A Appendix

A.1 Latent Space Visualization

Fig. 1: 3 principle components $p_{ci}$ of the latent space; red: all test encodings of scene 2, object 5, T-LESS (upper hemisphere); other colors: encodings of synthetic model views (whole view-sphere, color is defined by elevation)

A.2 Augmentation Parameters

Table 1: Augmentation Parameters; Scale and translation is in relation to image shape and occlusion is in proportion of the object mask

<table>
<thead>
<tr>
<th>add</th>
<th>contrast</th>
<th>multiply</th>
<th>invert</th>
<th>Gaussian blur</th>
<th>ambient</th>
<th>diffuse</th>
<th>specular</th>
<th>scale</th>
<th>translation</th>
<th>occlusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>50% chance $\mathcal{U}(-25, 25)$</td>
<td>$\mathcal{U}(0.4, 2.3)$</td>
<td>$\mathcal{U}(0.6, 1.4)$ per channel</td>
<td>$\mathcal{U}(0, 1.2\sigma)$</td>
<td>0.4</td>
<td>$\mathcal{U}(0.7, 0.9)$</td>
<td>$\mathcal{U}(0.2, 0.4)$</td>
<td>$\mathcal{U}(0.8, 1.2)$</td>
<td>$\mathcal{U}(-0.15, 0.15)$</td>
<td>$\mathcal{U}(0, 0.25)$</td>
<td></td>
</tr>
</tbody>
</table>
A.3 Failure Cases
Different kinds of failure cases:

Fig. 2: Failure cases; Blue: True poses; Green: Predictions; (a) Failed detections due to occlusions and object ambiguity, (b) failed AAE predictions of Glue (middle) and Eggbox (right) due to occlusion, (c) inaccurate AAE predictions due to occlusion.

A.4 ICP Refinement Details
We first project the depth image crops into a 3D point cloud and generate 3000 random points on the surface of the object CAD model. The Iterative Closest Point (ICP) is performed between these point sets based on the implementation of https://github.com/ClayFlannigan/icp. The refinement is first applied in direction of the camera axis where most of the RGB-based errors stem from and then on the full 6D pose. If the 3D orientation changes more than 20°, the latter refinement is discarded.

A.5 Example Sequences T-LESS
Fig. 3: Ground truth (blue) and estimated (green) 6D pose of T-LESS objects; tested on scene 15; no green box means that 2D detection failed
Fig. 4: Ground truth (blue) and estimated (green) 6D pose of T-LESS objects; tested on scene 11
Fig. 5: Incomplete IKEA mug 3D orientation estimation from webcam stream (left), nearest training neighbors (right)

Fig. 6: Ground truth (blue) and estimated (green) 6D pose of LineMOD objects, occlusion set (last row) tested with ground truth 2D bounding boxes