RefPose: Leveraging Reference Geometric Correspondences for Accurate 6D Pose Estimation of Unseen Objects

Supplementary Material

This Supplementary Material provides the following extended analyses, experimental findings, and additional details that complement our main paper:

- Additional training details, including dataset specifications, training duration, and memory requirements.
- A detailed description of the relative pose estimator used in our framework.
- Additional qualitative results demonstrating the effectiveness of our method across various datasets.

1. Additional Training Details

The training process consists of three stages: optical flow fine-tuning, a coarse stage, and a refinement stage. All stages were trained using image pairs constructed from the GSO [1] dataset. As noted in MegaPose [3], GSO plays a significant role in performance. We observed that training exclusively with GSO resulted in only a minor performance difference (-0.2 AR), despite requiring less training effort. The entire training process takes approximately 80 GPU hours on an RTX 3090, with a memory consumption of around 20GB.

2. Details of Relative Pose Estimator

The relative pose estimator follows the MRPE [4] framework and is composed of a three-layer CNN, two fully connected (FC) layers, and two separate FC branches for translation and rotation estimation. This network outputs a 3D translation vector and a 6D rotation vector [6].

3. Additional Qualitative Results

We present additional visualizations of the proposed method. Fig. 1 and Fig. 2 illustrate the coarse pose estimation results for the YCB-V [5] and LM-O [2] datasets, respectively, while Fig. 3 and Fig. 4 depict the corresponding pose refinement results. Although the coarse pose estimation stage yields promising results, a comparison between the last columns of Fig. 1 and Fig. 2 with those of Fig. 3 and Fig. 4 demonstrates that the refinement method proposed in this paper further improves pose estimation accuracy.

References

[1] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022.

[2] Alexander Krull, Eric Brachmann, Frank Michel, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Learning analysis-by-synthesis for 6d pose estimation in rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 954–962, 2015.

[3] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022.

[4] Jaewoo Park, Jaeguk Kim, and Nam Ik Cho. Leveraging positional encoding for robust multi-reference-based object 6d pose estimation. *arXiv preprint arXiv:2401.16284*, 2024.

[5] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[6] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019.

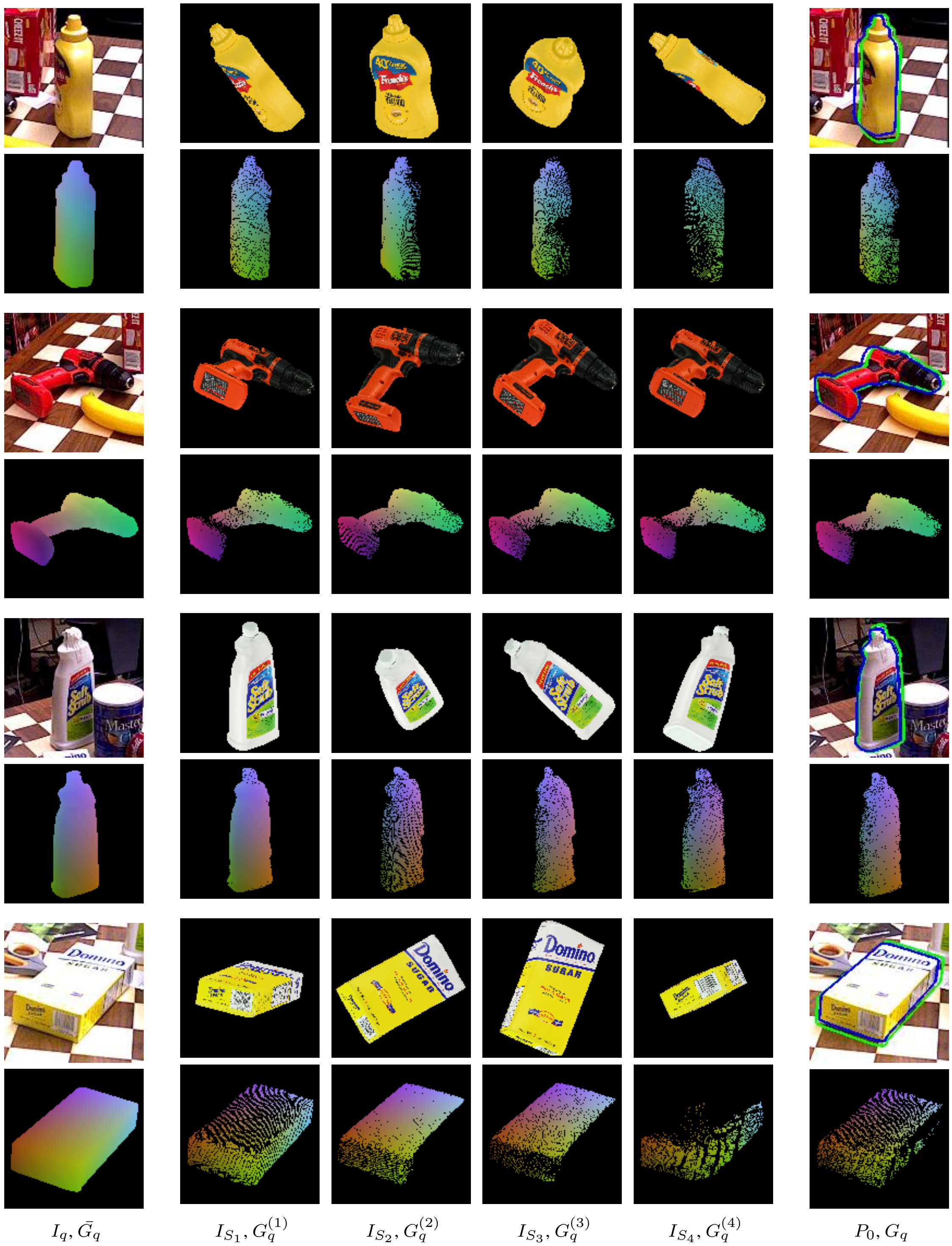
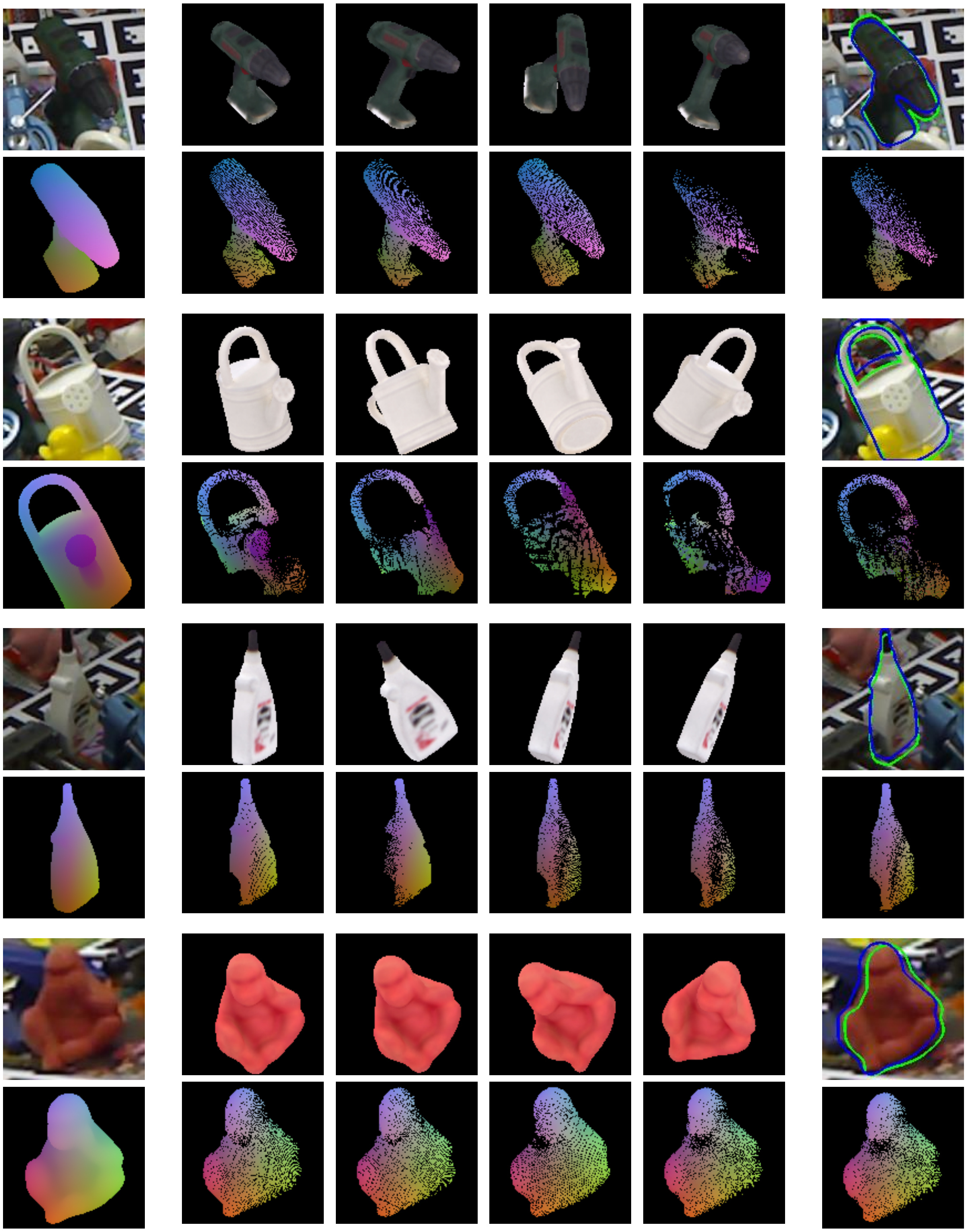
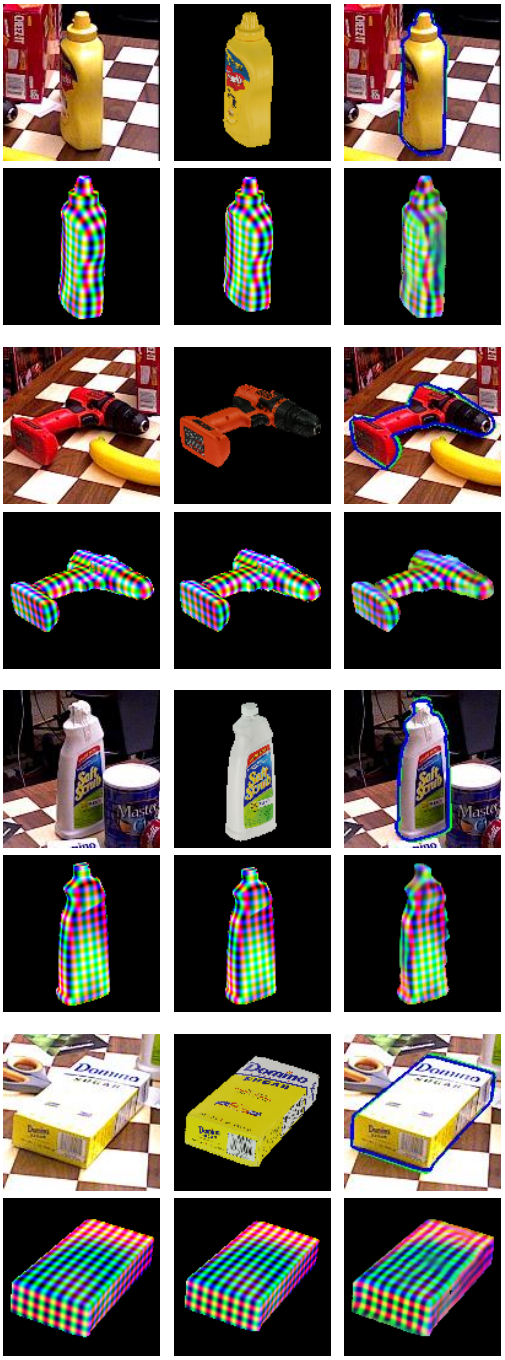


Figure 1. Qualitative results of coarse pose estimation on YCB-V. The first column shows the query image I_q along with the ground truth query geometry $\bar{G}_q$. The second to fifth columns display each template image I_{S_k} from the selected template set $\mathcal{S}$ along with the corresponding estimated geometry candidate $G_q^{(k)}$. In the last column, the image at the bottom illustrates the geometry G_q estimated through pixel-wise voting, and the image at the top shows the estimated initial pose P_0. In the top image, the green contour represents the ground truth pose, while the blue contour represents the estimated pose.



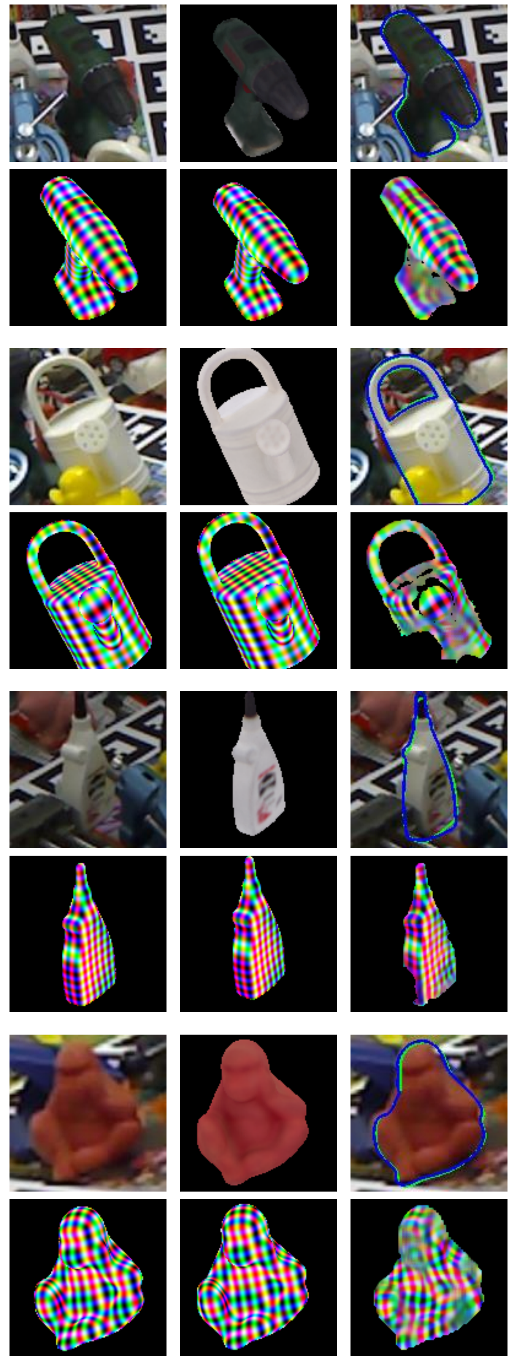
$I_q, \bar{G}_q$ $\quad I_{S_1}, G_q^{(1)}$ $\quad I_{S_2}, G_q^{(2)}$ $\quad I_{S_3}, G_q^{(3)}$ $\quad I_{S_4}, G_q^{(4)}$ $\quad P_0, G_q$

Figure 2. Qualitative results of coarse pose estimation on LM-O.



$I_q, \bar{G}_q^{pos}$ I_r, G_r^{pos} $P_{refined}, G_q^{pos}$

Figure 3. Qualitative results of pose refinement on YCB-V. The first column shows the query image I_q along with the ground truth query geometry $\bar{G}_q^{pos}$. The second column displays the reference image I_r and the reference geometry G_r^{pos}. In the last column, the image at the bottom illustrates the estimated geometry G_q^{pos}, and the image at the top shows the refined pose $P_{refined}$. We visualize G^{pos} for the low-frequency 3 channels.



$I_q, \bar{G}_q^{pos}$ I_r, G_r^{pos} $P_{refined}, G_q^{pos}$

Figure 4. Qualitative results of pose refinement on LM-O.