VITA: Video Instance Segmentation via Object Token Association

Miran Heo*  
Yonsei University

Sukjun Hwang*  
Yonsei University

Seoung Wug Oh  
Adobe Research

Joon-Young Lee  
Adobe Research

Seon Joo Kim  
Yonsei University

{miran, sj.hwang, seonjookim}@yonsei.ac.kr  
{seoh, jolee}@adobe.com

Video frames

Video instance annotations
VITA: Video Instance Segmentation via Object Token Association

Online VIS approaches & Offline VIS approaches
VITA: Video Instance Segmentation via Object Token Association

Online VIS approaches & Offline VIS approaches
Online VIS approaches & **Offline VIS approaches**

**Advantages:**
1) they have a greater receptive field to the temporal axis
2) they can avoid error propagation derived from hand-crafted association algorithms.

**Disadvantages:**
such methods show difficulties in handling **long sequences** as the myriad of dense reference features hinders the Transformer layers from retrieving relevant information.
VITA: Video Instance Segmentation via Object Token Association

**VITA approach**

Hypotheses:
1) an image object detector can fully embody the context of an object into a feature vector (or a token);
2) a video can be represented by the relationship between the objects.

*VITA aims to parse an input video from the collection of object tokens without the necessity of referencing dense spatio-temporal backbone*
VITA: Video Instance Segmentation via Object Token Association

- **Frame-level Detector**
  - frame-independent manner; no inter-computation between frames is involved
VITA: Video Instance Segmentation via Object Token Association

- **Object Encoder**
  - Build temporal communication by employing self-attention along the temporal axis

- **Object Decoder and Output heads**
  - **Q:** $N_v$ learnable queries
  - **K, V:** object tokens
VITA: Video Instance Segmentation via Object Token Association

• **Object Encoder**
  - Build temporal communication by employing self-attention along the temporal axis

• **Object Decoder and Output heads**
  - \(Q: N_v\) learnable queries
  - \(K, V:\) object tokens
VITA: Video Instance Segmentation via Object Token Association

- **Object Encoder**
  - Build temporal communication by employing self-attention along the temporal axis

- **Object Decoder and Output heads**
  - \( Q: N_v \) learnable queries
  - \( K, V: \) object tokens
• Similarity loss

• Embed the collection through a linear layer.
• Measure the similarity of all possible pairs using a simple matrix multiplication.
• Binary cross entropy

Figure 4: Similarity loss. ⊘ and □ indicate video query and frame query, respectively. Same color represents same GT instance ID.
The table below compares methods on YouTube-VIS 2019.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone [13]</th>
<th>AP</th>
<th>AP$_{50}$</th>
<th>AP$_{75}$</th>
<th>AR$_1$</th>
<th>AR$_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VITA (Ours)</td>
<td>ResNet-50</td>
<td>49.8</td>
<td>72.6</td>
<td>54.5</td>
<td>49.4</td>
<td>61.0</td>
</tr>
<tr>
<td></td>
<td>ResNet-101</td>
<td>51.9</td>
<td>75.4</td>
<td>57.0</td>
<td>49.6</td>
<td>59.1</td>
</tr>
<tr>
<td></td>
<td>Swin-L</td>
<td>63.0</td>
<td>86.9</td>
<td>67.9</td>
<td>56.3</td>
<td>68.1</td>
</tr>
</tbody>
</table>

The tendency of offline methods with higher accuracy.
**YouYube VIS2021:** We hypothesize that the object-oriented design of VITA is more effective than typical dense Transformer decoders in addressing such challenging scenes.

**OVIS:** VITA is the first complete-offline approach to evaluate on OVIS valid set. (maximum 292 frames)
### Table 3: Impact of local windows of varying sizes in Object Encoder.

<table>
<thead>
<tr>
<th>$W$</th>
<th>AP</th>
<th>AP$_{50}$</th>
<th>AP$_{75}$</th>
<th>AR$_{1}$</th>
<th>AR$_{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>49.4</td>
<td>72.2</td>
<td>54.4</td>
<td>48.6</td>
<td>60.9</td>
</tr>
<tr>
<td>6</td>
<td>49.8</td>
<td>72.6</td>
<td>54.5</td>
<td>49.4</td>
<td>61.0</td>
</tr>
<tr>
<td>12</td>
<td>50.0</td>
<td>73.0</td>
<td>54.7</td>
<td>49.0</td>
<td>60.8</td>
</tr>
<tr>
<td>All</td>
<td>50.1</td>
<td>72.4</td>
<td>54.7</td>
<td>49.0</td>
<td>60.6</td>
</tr>
</tbody>
</table>

### Table 4: Maximum number of frames that can be processed at once using a single Titan XP.

<table>
<thead>
<tr>
<th>Method</th>
<th>Max Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$360 \times 640$</td>
</tr>
<tr>
<td>VisTR [27]</td>
<td>46</td>
</tr>
<tr>
<td>IFC [14]</td>
<td>123</td>
</tr>
<tr>
<td>Mask2Former-VIS [6]</td>
<td>81</td>
</tr>
<tr>
<td>VITA (Ours) $W = 3$</td>
<td>2677</td>
</tr>
<tr>
<td>VITA (Ours) $W = 6$</td>
<td>1392</td>
</tr>
<tr>
<td>VITA (Ours) $W = 12$</td>
<td>741</td>
</tr>
</tbody>
</table>