LOCATE: Self-supervised Object Discovery via Flow-guided Graph-cut and Bootstrapped Self-training
Silky Singh, Shripad Deshmukh, Mausoom Sarkar, Balaji Krishnamurthy
Media and Data Science Research, Adobe
- Code available at: https://github.com/silky1708/LOCATE
- Scan the QR on the right for paper link

1. INTRODUCTION

- self-supervised framework for video object segmentation (VOS)
- matches state-of-the-art (SOTA) performance on DAVIS16
- establishes a new SOTA on SegTrackv2 (~1 mIoU)
- inference with single images (no additional inputs required!)
- no post-processing! (deployable in real-world applications)
- trained on videos; exemplary zero-shot performance on images

2. METHOD

1. Graph-cut
   - Given video frame \( f \), divide \( f \) into square patches \( v_i \) of size \( p_x \times p_y \)
   - Build a fully-connected graph \( G=(V,E) \) on these image patches. \( V=(v_i) \)
   - \( E(v_i,v_j) \) is given by: cosine similarity (S) scores of the patch features from DINO (\( \phi \)) [5].
   - Specifically, \( E(v_i,v_j) = \alpha S(\phi(v_i), \phi(v_j)) + (1-\alpha) S(\phi(\text{flow}_i), \phi(\text{flow}_j)) \)
     where \( \alpha \in [0,1] \), flow, \( \text{flow} \), are the corresponding optical flow patches.

2. Bootstrapped self-training
   - Given \( N \) video frames, \( x_i \in \mathbb{R}^{W \times H \times 3} \), with corresponding graph-cut masks \( m_i \in \mathbb{R}^{W \times H} \)
   - We train a segmentation network, \( g_\theta \), minimizing cross-entropy (CE):
     \[ \theta^*_i = \arg \min_{\theta_i} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{CE}(m_i, g_\theta(x_i)) \]
   - Next, we iteratively train \( g_\theta \) with its outputs from previous rounds as supervisory signal
     \[ \theta^*_j = \arg \min_{\theta_j} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{CE}(g_{\theta^*_i}(x_i), g_\theta(x_i)) \]

3. QUANTITATIVE RESULTS

<table>
<thead>
<tr>
<th>Method</th>
<th>Supervision</th>
<th>Post-processing</th>
<th>DAVIS16 (mIoU)</th>
<th>SegTrackv2 (mIoU)</th>
<th>FBMS59 (mIoU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ponimatkin et. al. [1]</td>
<td>None</td>
<td>CRF</td>
<td>80.2</td>
<td>74.9</td>
<td>70.0</td>
</tr>
<tr>
<td>OCLR [2]</td>
<td>Synth</td>
<td>DINO-based TTA</td>
<td>80.9</td>
<td>72.3</td>
<td>72.7</td>
</tr>
<tr>
<td>DyStab [3]</td>
<td>Sup. feats</td>
<td>CRF</td>
<td>80.0</td>
<td>74.2</td>
<td>73.2</td>
</tr>
<tr>
<td>GWM [4]</td>
<td>Sup. feats</td>
<td>CRF + DINO</td>
<td>80.7</td>
<td>78.9</td>
<td>78.4</td>
</tr>
<tr>
<td>LOCATE (Ours)</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>

For detailed comparison, please check out the paper here: https://arxiv.org/abs/2308.11239

4. QUALITATIVE RESULTS

5. REFERENCES