Fusing Local Similarities for Retrieval-based 3D Orientation Estimation of Unseen Objects

Chen Zhao[1], Yinlin Hu[1,2], and Mathieu Salzmann[1,2]

[1]CVLab EPFL, [2]ClearSpace SA
{chen.zhao, yinlin.hu, mathieu.salzmann}@epfl.ch

1 Network Architecture

Fig. 1 illustrates the network architectures of our multi-scale feature extraction module and adaptive fusion module. The multi-scale feature extraction module generates three feature maps of different sizes by default, i.e., $\mathbf{F}_1 \in \mathbb{R}^{13\times13\times128}$, $\mathbf{F}_2 \in \mathbb{R}^{16\times16\times128}$, and $\mathbf{F}_3 \in \mathbb{R}^{32\times32\times128}$. Given a pair of source image and reference image, the corresponding feature maps are paired to compute $\mathbf{F}_1^*$, $\mathbf{F}_2^*$, and $\mathbf{F}_3^*$, respectively. The local similarities are then adaptively fused into an image similarity score via our adaptive fusion module. Every convolution layer (conv) and fully connected layer (FC) is followed by ReLU, except for the ones in grey.

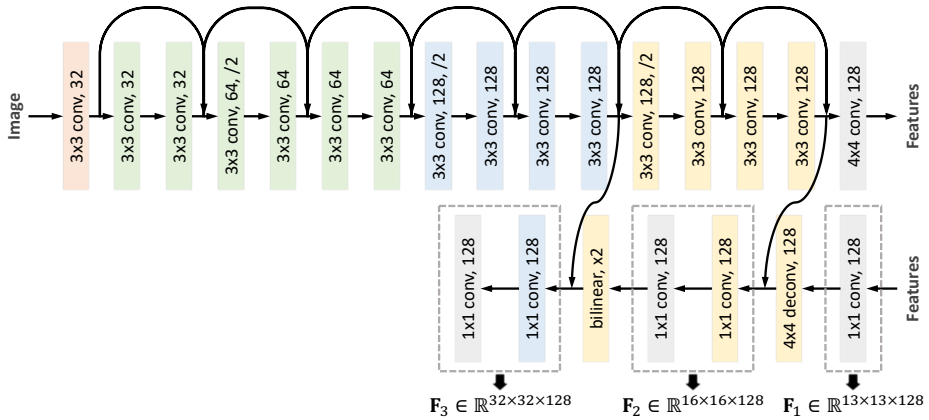
2 Canonical Frame

Fig. 2 shows the importance of a canonical frame to the 3D object orientation estimation. Specifically, the expected $\mathbf{R}_{src}$ is the relative rotation between the camera frame and the object frame. Since the object can be placed in the real world with arbitrary poses, multiple possible object frames could exist. These object frames correspond to different relative rotations. Therefore, $\mathbf{R}_{src}$ is ill-defined without a canonical frame. To address this issue, one could define a common canonical frame for all objects, but the shape variation among different objects makes this definition unreasonable. Consequently, we assume the 3D object models to be known in our experiments, which are employed to define the canonical frames dependently. Holding this assumption, we also use the 3D models to generate synthetic reference images for our retrieval-based 3D object orientation estimation.

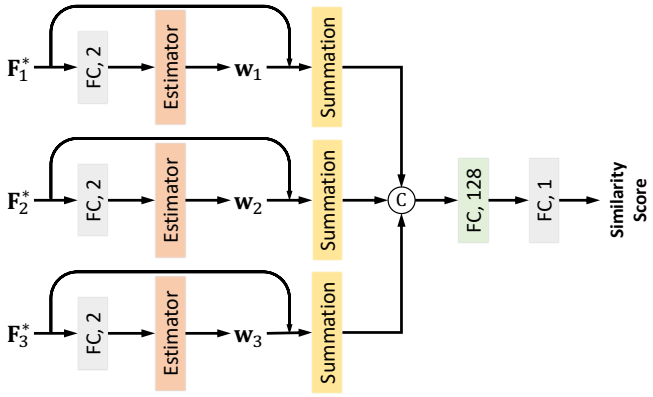
3 Experimental Setup

Table 1 shows the data splitting of LineMOD and LineMOD-O. The cropped data are assigned to three different groups according to the depicted objects.

Recall that the reference images are generated by rendering the corresponding 3D object model with different 3D orientation. The 3D orientation is formalized as a 3D rotation matrix. In our experiments, the 3D rotation matrix is randomly



(a) Multi-Scale Feature Extraction



(b) Adaptive Fusion

Fig. 1. Network Architecture. $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{F}_3$ indicate the three feature maps used to estimate local similarities. The estimator is used for the confidence map estimation.

Table 1. Data splitting of LineMOD and LineMOD-O.

Dataset	**Split #1**	**Split #2**	**Split #3**
LineMOD	ape, benchvise camera, can	cat, driller duck, eggbox	glue, holepuncher iron, lamp, phone
LineMOD-O	ape, can	cat, driller, duck	eggbox, glue, holepuncher

sampled, using the 6D representation [9]. Fig. 3 illustrates the distribution of the sampled 3D rotation matrices, which is visualized by using t-SNE [5]. Compared with the samples using the representation of Euler angles [7] (Fig. 3(b)), the ones based on the 6D representation (Fig. 3(a)) are scattered in the sample space more evenly.

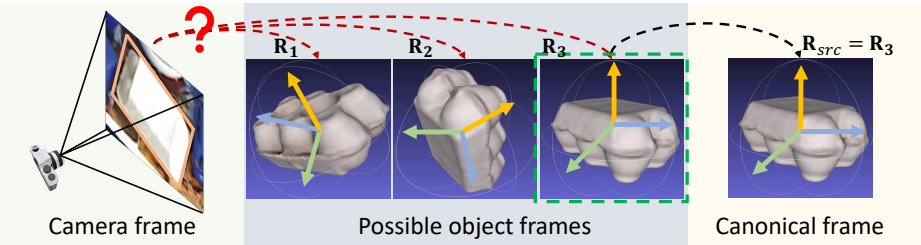
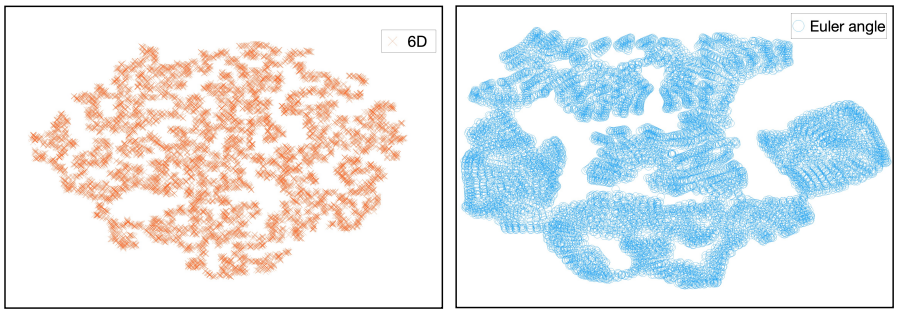


Fig. 2. Importance of a canonical frame. Different object frames correspond to different relative 3D rotations, which makes the 3D object orientation in the camera frame ill-defined. A canonical frame is required to uniquely define $\mathbf{R}_{src}$.

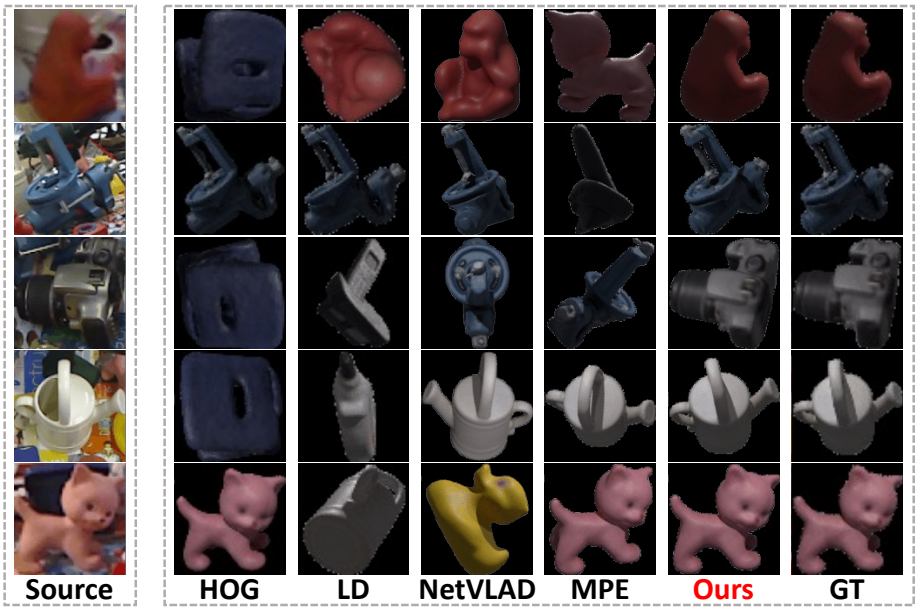


(a) 6D representation (b) Euler angle

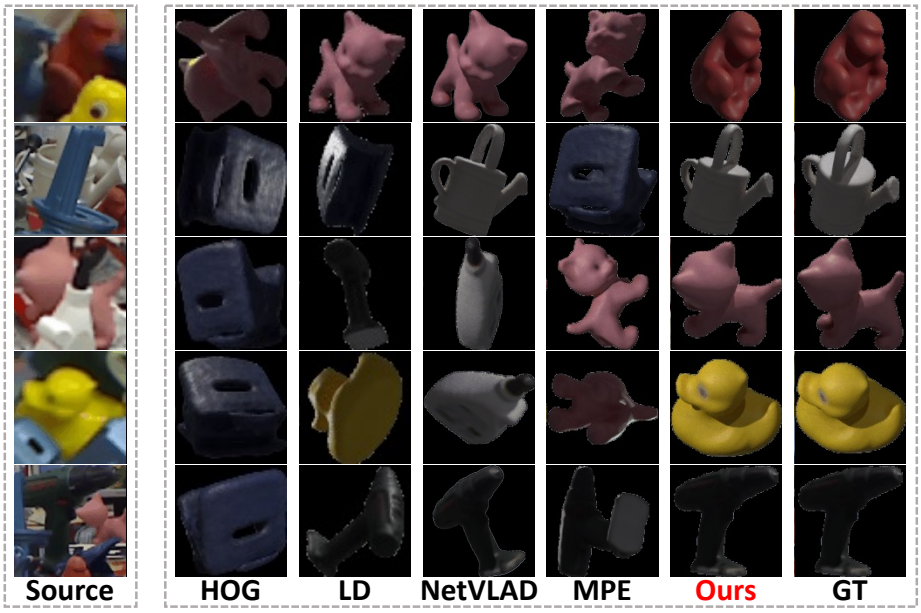
Fig. 3. Distribution of the sampled 3D rotation matrices. The 3D object orientation (3D rotation matrix) of the reference images is sampled using 6D representation [9] (a) and Euler angles [7] (b), respectively. The distribution is visualized by using t-SNE [5].

4 Qualitative Results

Fig. 4 shows some qualitative results in the presence of unseen objects on LineMOD and LineMOD-O. The images in the leftmost column are the real source images and the ones in the rightmost column are the most similar synthetic references. The other images are the retrieved results of the evaluated methods. One can observe that given some unseen objects, the previous approaches either select wrong objects or pick up the correct objects but with incorrect 3D orientation. By contract, our method is capable of robustly retrieving the reference similar to the source image, and thus ensures the 3D orientation estimation accuracy for unseen objects.



(a) LineMOD



(b) LineMOD-O

Fig. 4. Qualitative Results in the presence of unseen objects.

Table 2. Rota. Acc. (%) on LineMOD [4] in the case of unseen objects.

	Split #1	Split #2	Split #3	Mean
HOG [3]	71.43	66.24	53.75	63.81
LD [8]	14.39	19.46	13.32	15.72
NetVLAD [1]	42.84	38.04	41.31	40.73
MPE [6]	53.15	44.77	67.76	55.23
Ours	**90.37**	**82.00**	**79.17**	**83.85**

Table 3. Rota. Acc. (%) on LineMOD-O [2] in the case of unseen objects.

	Split #1	Split #2	Split #3	Mean
HOG [3]	34.51	34.92	28.01	32.48
LD [8]	7.85	4.30	6.98	6.38
NetVLAD [1]	37.33	24.31	23.37	28.34
MPE [6]	27.81	7.52	21.40	18.91
Ours	**64.43**	**54.69**	**39.64**	**52.92**

5 Quantitative Results

In the main paper, we assume that the object category is unknown on LineMOD and LineMOD-O, and the evaluated methods are used to both classify the object and estimate its 3D orientation. Therefore, the 3D object orientation estimation accuracy is related to the object classification accuracy (please refer to Eq. 9 in our main paper). To further evaluate the unseen-object generalization in the context of pure 3D object orientation estimation, we conduct another experiment on LineMOD and LineMOD-O, assuming the object category is known. In this case, Rota. Acc. is computed as

$$\text{Rota. Acc.} = \begin{cases} 1 & \text{if } d(\hat{\mathbf{R}}_{ref}, \mathbf{R}_{src}) < \lambda \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

We report Rota. Acc. (%) for unseen objects in Table 2 and Table 3. As this benchmark is less challenging, all methods yield better results compared with the ones reported in Sec. 4.3 of the main paper. In this context, our method still surpasses the competitors by a significantly large margin. This observation indicates the better distinctiveness of our method towards the 3D orientation of unseen objects. The superior results on LineMOD-O also evidence the robustness of our method to occlusions.

6 Failure Cases

Fig. 6 illustrates some failure cases of our method on LineMOD. One can observe that there is a flipping issue in these cases. Taking the benchvise as an example, it is difficult to distinguish between our result and the ground-truth one,

Fig. 5. Failure cases on LineMOD [4] in the presence of unseen objects.

which makes this problem challenging. To address this issue, the network should be able to extract fine-grained information. Ideally, our patch-level solution is more capable of capturing such information than image-level one. Therefore, in our future work, we plan to utilize hard example mining over the patch-level comparison to make our method pay more attention to fine-grained differences.

References

1. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 5297–5307 (2016)
2. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6d object pose estimation using 3d object coordinates. In: Proceedings of the European Conference on Computer Vision. pp. 536–551. Springer (2014)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 886–893. Ieee (2005)
4. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: Asian Conference on Computer Vision. pp. 548–562. Springer (2012)
5. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(11) (2008)
6. Sundermeyer, M., Durner, M., Puang, E.Y., Marton, Z.C., Vaskevicius, N., Arras, K.O., Triebel, R.: Multi-path learning for object pose estimation across domains. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 13916–13925 (2020)
7. Sundermeyer, M., Marton, Z.C., Durner, M., Brucker, M., Triebel, R.: Implicit 3d orientation learning for 6d object detection from rgb images. In: Proceedings of the European Conference on Computer Vision. pp. 699–715 (2018)
8. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3d pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3109–3118 (2015)
9. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5745–5753 (2019)