Supplementary Material for Omni6DPose: A Benchmark and Model for Universal 6D Object Pose Estimation and Tracking

This supplementary material provides additional statistics of Omni6DPose, visualization of SOPE dataset, the details of GenPose++ implementation, and additional experiments and analysis.

A Additional Information of Omni6DPose

In this section, we provide additional statistical details about the Omni6DPose dataset, showcasing its characteristics and diversity. Specifically, it includes the statistics of full object categories, object symmetries, and materials. Additionally, to demonstrate the annotation quality of the ROPE dataset, we visualize the annotated poses of objects.

A.1 Statistics of Full Object Categories

Figure 1 illustrates the frequency of occurrence for all object categories within the Omni6DPose dataset, demonstrating the diversity of object categories covered. This diversity poses new challenges for universal 6D object pose estimation and is conducive to facilitating downstream applications, such as object rearrangement [3] and robot manipulation [1].

A.2 Statistics of Object symmetries

In the domain of 6D object pose estimation, one of the principal challenges is mitigating the ambiguity issue arising from object symmetry. Omni6DPose includes a spectrum of objects characterized by distinct symmetry attributes, broadly classified into three categories: asymmetric objects such as cameras, continuously symmetric objects exemplified by bottles, and discretely symmetric objects, typical examples being boxes. Further delineation within Omni6DPose segregates discretely symmetric objects based on the count of peaks in the distribution of the objects' poses, categorized into Bimodal, 4-peak, 8-peak, and 24-peak classifications. The detailed statistical outcomes are illustrated in the left section of Figure 2. This vast diversity of object symmetries compels the development of new strategies and techniques for precise 6D object pose estimation.

A.3 Statistics of Object Materials

Objects in daily life are made from diverse materials, such as transparent glass mugs and reflective knives. Precise 6D object pose estimation across different

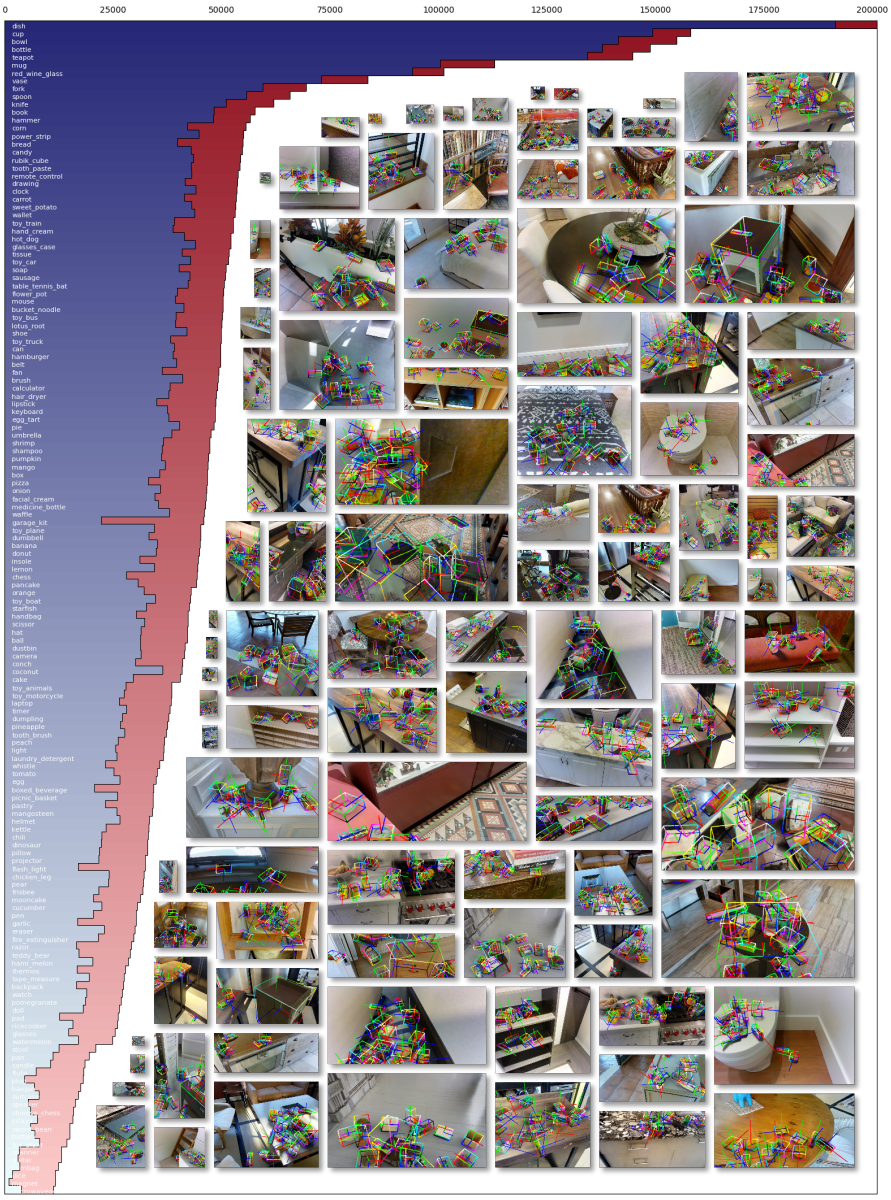

Fig. 1: Frequency of occurrence for all object categories within **Omni6DPose**.

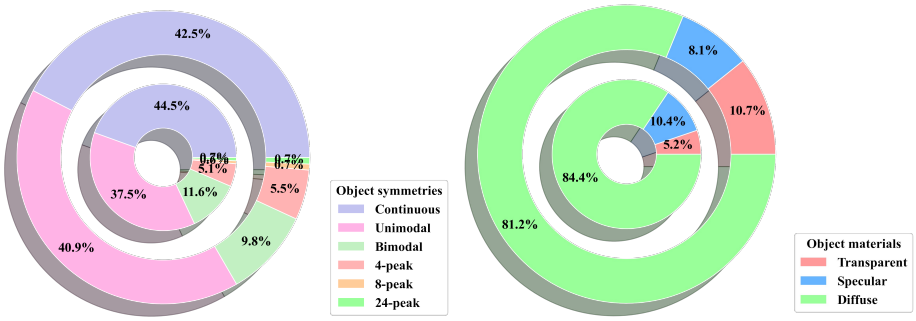

Fig. 2: Statistics of object symmetries and materials within **Omni6DPose**. Inner rings denote **ROPE** statistics and outer rings denote **SOPE**. The left chart categorizes object symmetries into 'continuous' for objects with continuous symmetry, 'unimodal' for objects with no symmetry attributes, and 'bimodal', '4-peak', '8-peak', and '24-peak' for objects with respective counts of discrete symmetry attributes. The right chart details object material distributions: transparent, specular, and diffuse.

materials is crucial for the application of pose estimation in real-world scenarios. Omni6DPose, serving as a comprehensive large-scale 6D object pose dataset, includes a diverse range of materials, categorized into three main types: Diffuse objects, Transparent objects, and Specular objects. The distribution of each material type within the ROPE and SOPE subsets of the dataset is detailed in the right of Figure 2. This variety provides a significant dataset for research into 6D pose estimation of objects with challenging material properties.

A.4 Visualization of ROPE Annotation Quality

To demonstrate the quality of the annotations in the ROPE dataset, we visualize the annotated poses of objects. Figure 3 displays the projected edges of annotated CAD models overlaid onto the corresponding images. These visualizations highlight the precision and accuracy of the pose annotations, underscoring the reliability of our dataset for evaluation.

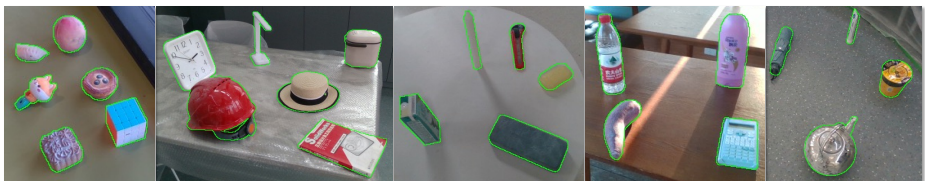

Fig. 3: ROPE annotation quality visualization.

B GenPose++ Implementation Details

B.1 Training Details

In the training phases, both the ScoreNet and Energy models are subjected to training with a batch size of 128, employing the Adam optimizer. The initial learning rate is established at 10^{-3}, subsequently decaying to 10^{-4} to foster optimal convergence. Specifically, ScoreNet is trained for a total of 28 epochs, whereas the Energy model undergoes 25 epochs of training.

B.2 Network Details

In this section, we detail the feature encoder of GenPose++, which processes data from two modalities: RGB and pointcloud. For the RGB modality, we utilize a frozen, pre-trained DINOv2 [2] to extract the semantic features. Specifically, we begin by cropping the object region from the original image based on the object mask and resizing this crop to 224×224 pixels. This resized region is then passed through DINOv2 to produce a feature map of dimensions 16×16. Each feature vector in this map is 384 elements long and represents a 14×14 patch from the original RGB image. To streamline the process, we employ the 'ViT-S/14' variant of DINOv2, which reduces the number of parameters and enhances inference speed. For the pointcloud modality, the object's point cloud is extracted directly using the object mask. We then apply Farthest Point Sampling (FPS) to sample 1024 points and extract global features using pointnet++ [4]. During the feature extraction process for the point cloud, the RGB features are point-wise concatenated onto the corresponding points, integrating data from both modalities to enrich the feature representation.

C Visualization of SOPE

We synthesize 475K frames for training by integrating context-aware mixed reality with physics-based depth sensor simulation. To enhance the generalization capability of SOPE, we systematically apply domain randomization during the data generation process, specifically targeting variations in illumination and object material properties. Considering the relatively lower instance numbers of transparent and reflective objects among all types of objects, we increase their occurrence probability in SOPE. Consequently, Figure 4 exhibits selected examples from the Synthetic Objects in SOPE, showcasing the diversity and realism of the simulated dataset.

D Additional Experiments and Results

In this section, we analyze the necessity of physics-based deep simulation and the distance in feature space between context-aware mixed reality generated RGB

Fig. 4: SOPE dataset visualization. In the figure, bounding boxes are colored according to the coordinates in the object's coordinate system.

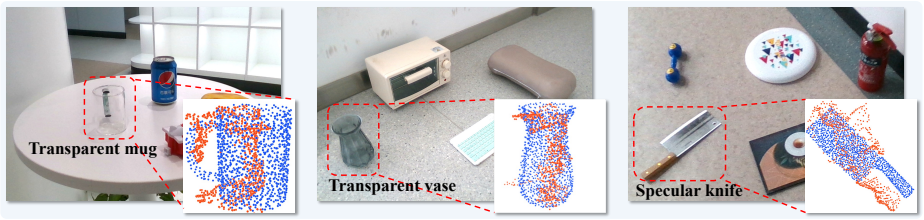

Fig. 5: Visualization of structured-light depth sensor noise on transparent and specular areas. The visualization presents the discrepancy between the ground truth point cloud (blue) and the captured point cloud (red) by the depth sensor. The examples include a transparent mug, a transparent vase, and a specular knife.

images and real images. This elucidates why the SOPE dataset enhances sim-to-real generalization capabilities. Additionally, to demonstrate the robustness of our method to mask predictions, we present the performance of our pose estimation method using SAM segmentation results.

D.1 Physical-based Depth Sensor Simulation.

Structured light-based depth sensors typically introduce noise into the captured depth images, which is particularly pronounced in regions with transparent and reflective objects. This results in a considerable sim-to-real gap when training on perfect synthetic point clouds. Our ablation experiments, as discussed in the main manuscripts, have already established that physics-based depth sensor simulations can significantly bridge the sim-to-real gap. To more vividly demonstrate the divergence between the point clouds captured by the depth sensor and the ideal synthetic ones, Figure 5 shows the depth noise in the transparent and reflective regions from a subset of the ROPE dataset. These visualizations clearly articulate the necessity of physics-based depth sensor simulation.

D.2 Context-Aware Mixed Reality RGB.

Previous synthetic datasets employ rasterization to integrate manually created object models into a real scene, an approach that falls short in terms of overall image fidelity and the realism of individual objects. In contrast, our method leverages ray-tracing and real scanned objects to produce highly realistic imagery. As noted in the main manuscript, the inclusion of RGB information markedly enhances performance. To delve deeper into this, in Figure 6, we showcase the comparison of features extracted using DINOv2 from both synthetic and real RGB images. It demonstrates that the features within the synthetic data set significantly overlap with those in the real data, which bridges the semantic sim-to-real gap.

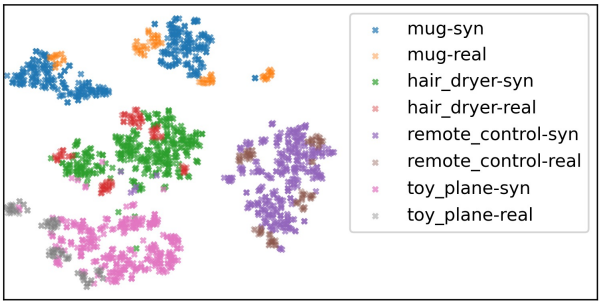

Fig. 6: Visualization of the features of RGB images extracted by DINOv2, reduced to 2D plane using t-SNE.

Table 1: Ablation study of object segmentation mask.

Mask	IoU_{25}	IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm
GT	**39.0**	**19.1**	**2.0**	10.0	15.1	**19.5**	29.4
SAM	38.7	18.9	1.9	**10.3**	**15.6**	**19.5**	**29.8**

D.3 Performance with SAM Segmentation Results.

To demonstrate the robustness of our method to mask predictions, we present the performance of our pose estimation method using SAM segmentation results. The SAM segmentation model generates masks that may not always align perfectly with the ground truth. However, as shown in Table 1 our method maintains high accuracy in pose estimation despite these variations.

References

1. An, B., Geng, Y., Chen, K., Li, X., Dou, Q., Dong, H.: Rgbmanip: Monocular image-based robotic manipulation through active object pose estimation (2023)
2. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
3. Wu, M., Zhong, F., Xia, Y., Dong, H.: Targf: Learning target gradient field to rearrange objects without explicit goal specification. Advances in Neural Information Processing Systems **35**, 31986–31999 (2022)
4. Zhang, J., Wu, M., Dong, H.: Generative category-level object pose estimation via diffusion models. Advances in Neural Information Processing Systems **36** (2024)