PMSNet: Propagated Masks Selection Network for Video Object Segmentation

Huaijia Lin\textsuperscript{1*}, Ruizheng Wu\textsuperscript{1,2*}, Xiaogang Xu\textsuperscript{1}, Xiaojuan Qi\textsuperscript{1}, Jiaya Jia\textsuperscript{1,2}

\textsuperscript{1}The Chinese University of Hong Kong \hspace{1cm} \textsuperscript{2}Tencent YouTu Lab

*indicates equal contribution
Problem Definition

• Separating an object from the background in a video, given the mask of the first frame.
Challenges

• Missing Objects Reappear

• Substantial Appearance Variations

• Multiple Similar Objects Occluding
Framework – Inference Pipeline

\[ L_{x-1} \]

\[ I_{x-1} \] → Propagation Module → Propagated Mask

\[ I_x \] → Mask RCNN → Proposed Masks

Reid Selection → \( L_x \)
Framework – Inference Pipeline

\[ L_{x-1} \]

\[ I_{x-1} \]

Propagation Module

Propagated Mask

Reid Selection

\[ L_x \]

\[ I_x \]

Mask RCNN

Proposed Masks
Framework – Propagation Module

- Overview of Propagation Module

\[ \text{Loss} = \frac{\sum_i (GT_i \ast P_i)}{\sum_i (GT_i + P_i - GT_i \ast P_i)} \]
Framework – Propagation Module

- Overview of Propagation Module

\[
Loss = \frac{\sum_i (GT_i * P_i)}{\sum_i (GT_i + P_i - GT_i * P_i)}
\]
Framework – Propagation Module

- **Motion Feature extract**
  - Adopt FlowNet2C\textsuperscript{[1]} structure.
  - Load Flownet2C pre-trained weight.
  - The magnitude of optical flow $\|F_{x\rightarrow x-1}\|^2_2$ will be the motion features for subsequent network.
  - Learn motion features end-to-end.

Framework – Propagation Module

• Overview of Propagation Module

\[
Loss = \frac{\sum_i (GT_i * P_i)}{\sum_i (GT_i + P_i - GT_i * P_i)}
\]

\[
\begin{align*}
I_x & \quad \text{Appearance Branch} \\
L_{x-1} & \quad \text{Deep motion Branch} \\
F_{x\rightarrow x-1} & \quad \text{Shallow motion Branch} \\
\end{align*}
\]
Framework – Propagation Module

- Overview of Propagation Module

\[ \text{Loss} = \frac{\sum_i (GT_i \ast P_i)}{\sum_i (GT_i + P_i - GT_i \ast P_i)} \]
Framework – Propagation Module

- **Motion branches**
  - **Input:**
    - Last frame label $L_{x-1}$.
    - Motion features from current frame to last frame $F_{x \rightarrow x-1}$.
  - **Deep Motion Branch**
    - Adopt OSVOS\(^1\) network structure.
    - Load with VGG16 pre-trained weight.
  - **Shallow Motion Branch**
    - Several Convolution-Relu blocks.
    - No down-sample operation.
    - It improves **10.56** in validation set overall score.

One-Shot Video Object Segmentation, Computer Vision and Pattern Recognition (CVPR), 2017.
Framework – Propagation Module

- Overview of Propagation Module

\[
Loss = \frac{\sum_i (GT_i \cdot P_i)}{\sum_i (GT_i + P_i - GT_i \cdot P_i)}
\]
Framework – Propagation Module

• Overview of Propagation Module

\[ \text{Loss} = \frac{\sum_i (GT_i \times P_i)}{\sum_i (GT_i + P_i - GT_i \times P_i)} \]
Framework – Propagation Module

• **Appearance Branch**
  • **Input:**
    • Current frame RGB image $I_x$
    • Last frame label $L_{x-1}$
  • **Network setting**
    • Adopt OSVOS network structure.
    • Load with VGG16 pre-trained weight.
Framework – Propagation Module

• Overview of Propagation Module

\[ \text{Loss} = \frac{\sum_i (GT_i \cdot P_i)}{\sum_i (GT_i + P_i - GT_i \cdot P_i)} \]
Framework – Propagation Module

- Overview of Propagation Module

\[
Loss = \frac{\sum_i (GT_i \times P_i)}{\sum_i (GT_i + P_i - GT_i \times P_i)}
\]
Framework – Propagation Module

- **Fusion & Loss function**
  - We fuse three branch prediction.
  - IoU loss is set as our loss function.

\[
Loss = \frac{\sum_i (GT_i \times P_i)}{\sum_i (GT_i + P_i - GT_i \times P_i)}
\]

- The propagated mask is utilized for subsequent selection.
Framework – Propagation Module

• Inference strategy
  • Multi-frame ensemble

• We ensemble the prediction from previous 5 frames.
• Only the results of $I_{t-5}$ and $I_t$ will be saved for validation, all frame results will be saved for testing.
• It improves 2.52 in validation set overall score.
We directly use the pre-trained model\cite{1} of coco dataset.

Framework – Inference Pipeline

\[ L_{x-1} \]

\[ I_{x-1} \]

\[ I_x \]

Propagation Module

Propagated Mask

Reid Selection

Proposed Masks

\[ L_x \]
Framework – Reid Selection

Proposed Masks

Motion Consistency With \( L_{x-1} \)

Appearance Consistency With \( L_{x-1} \)

Best Proposed Masks

Best Proposed Masks

Propagated Mask

Appearance Consistency With \( L_0 \)

\( L_x \)

\( L_0 \) denotes the given mask.
Framework – Reid Selection

Proposed Masks

Motion Consistency With $L_{x-1}$

Appearance Consistency With $L_{x-1}$

Best Proposed Masks

Best Proposed Masks

Appearance Consistency With $L_0$

$L_x$

$L_0$ denotes the given mask.
Reid Selection – Motion Consistency

• **Motion Consistency**

\[
S_{MC}(\text{Mask}, L_{x-1}) = \frac{1}{2} (\text{IoU}(F_{x \rightarrow x-1}(\text{Mask}), L_{x-1}) + \text{IoU}(F_{x-1 \rightarrow x}(L_{x-1})), \text{Mask}))
\]

- \(F_{x \rightarrow x-1}\) and \(F_{x-1 \rightarrow x}\) is the warp operation with the optical flow
- For saving computation, the masks with \(S_{MC}(\text{Mask}, L_{x-1})\) smaller than a threshold (=0.2 in the experiment) are abandoned.
Framework – Reid Selection

Proposed Masks

Motion Consistency With $L_{x-1}$

Appearance Consistency With $L_{x-1}$

Appearance Consistency With $L_0$

Best Proposed Masks

$L_x$

$L_0$ denotes the given mask.
Reid Selection – Appearance Consistency

- **Appearance Consistency**
  - Realize it by Pixel Embedding Network, inspired by [1]
  - For saving computation, the size of $E^F_x/E^B_x$ is not more than a fixed number (=512 in the experiment).
  - For sampling evenly, down-sample the FG/BG mask to the corresponding area.

Reid Selection – Appearance Consistency

- **Appearance Consistency**
  - Use the Embedding Vectors to calculate the number of valid vectors with function $\varphi(\cdot, (\cdot, \cdot))$

$$
\varphi(V, (F, B)) = \sum_{v \in V} I(min_{f \in F} ||f^V - f^F|| - min_{f \in B} ||f^V - f^B|| < 0)
$$
Reid Selection – Appearance Consistency

- **Appearance Consistency**
  - Use the Embedding Vectors to calculate the number of valid vectors with function $\varphi(\cdot, (\cdot, \cdot))$
    \[
    \varphi(V, (F, B)) = \Sigma_{f' \in V} I(\text{min}_{f' \in F} |f' - f| - \text{min}_{f' \in B} |f' - f| < 0)
    \]
  - Use the valid number to calculate the appearance consistency with $L_{x-1}/L_0$
    \[
    S_{AC}(\text{Mask}, L_{x-1}) = \frac{\varphi\left(E^{FG}_x, (E^{FG}_{x-1}, E^{BG}_{x-1})\right) + \varphi\left(E^{FG}_{x-1}, (E^{FG}_x, E^{BG}_x)\right)}{|E^{FG}_x| + |E^{FG}_{x-1}|}
    \]

**Diagram:**
- $L_x$
- $I_x$
- Embedding Network
- Embedding Vectors
- Foreground Embedding Vectors $E^{FG}_x$
- Background Embedding Vectors $E^{BG}_x$
- Pixel Embedding Network
Framework – Reid Selection

Proposed Masks

Motion Consistency With $L_{x-1}$

Appearance Consistency With $L_{x-1}$

Appearance Consistency With $L_0$

Best Proposed Masks

– Why not check the motion consistency of propagated Mask?
– Because propagated mask has high motion consistency already

$L_0$ denotes the given mask.
Framework – Reid Selection

Proposed Masks

Motion Consistency With $L_{x-1}$

Appearance Consistency With $L_{x-1}$

Best Proposed Masks

If $S_{AC}(Mask, L_0)$ is smaller than a threshold (=0.1 in the experiment), then the object is considered to be missing.

Appearance Consistency With $L_0$

$L_x$

$L_0$ denotes the given mask.
If the object is missing, we enumerate $I_k/L_k$ from $k = 0$ to $k = x - 1$ to obtain the propagated mask. This is based on the observation that objects usually reappear at close positions in the image.
Results

• Summary of performance with different components
Results

• Our final results (rank 3rd)

• Validation set:

<table>
<thead>
<tr>
<th>#</th>
<th>User</th>
<th>Entries</th>
<th>Date of Last Entry</th>
<th>Overall ▲</th>
<th>J_seen ▲</th>
<th>J_unseen ▲</th>
<th>F_seen ▲</th>
<th>F_unseen ▲</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>speeding_zZ</td>
<td>62</td>
<td>08/28/18</td>
<td>0.710 (1)</td>
<td>0.725 (2)</td>
<td>0.644 (1)</td>
<td>0.761 (2)</td>
<td>0.711 (1)</td>
</tr>
<tr>
<td>2</td>
<td>Jono</td>
<td>8</td>
<td>08/27/18</td>
<td>0.703 (2)</td>
<td>0.744 (1)</td>
<td>0.606 (4)</td>
<td>0.789 (1)</td>
<td>0.675 (3)</td>
</tr>
<tr>
<td>3</td>
<td>linhj</td>
<td>26</td>
<td>08/29/18</td>
<td>0.697 (3)</td>
<td>0.723 (3)</td>
<td>0.631 (2)</td>
<td>0.736 (4)</td>
<td>0.698 (2)</td>
</tr>
</tbody>
</table>

• Test set:

<table>
<thead>
<tr>
<th>#</th>
<th>User</th>
<th>Entries</th>
<th>Date of Last Entry</th>
<th>Overall ▲</th>
<th>J_seen ▲</th>
<th>J_unseen ▲</th>
<th>F_seen ▲</th>
<th>F_unseen ▲</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Jono</td>
<td>4</td>
<td>09/01/18</td>
<td>0.722 (1)</td>
<td>0.737 (1)</td>
<td>0.648 (2)</td>
<td>0.778 (1)</td>
<td>0.725 (2)</td>
</tr>
<tr>
<td>2</td>
<td>speeding_zZ</td>
<td>8</td>
<td>09/01/18</td>
<td>0.720 (2)</td>
<td>0.725 (3)</td>
<td>0.663 (1)</td>
<td>0.752 (3)</td>
<td>0.741 (1)</td>
</tr>
<tr>
<td>3</td>
<td>mikirui</td>
<td>8</td>
<td>09/01/18</td>
<td>0.699 (3)</td>
<td>0.736 (2)</td>
<td>0.621 (4)</td>
<td>0.755 (2)</td>
<td>0.684 (4)</td>
</tr>
</tbody>
</table>
Visual Results

$I_x$

$L_x$
Future Direction

• Small object
• Learned parameters instead of fixed threshold
• Key multiple previous frames
• Long term understanding for retrieving missing object
• Speedup
• ...

Thanks & Questions