’ in Person Search[1]
  • Trained on the ImageNet VID

• Retrieve an instance in a single frame each time

Mask Propagation Module

Video Object Segmentation with Re-identification (VS-ReID)
VS-ReID

- Mask Initialization

Mask Initialization

1  8  20  37  52  64  82

Input Frames

Mask Propagation
Input Frames

1 8 20 37 52 64 82

Mask
Propagation

Re-Identification

1st Round

21
Input Frames

Re-Identification

1st Round

Frames: 1, 8, 20, 37, 52, 64, 82
Re-Identification

Mask Propagation

Input Frames

1st Round

\( \hat{x} \)
## Performance

<table>
<thead>
<tr>
<th></th>
<th>J Mean</th>
<th>F Mean</th>
<th>Global Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voigt</td>
<td>54.8</td>
<td>60.5</td>
<td>57.7</td>
</tr>
<tr>
<td>Haamo</td>
<td>59.8</td>
<td>63.2</td>
<td>61.5</td>
</tr>
<tr>
<td>Vanta</td>
<td>61.5</td>
<td>66.2</td>
<td>63.8</td>
</tr>
<tr>
<td>Apata</td>
<td>65.1</td>
<td>70.6</td>
<td>67.8</td>
</tr>
<tr>
<td>Ours</td>
<td>67.9</td>
<td>71.9</td>
<td>69.9</td>
</tr>
</tbody>
</table>

(DAVIS 2017 Challenge test-challenge set)
Visualization
Thanks!