# Templates for 3D Object Pose Estimation Revisited: Generalization to New objects and Robustness to Occlusions Supplementary Material


Van Nguyen Nguyen[1], Yinlin Hu[2], Yang Xiao[1], Mathieu Salzmann[2], Vincent Lepetit[1]

[1]LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, France

[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.

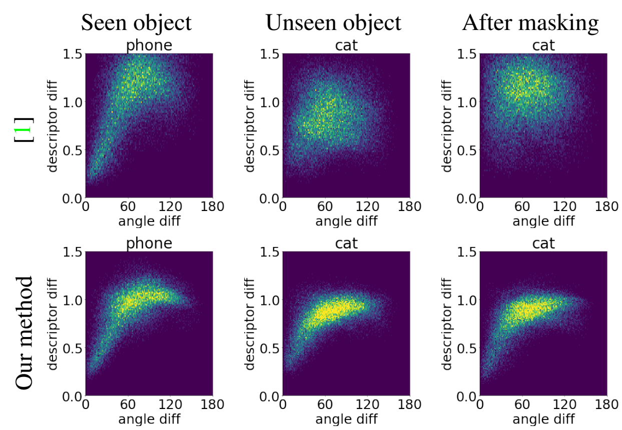


Figure 1: **Visualization of the correlation between pose distances and representation distances,** for understanding the discriminative power of different image representations for pose retrieval. First row is for [1], second row is for our method. Please see Section 1.

## 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.

| Method | Backbone | Features | Loss | Seen LM | | | | Seen O-LM | | | | Unseen LM | | | | Unseen O-LM | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | #1 | #2 | #3 | Avg. | #1 | #2 | #3 | Avg. | #1 | #2 | #3 | Avg. | #1 | #2 | #3 | Avg. |
| [9] | Base [9] | Global | [9] | 12.1 | 13.2 | 12.0 | 12.4 | 49.5 | 51.1 | 52.3 | 50.9 | 54.6 | 55.7 | 59.0 | 56.4 | 59.4 | 57.2 | 56.0 | 57.5 |
| [9] | Base [9] | Global | InfoNCE [5] | 6.6 | 6.5 | 6.7 | 6.6 | 47.9 | 45.2 | 52.9 | 58.6 | 58.6 | 48.5 | 48.1 | 51.7 | 61.4 | 56.3 | 54.6 | 57.3 |
| [1] | Base [9] | Global | [1] | 11.2 | 11.8 | 12.9 | 12.0 | 49.5 | 51.1 | 52.3 | 51.0 | 54.6 | 55.7 | 59.0 | 56.4 | 60.0 | 53.8 | 60.1 | 57.9 |
| [1] | Base [9] | Global | InfoNCE [5] | 6.4 | 6.4 | 6.5 | 6.4 | 46.6 | 47.2 | 50.4 | 48.0 | 67.9 | 48.6 | 50.8 | 55.7 | 73.4 | 56.1 | 53.4 | 60.9 |
| Ours | Base [9] | Local | [9] | 15.2 | 15.8 | 14.9 | 15.3 | 32.6 | 31.9 | 31.0 | 31.8 | 27.1 | 27.4 | 25.3 | 26.5 | 41.5 | 41.2 | 42.3 | 41.6 |
| Ours | Base [9] | Local | InfoNCE [5] | 4.8 | 5.1 | 7.9 | 5.9 | 12.3 | 18.5 | 21.8 | 18.5 | 15.4 | 9.9 | 20.3 | 15.2 | 32.3 | 21.3 | 17.6 | 23.7 |
| [9] | ResNet50 [4] | Global | InfoNCE [5] | 3.6 | **4.3** | 4.7 | 4.2 | 26.7 | 29.8 | 34.5 | 30.3 | 43.1 | 42.7 | 40.5 | 42.1 | 45.8 | 51.8 | 44.0 | 47.1 |
| [1] | ResNet50 [4] | Global | InfoNCE [5] | 3.7 | 4.5 | 5.1 | 4.4 | 35.7 | 29.8 | 39.7 | 35.1 | 51.1 | 50.0 | 39.3 | 46.8 | 64.5 | 61.0 | 49.8 | 58.4 |
| Ours | ResNet50 [4] | Local | InfoNCE [5] | **3.3** | 4.6 | **3.4** | **3.7** | **9.7** | **11.1** | **11.5** | **10.7** | **7.5** | **3.1** | **10.0** | **6.9** | **17.5** | **11.5** | **7.5** | **12.2** |

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.

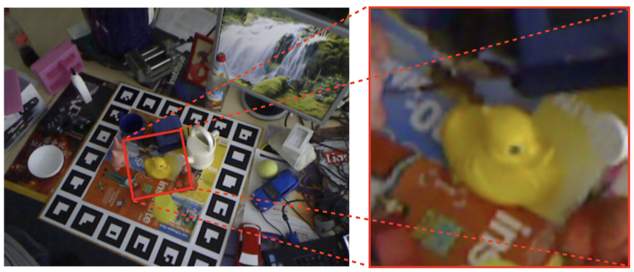

Figure 2: **Cropping on LINEMOD.** the cropping on LINEMOD and Occlusion-LINEMOD is done by virtually setting a box centered at the object. Please see Section 2.

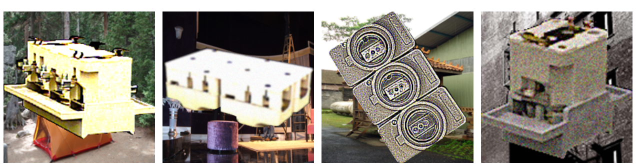

Figure 3: **Examples of training images.** We use randomized background from SUN397 [10] with data augmentation, including Gaussian blur, contrast, brightness, color and sharpness filters from the Pillow library [3].

**Data augmentation on T-LESS.** As done in [7, 6], we also apply data augmentation to the input images of T-LESS during training. We use Gaussian blur, contrast, brightness, color, and sharpness filters with the Pillow library [3]. Some training samples can be seen in Figure 3.

**Pre-trained Features.** We intialize the network ResNet50 with MOCOv2's features [2]. It has been shown in [11] that this initialization can improve both the convergence and the performance. We show in Table 2 our comparison on T-LESS dataset when training the network ResNet50 from scratch and with initializing pre-trained features of MOCOv2 [2].

| Initialization | Number templates | Recall VSD | | |
|---|---|---|---|---|
| | | Obj. 1-18 | Obj. 19-30 | Avg |
| From scratch | 21K | 55.42 | 51.40 | 53.81 |
| MOCOv2 [2] | 21K | **59.14** | **56.91** | **58.25** |

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

## 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 $\hat{t}_{query,z}$ of real image:

$$\hat{t}_{query,z} = t_{temp,z} \times \frac{||bb_{temp}||}{||bb_{query}||} \times \frac{f_{query}}{f_{temp}} \quad (1)$$

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

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

$$\Delta\hat{t} = \hat{t}_{query,z} K_{query}^{-1} bb_{query,c} - \hat{t}_{temp,z} K_{temp}^{-1} bb_{temp,c} \quad (2)$$

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

Finally, the 3D translation in the query image $\hat{t}_{query}$ can be estimated as :

$$\hat{t}_{query} = \hat{t}_{temp} + \Delta\hat{t} \quad (3)$$

where $\hat{t}_{temp} = (0, 0, \hat{t}_{temp,z})$, 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


[1] Vassileios Balntas, Andreas Doumanoglou, Caner Sahin, Juil Sock, Rigas Kouskouridas, and Tae-Kyun Kim. Pose Guided RGBD Feature Learning for 3D Object Pose Estimation. In *International Conference on Computer Vision (ICCV)*, 2017.

[2] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *ArXiv*, 2020.

[3] Alex Clark et al. The pillow imaging library. https://github.com/python-pillow/pillow. *IEEE Robot. Autom. Mag.*, 22(3):36–52, Sept. 2015.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] Aaron Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *ArXiv*, 2018.

[6] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O. Arras, and Rudolph Triebel. Multi-Path Learning for Object Pose Estimation Across Domains. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. In *European Conference on Computer Vision (ECCV)*, pages 699–715, 2018.

[8] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, and Rudolph Triebel. Augmented Autoencoders: Implicit 3D Orientation Learning for 6D Object Detection. *IJCV*, 128(3):714–729, 2020.

[9] Paul Wohlhart and Vincent Lepetit. Learning Descriptors for Object Recognition and 3D Pose Estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[11] Yang Xiao, Yuming Du, and Renaud Marlet. PoseContrast: Class-Agnostic Object Viewpoint Estimation in the Wild with Pose-Aware Contrastive Learning. In *International Conference on 3D Vision (3DV)*, 2021.

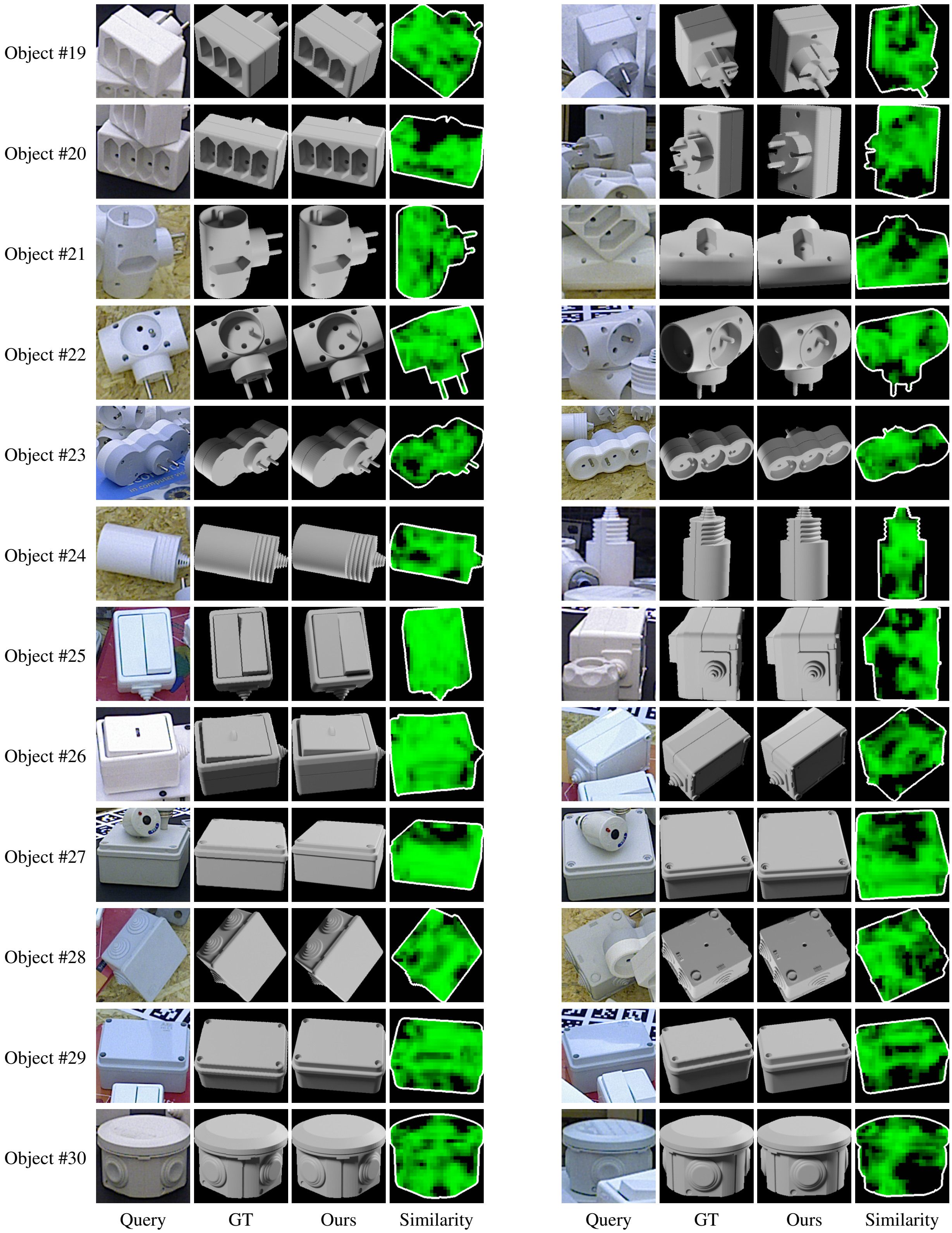


Figure 4: **Qualitative results on unseen objects of T-LESS dataset.**

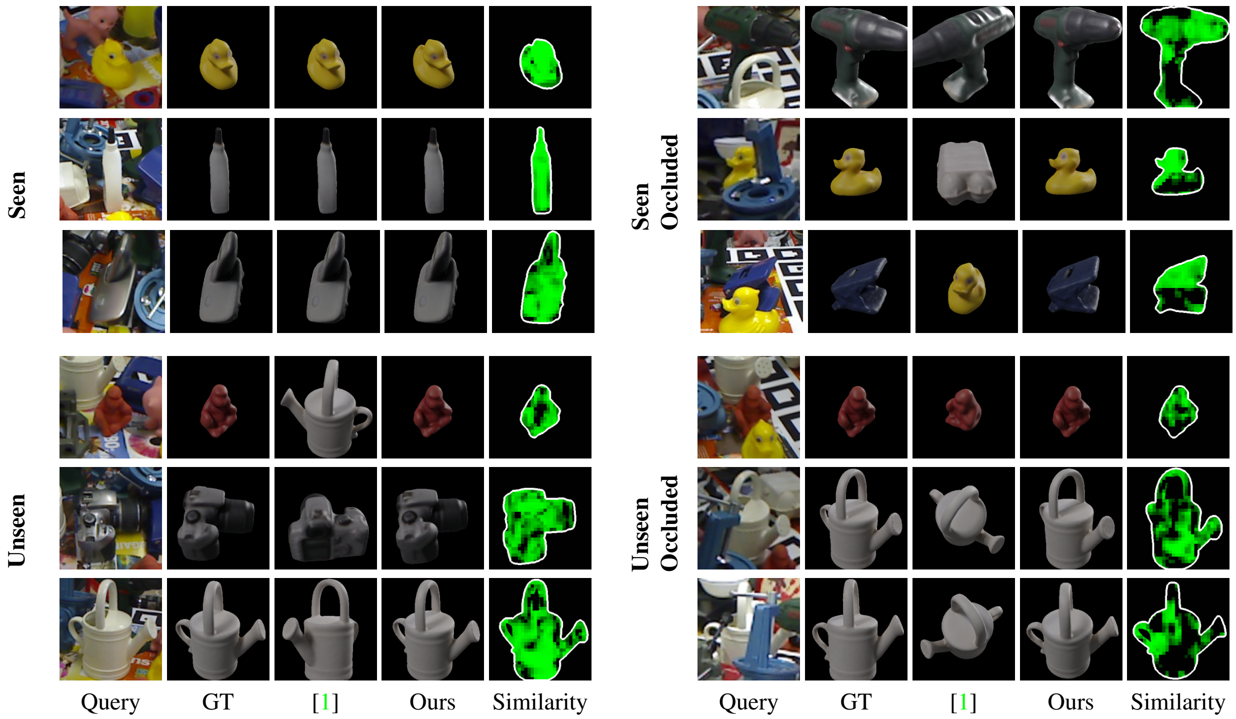


Figure 5: **Qualitative results four test sets on Split #1:** of seen objects of LINEMOD (top left), seen objects of Occlusion-LINEMOD (top right), unseen objects of LINEMOD (bottom left) and unseen objects of Occlusion-LINEMOD (bottom right).

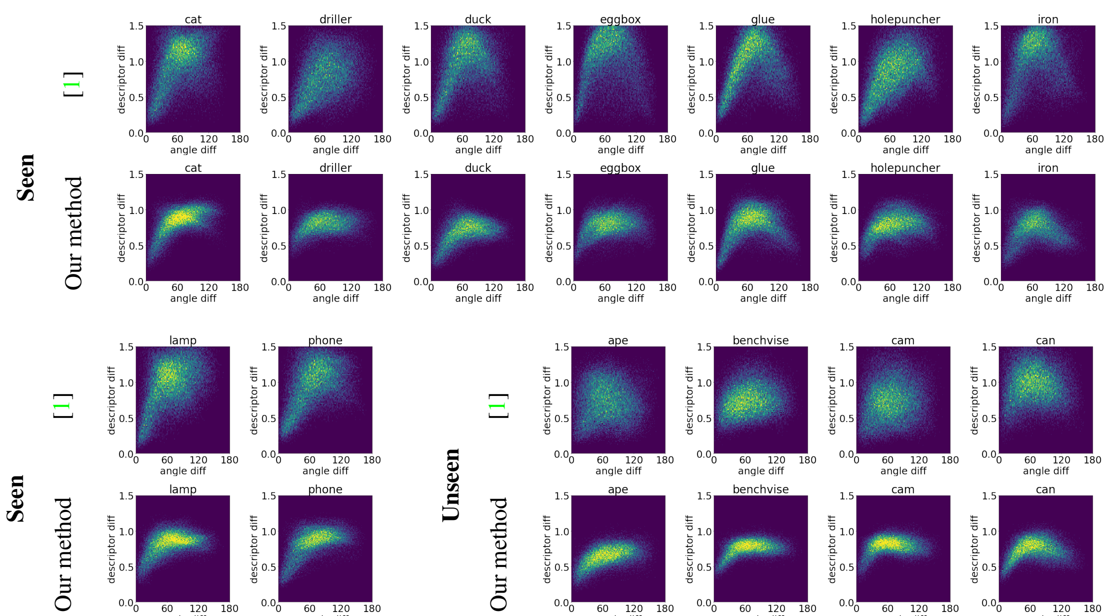


Figure 6: **Visualization of the correlation between pose distances and representation distances on Split #1**: of unseen objects of LINEMOD (two first rows and two first columns from the left of two last rows) and unseen objects of LINEMOD (fours last columns from the left of two last rows). Ideally, 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.

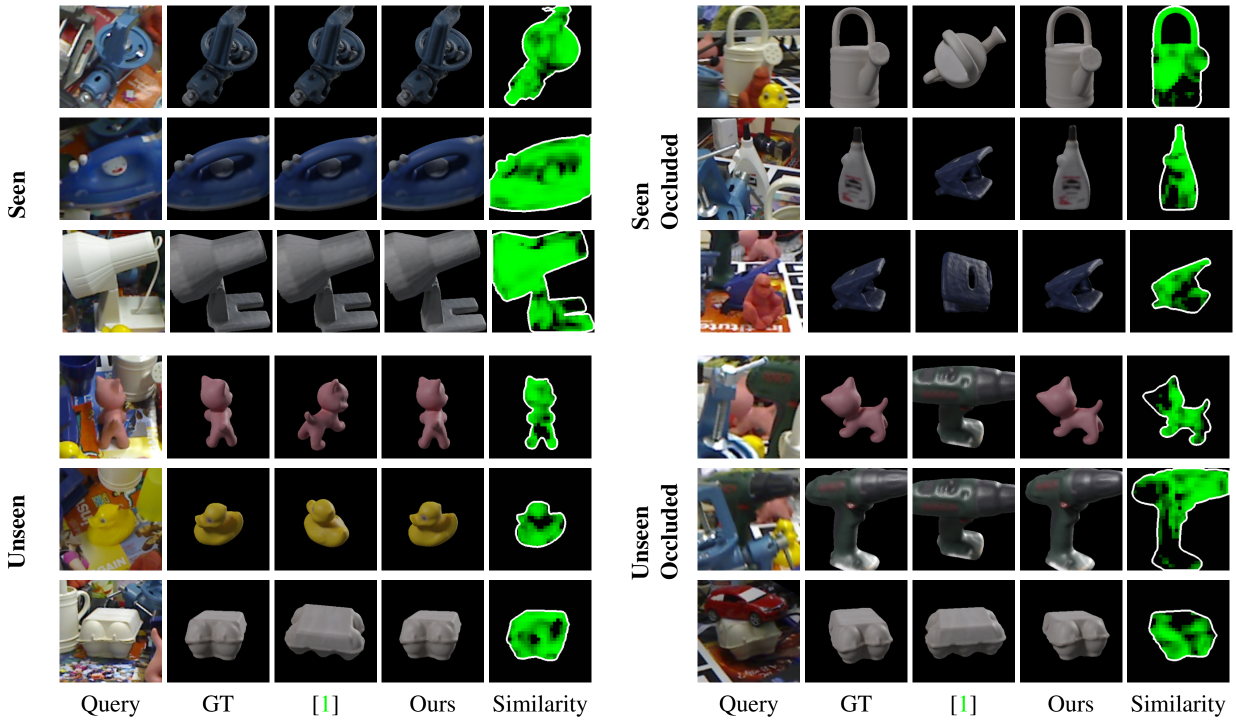


Figure 7: **Qualitative results four test sets on Split #2:** of seen objects of LINEMOD (top left), seen objects of Occlusion-LINEMOD (top right), unseen objects of LINEMOD (bottom left) and unseen objects of Occlusion-LINEMOD (bottom right).

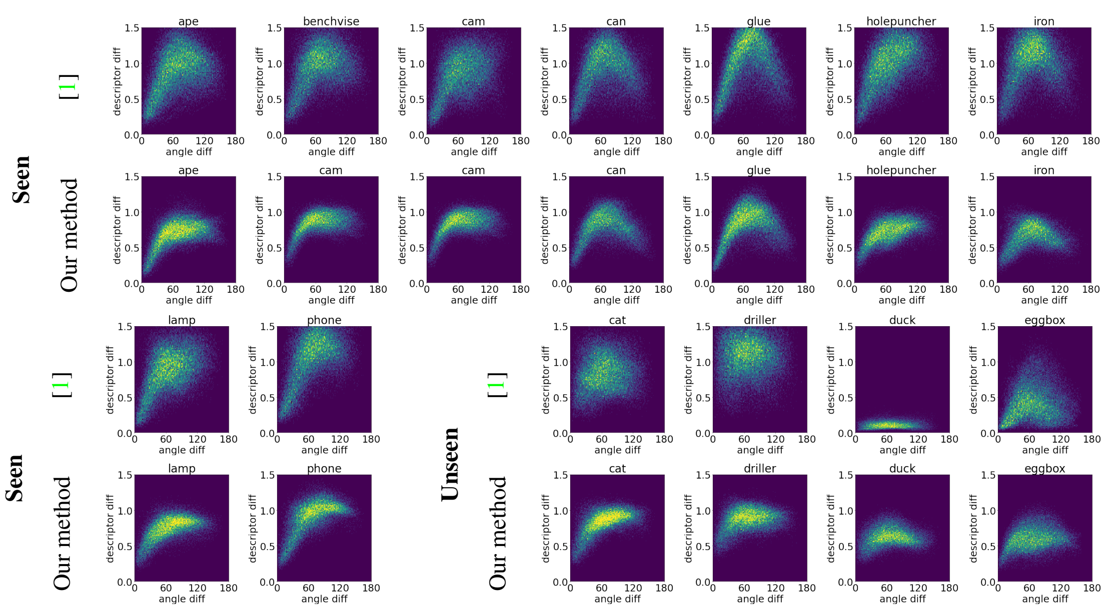


Figure 8: **Visualization of the correlation between pose distances and representation distances on Split #2**: seen objects of LINEMOD (two first rows and two first columns from the left of two last rows) and unseen objects of LINEMOD (fours last columns from the left of two last rows).

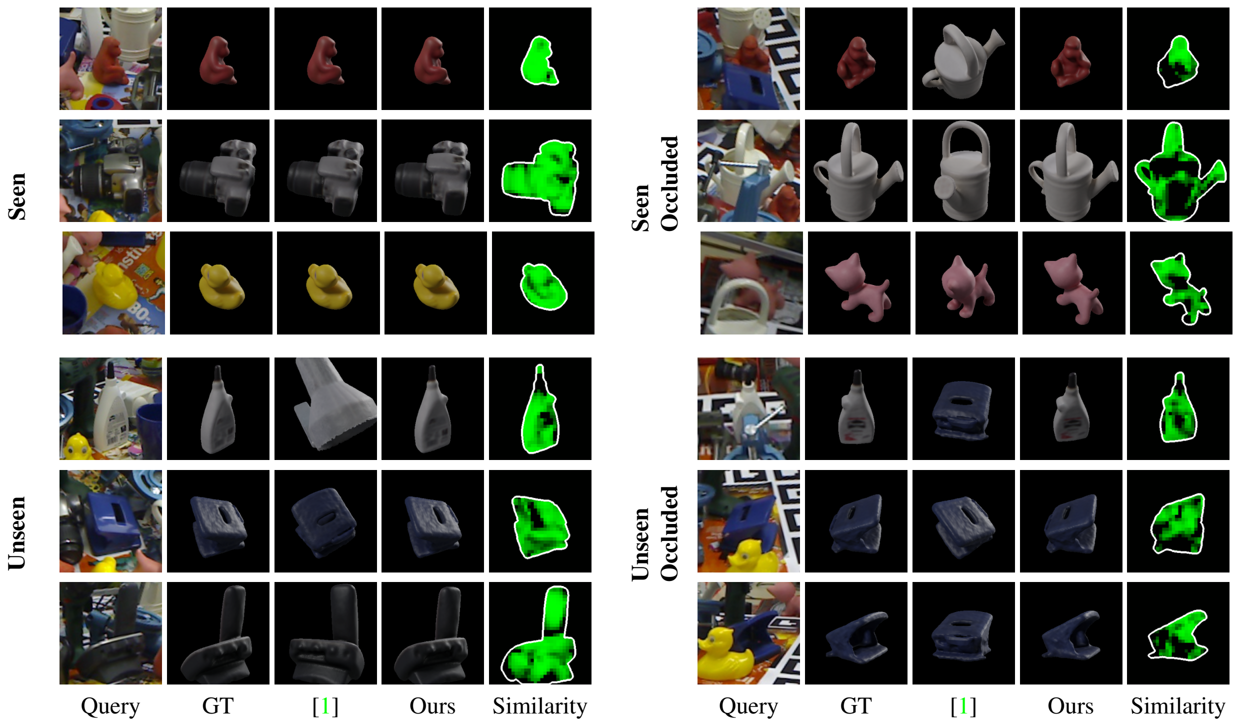


Figure 9: **Qualitative results four test sets on Split #3:** of seen objects of LINEMOD (top left), seen objects of Occlusion-LINEMOD (top right), unseen objects of LINEMOD (bottom left) and unseen objects of Occlusion-LINEMOD (bottom right).

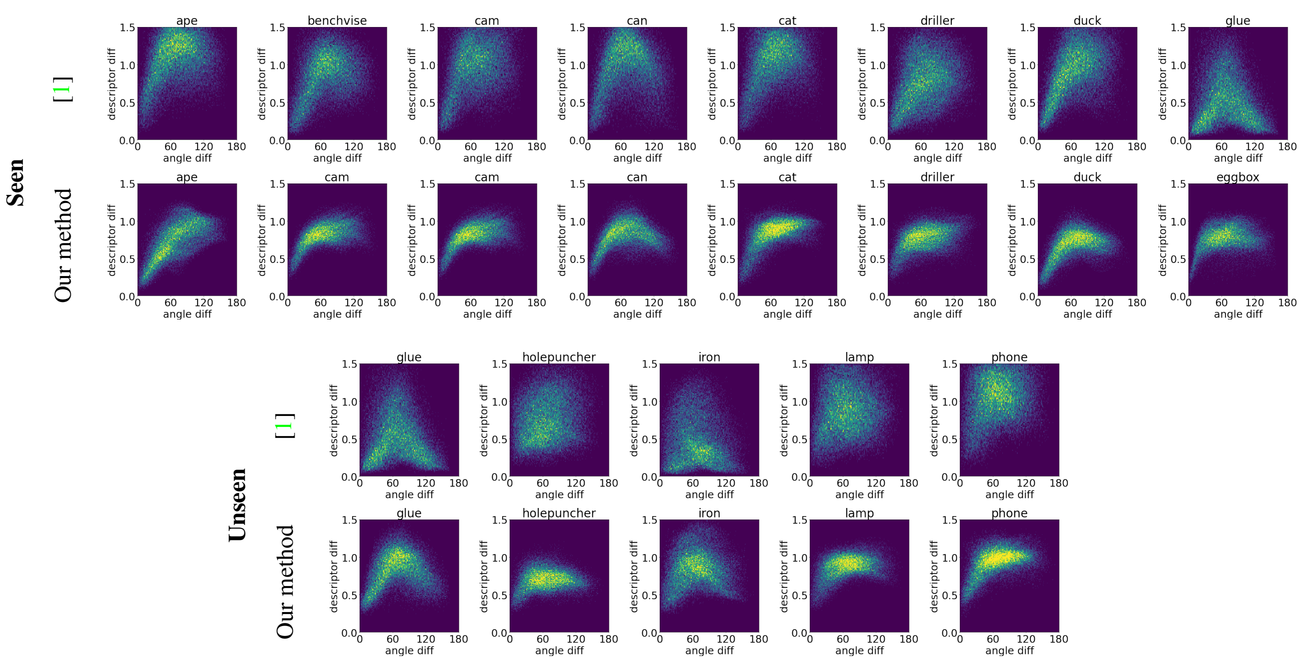


Figure 10: **Visualization of the correlation between pose distances and representation distances on Split #3**: seen objects of LINEMOD (two first rows and two first columns from the left of two last rows) and unseen objects of LINEMOD (fours last columns from the left of two last rows).