Proposal Tracking and Segmentation (PTS): A cascaded network for video object segmentation

Zilong Huang*, Qiang Zhou*, Lichao Huang, Han Shen, Yongchao Gong, Chang Huang, Wenyu Liu, Xinggang Wang

speeding_zZ team

Huazhong University of Science and Technology (HUST) & Horizon Robotics

*equal contribution & interns of Horizon Robotics
PTS: A cascaded network for video object segmentation

- **RPN**: Region Proposal Network (2000 boxes)
- **OTN**: Object Tracking Network (1 box)
- **RGSN**: Reference-Guided Segmentation Network
RPN: Region Proposal Network

The Region Proposal Network is pre-trained on COCO and provides class-agnostic object candidate boxes. RPN could encode the instance(object) information into framework.

OTN: Object Tracking Network

Inspired by MDNet, Object Tracking Network is designed to score the candidate boxes and updated online for adapting to large and fast changes in object appearance.

Online Object Tracking Network

- **Long-term updates** are performed in regular intervals using the positive samples collected for a long period.
- **Short-term updates** are conducted whenever potential tracking failures are detected—when the score of the estimated target is less than 0.5 — using all the positive samples in the short-term period.

To estimate the target state in each frame, \( N=256 \) target candidates \( x^1, \ldots, x^N \) sampled from candidate bounding boxes which are around the previous target state are evaluated using the network, and we obtain their scores \( f(x^i) \). The optimal target state \( x^* \) is given by finding the example with the maximum score as

\[
x^* = \arg \max_{x^i} f(x^i)
\]
Then, the box with the highest score evaluated by OTN is selected to crop and resize the frame for normalizing the scale variation of objects.

Reference-Guided Segmentation Network will make use of both cropped region with previous mask and the reference frame to segment target object.

Offline Training

RPN adapts Resnet-152 as backbone and is trained on COCO

RGSN adapts Resnet-50 as backbone and is trained on YouTube-VOS training dataset

AUG:
1. Random select two frames as a current frame and a reference frame.
2. Sample bounding boxes around the ground truth box and random scale from 1.5~2.0
3. Encode the previous mask as a heatmap with a two-dimensional Gaussian distribution
Online Training

Video sequences

RPN

OTN

RGSN

Update model during inference

Fine-tune with first annotated frame before inference for only one time

AUG:
1. Sample bounding boxes around the ground truth box and random scale from 1.5~2.0
2. Encode the previous mask as a heatmap with a two-dimensional Gaussian distribution
The influence of Reference-Guided Segmentation Network

<table>
<thead>
<tr>
<th>Method</th>
<th>J seen</th>
<th>J unseen</th>
<th>F seen</th>
<th>F unseen</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>P + T+ naïve segmentation</td>
<td>61.3</td>
<td>50.5</td>
<td>61.9</td>
<td>55.3</td>
<td>57.1</td>
</tr>
<tr>
<td>P + T+ RGSN</td>
<td>66.3</td>
<td>51.2</td>
<td>69.2</td>
<td>57.2</td>
<td><strong>61.0</strong></td>
</tr>
</tbody>
</table>

Reference-Guided Segmentation Network outperforms naïve segmentation Network
The influence of tracked box expansion

<table>
<thead>
<tr>
<th>Method</th>
<th>J seen</th>
<th>J unseen</th>
<th>F seen</th>
<th>F unseen</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTS+1.0x tracked box</td>
<td>66.3</td>
<td>51.2</td>
<td>69.2</td>
<td>57.2</td>
<td>61.0</td>
</tr>
<tr>
<td>PTS+1.4x tracked box</td>
<td>67.9</td>
<td>52.7</td>
<td>70.6</td>
<td>58.6</td>
<td>62.4</td>
</tr>
<tr>
<td>PTS+1.5x tracked box</td>
<td>68.4</td>
<td>52.5</td>
<td>70.9</td>
<td>58.3</td>
<td><strong>62.5</strong></td>
</tr>
<tr>
<td>PTS+1.6x tracked box</td>
<td>68.5</td>
<td>52.3</td>
<td>70.9</td>
<td>57.8</td>
<td>62.4</td>
</tr>
<tr>
<td>PTS+1.7x tracked box</td>
<td>68.5</td>
<td>52.1</td>
<td>70.9</td>
<td>57.2</td>
<td>62.2</td>
</tr>
</tbody>
</table>

The proper box expansion can improve the result consistently
**Summary**

*Baseline: RPN + OTN + naïve segmentation network*
Visualization
Speed

- 30 hours for offline-training (RGSG)
- 0.9 second per frame for online-learning and inference
- Hardware: a single Titan X Pascal GPU
- Implemented using PyTorch
Conclusions

1. PTS is a unified, simple yet effective framework for video object segmentation.

2. The proposal network helps to bring objectness info for VOS by supervised pre-training.

3. PTS utilizes the SOTA video object tracking and video segmentation methods.
Future directions

1. Integrate long-term temporal features of OTN into RGSN

2. Joint training of three networks

3. Speedup
Thanks & Questions