Templates for 3D Object Pose Estimation Revisited: Generalization to New objects and Robustness to Occlusions
Supplementary Material
Van Nguyen Nguyen\textsuperscript{1}, Yinlin Hu\textsuperscript{2}, Yang Xiao\textsuperscript{1}, Mathieu Salzmann\textsuperscript{2}, Vincent Lepetit\textsuperscript{1}
\textsuperscript{1}LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, France
\textsuperscript{2}CVLab, EPFL, Switzerland
\{van-nguyen.nguyen, yang.xiao, vincent.lepetit\}@enpc.fr
\{yinlin.hu, mathieu.salzmann\}@epfl.ch

1. Pose Discrimination

As discussed in Section 3.1.2 of the main paper, a drawback of global representations is their poor reliability to represent the real image of a new object even when the object identity is known and the background is uniform. To illustrate this, we show in Figure 1 the correlation between pose distances and representation distances as in [9, 1]. [9, 1] provided such plots only for seen objects and RGB-D data, we consider here objects that have been seen or unseen and we use RGB data only. As in [9, 1], the plots of Figure 1 are obtained by considering all possible pairs made of real images and synthetic images for a given object. Ideally, the plots should exhibit a diagonal pattern, in the region closed to the (0, 0) point on the bottom-left of the graph. This region corresponds to the critical region for correct image/template matching. A diagonal pattern corresponds to a strong correlation between pose differences and distances between representations.

More plots are given in Section 4 and they all yield to the same conclusion:

- The first column of Figure 1 shows that both representations result in a strong correlation for an seen object.
- The second column shows this correlation is lost when considering a new object for the global representation but not with ours.
- To check if this was due to the presence of clutter in the background of the real images, we removed the background by using the ground truth mask of the objects. The third column of Figure 1 shows that even without background, the correlation is still very poor for global representations. This can be explained by the fact that the pooling layers remove important information for unseen objects. We postulate that the rest of the architecture, in particular the fully connected layers learns to compensate this loss of information for seen objects, but such compensation is not possible for unseen objects.

2. Training details

Cropping on LINEMOD. Unless otherwise stated in previous works [9, 1], the cropping on LINEMOD and Occlusion-LINEMOD is done by virtually setting a box, 40 cm in each dimension, centered at the object as shown in Figure 2. When all the patches are extracted, we normalize them to the desired crop size. Please note that with this cropping, we do not consider in-plane rotations, in other words, we omit one additional degree of freedom.
Table 1: Comparison of our method with [9] and [1] on seen and unseen objects of LINEMOD (LM) and Occlusion-LINEMOD (O-LM) for the three splits detailed in Section 4.1 of the main paper. We report here the pose error, measured by the angle between the positions on the half-sphere for the ground truth pose and the predicted pose.

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>Features</th>
<th>Loss</th>
<th>Seen LM</th>
<th>Seen O-LM</th>
<th>Unseen LM</th>
<th>Unseen O-LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>[9] Base [9] Global [9]</td>
<td>12.1</td>
<td>13.2</td>
<td>12.0</td>
<td>12.4</td>
<td>49.5</td>
<td>51.1</td>
<td>52.3</td>
</tr>
<tr>
<td>[9] Base [9] Global InfoNCE [5]</td>
<td>6.6</td>
<td>6.5</td>
<td>6.7</td>
<td>6.6</td>
<td>47.9</td>
<td>45.2</td>
<td>52.9</td>
</tr>
<tr>
<td>[1] Base [9] Global [1]</td>
<td>11.2</td>
<td>11.8</td>
<td>12.9</td>
<td>12.0</td>
<td>49.5</td>
<td>51.1</td>
<td>52.3</td>
</tr>
<tr>
<td>[1] Base [9] Global InfoNCE [5]</td>
<td>6.4</td>
<td>6.4</td>
<td>6.5</td>
<td>6.4</td>
<td>46.6</td>
<td>47.2</td>
<td>50.4</td>
</tr>
<tr>
<td>Ours Base [9] Local [9]</td>
<td>15.2</td>
<td>15.8</td>
<td>14.9</td>
<td>15.3</td>
<td>32.6</td>
<td>31.9</td>
<td>31.0</td>
</tr>
<tr>
<td>Ours Base [9] Local InfoNCE [5]</td>
<td>4.8</td>
<td>5.1</td>
<td>7.9</td>
<td>5.9</td>
<td>12.3</td>
<td>18.5</td>
<td>21.8</td>
</tr>
<tr>
<td>[9] ResNet50 [4] Global InfoNCE [5]</td>
<td>3.6</td>
<td>4.3</td>
<td>4.7</td>
<td>4.2</td>
<td>26.7</td>
<td>29.8</td>
<td>34.5</td>
</tr>
<tr>
<td>[1] ResNet50 [4] Global InfoNCE [5]</td>
<td>3.7</td>
<td>4.5</td>
<td>5.1</td>
<td>4.4</td>
<td>35.7</td>
<td>29.8</td>
<td>39.7</td>
</tr>
<tr>
<td>Ours ResNet50 [4] Local InfoNCE [5]</td>
<td>3.3</td>
<td>4.6</td>
<td>3.4</td>
<td>3.7</td>
<td>9.7</td>
<td>11.1</td>
<td>11.5</td>
</tr>
</tbody>
</table>

Table 2: Network initializations evaluated on T-LESS. Using pre-trained features from MOCOv2 [2] brings some improvement comparing to training from scratch.

<table>
<thead>
<tr>
<th>Initialization</th>
<th>Number templates</th>
<th>Recall VSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>From scratch</td>
<td>21K</td>
<td>55.42</td>
</tr>
<tr>
<td>MOCOv2 [2]</td>
<td>21K</td>
<td>59.14</td>
</tr>
<tr>
<td>Obj. 1-18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obj. 19-30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Projective distance estimation

As done in [7, 6], we estimate 3D translation in the query image from the retrieved template and the input bounding box as detailed in Section 3.6.2 of [8]. More precisely, given known camera intrinsic of both real sensor $K_{query}$ and of the synthetic view $K_{temp}$, we estimate the distance $t_{query,z}$ of real image:

$$t_{query,z} = t_{temp,z} \times \frac{||bb_{temp,z}||}{||bb_{query,z}||} \times \frac{f_{query}}{f_{temp}}$$

(1)

where $||bb_{(i)}||$ is the diagonal of the bounding box and $||f_{(i)}||$ is the focal length.

Then, we can estimate the vector to transform from the object center in the synthetic view to the query image:

$$\Delta t = i_{query,z}K_{query}^{-1}bb_{query,c} - i_{temp,z}K_{temp}^{-1}bb_{temp,c}$$

(2)

where $bb_{(i),c}$ is the bounding box centers in homogeneous coordinates.

Finally, the 3D translation in the query image $i_{query}$ can be estimated as:

$$i_{query} = i_{temp} + \Delta t$$

(3)

where $i_{temp} = (0,0,i_{temp,z})$, the translation from camera to object center in the synthetic view.
4. Additional Results

4.1. Quantitative Results

We show in Table 1 quantitative results with pose error, measured by the angle between the two positions on the viewing half-sphere.

4.2. Qualitative Results

We show in Figures 4, 5, 6, 7, 8, 9, 10 additional qualitative results on T-LESS and each split of LINEMOD and Occlusion-LINEMOD.

References


Figure 4: **Qualitative results on unseen objects of T-LESS dataset.**
Figure 5: **Qualitative results four test sets on Split #1:** of seen objects of LINEMOD (top left), seen objects of Occlusion-LINEMOD (top right), unseen objects of LINEMOD (bottom left) and unseen objects of Occlusion-LINEMOD (bottom right).

Figure 6: **Visualization of the correlation between pose distances and representation distances on Split #1:** of unseen objects of LINEMOD (two first rows and two first columns from the left of two last rows) and unseen objects of LINEMOD (fours last columns from the left of two last rows). Ideally, the plots should exhibit the diagonal pattern at the region closed to the (0, 0) point on the bottom-left that corresponds to the critical region for correct image/template matching, showing a strong correlation between pose differences and representation distances. The plots of seen objects of LINEMOD show that both representations result in a strong correlation for training objects. The plots of unseen objects of LINEMOD show this correlation is lost when considering a new object for the global representation [1] but not with ours.
Figure 7: **Qualitative results four test sets on Split #2**: of seen objects of LINEMOD (top left), seen objects of Occlusion-LINEMOD (top right), unseen objects of LINEMOD (bottom left) and unseen objects of Occlusion-LINEMOD (bottom right).

Figure 8: **Visualization of the correlation between pose distances and representation distances on Split #2**: seen objects of LINEMOD (two first rows and two first columns from the left of two last rows) and unseen objects of LINEMOD (fours last columns from the left of two last rows).
Figure 9: **Qualitative results four test sets on Split #3**: of seen objects of LINEMOD (top left), seen objects of Occlusion-LINEMOD (top right), unseen objects of LINEMOD (bottom left) and unseen objects of Occlusion-LINEMOD (bottom right).

Figure 10: **Visualization of the correlation between pose distances and representation distances on Split #3**: seen objects of LINEMOD (two first rows and two first columns from the left of two last rows) and unseen objects of LINEMOD (fours last columns from the left of two last rows).