ShAPO: Implicit Representations for Multi-Object Shape, Appearance, and Pose Optimization (Supplementary material)

Muhammad Zubair Irshad*[1], Sergey Zakharov*[2], Rares Ambrus[2], Thomas Kollar[2], Zsolt Kira[1], and Adrien Gaidon[2]
* denotes equal contribution

[1] Georgia Institute of Technology
[2] Toyota Research Institute
[1]{mirshad7, zkira}@gatech.edu,[2]{first.last}@tri.global

1 Appendix A: Result Visualization

Here we provide more visual qualitative result of superior single-view multi-object *Shape Reconstruction*, *6D pose and size estimation* and *Appearance Reconstruction* done using our technique, **ShaPO**. Our method shows very promising results for superior 6D pose and size estimation compared to the strong baseline NOCS [9] (Figure 1). Our network also performs more accurate shape and texture reconstruction compared to the strong-baseline, CenterSnap [2] (Figure 3), which only performs shape reconstruction (i.e. meshes obtained through surface reconstruction of coarse pointcloud predictions i.e. 2048 points). We also visualize the improved pose estimation performance of our method after inference-time optimization (Figure 2). Figure 5 also shows zero-shot generalization results on HSR robot i.e. no-retraining was done.

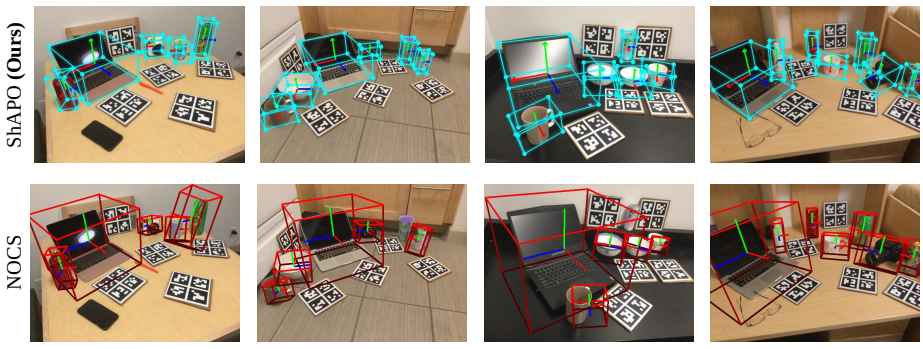

Fig. 1: **ShAPO Qualitative 6D pose estimation comparison with NOCS:** Our method's 6D pose estimation in comparison to the best pose estimation configuration i.e. 32-bin classification on the NOCS dataset. We show accurate 3D bounding box predictions and 6D pose and size estimation of multiple novel object categories than the strong baseline.

Table 1: **Texture quality ablation.** We compare texture quality using the PSNR metric between three modalities: network prediction, optimization, and fine-tuning of the t_θ network.

	Inference	Optimization	Fine-tuning
PSNR	11.41	20.64	24.32

2 Appendix B: Texture Quality Ablation on NOCS Real275

In this section, we provide an ablation on the output texture quality on NOCS Real275 test-scenes. In particular, we compare the direct network texture prediction with the result after our differentiable optimization, and the result after our differentiable optimization with additional fine-tuning of the t_θ network weights. We use the learning rate of 10^{-5} for the weight fine-tuning. Table 1 demonstrates that our optimization procedure almost doubles the texture quality in terms of PSNR. Additional fine-tuning of the network weights allows us to improve texture reconstruction results even further. For qualitative results see Figure 4.

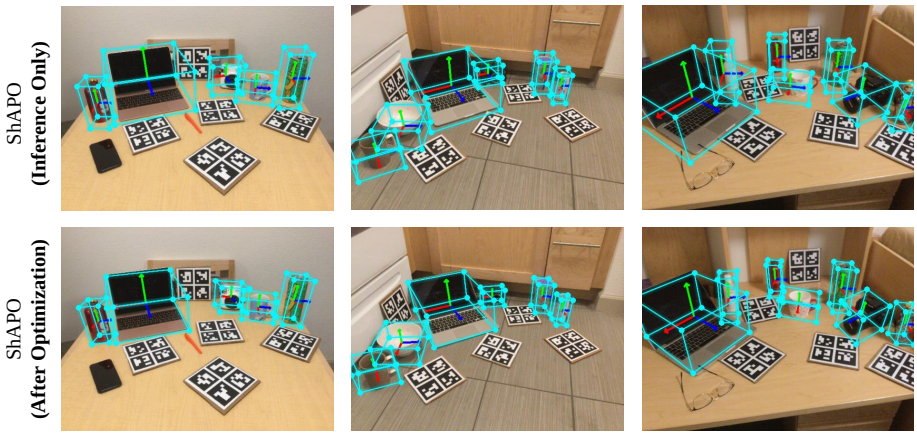

Fig. 2: **ShAPO Qualitative Comparison of 6D pose and size Inference and Optimization:** Our method's 6D pose and size comparison shown on 3 novel scenes in NOCS Real275 test-set. After optimization, our method predicts accurate bounding boxes as shown by the bottom row in the figure.

3 Appendix C: Network Architecture details and Training

Our backbone is implemented as Feature pyramid network [3] with takes as input Resnet [1] outputs at various spatial resolutions and adds lateral connections

Fig. 3: **ShAPO Qualitative Reconstruction Comparison with CenterSnap [2]:** The figure qualitatively shows the superior reconstruction performance of our method with the strong state of art i.e. CenterSnap [2] on novel scene in NOCS Real275 test-set. Our method produces finer reconstruction surfaces both in terms of shape accuracy and textures with details such as mug-handle and camera lens.

Fig. 4: **ShAPO Qualitative Inference, Optimization and Finetuning Comparison:** The figure qualitatively shows the inference, latent-only optimization and latent with appearance network optimization. Note that as noted earlier, we let the appearance network weights to change to allow for finer level of reconstruction.

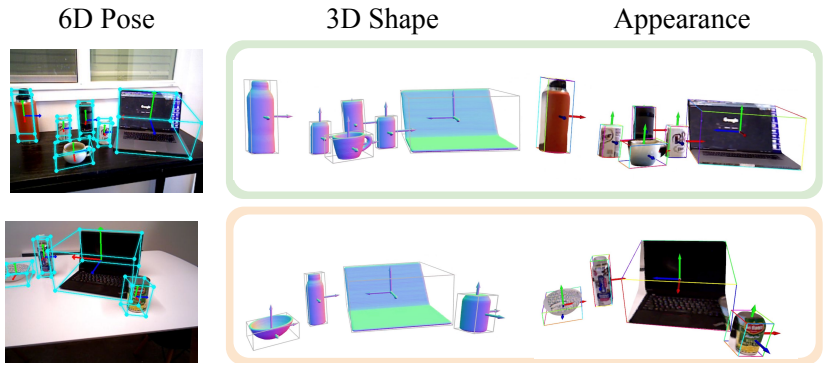

Fig. 5: **Real-world generalization experiments on HSR robot**

with a top-down pathway. Each of our specialized heads comprises of a series of convolution layers (i.e. up-sampling stages), as described in [4] with the final prediction layer comprising of a 1×1 convolution and $4x$ bilinear up-sampling. We train the combined backbone and heads network for 30 epochs with early stopping based on the performance on the validation set. We use a learning rate of $6e^{-4}$ and a polynomial weight decay with a co-efficient of $1e^{-4}$. Our texture network t_θ is a Siren-based [8] 6-layer MLP consisting of 512-dimensional hidden layers and with ω_0 set to 128. Siren networks demonstrate superior results at representing fine details when compared to standard ReLU-based MLPs thanks to the used periodic activation functions. After we train the shape MLP (G) and texture MLP (t_θ), we freeze the networks for single-shot supervision at the Gaussian center locations. During inference, we use the frozen networks (G) and (t_θ) to optimize for shape, pose, size and appearance latent codes.

4 Appendix D: Related works

Our method, ShAPO, relates to multiple key areas in 3D scene reconstruction, object understanding and pose estimation. In essence, we provide more qualitative and quantitative comparison to four related works i.e. CenterSnap [2], MeshSDF [7], Occ-Nets [5] and TextureFieldd [6].

CenterSnap [2]: In particular, Figure 3 qualitatively shows the superior reconstruction quality of our method compared to the strong state of the art i.e. CenterSnap [2]. Furthermore, Our shape representation (implicit SDF vs point-clouds in CS), addition of texture network and texture codes, differen- tiable iso-surface extraction, optimization and joint shape, appearance, and textures warping all make our work sig- nificantly different from CS. While our work does share a common backbone with CS, being able to leverage the test- time observations to warp the latent shape, appearance, and poses of the model beyond network inference is precisely our contribution. Hence, our technique is able to model large intra-class variations (25.4% and 7.1% absolute improvement in 6D pose) over CS (cf. Tbl 2 in the main text).

Table 2: **Quantitative Comparison with MeshSDF** We compare the computational time to extract surface for our method, ShAPO, in comaprison to MeshSDF [7]

# Points	704	3228	13023	56041	224680
MeshSDF	0.017 s	0.032 s	0.091 s	0.654 s	4.396 s
ShaPO	0.010 s	0.013 s	0.016 s	0.025 s	0.093 s

MeshSDF [7] pro- poses a solution to extracting surface meshes while preserving end-to-end differentiability. We instead extract dense surface point clouds using our octree-based sampling abstaining from the expensive Marching Cubes computation at every optimization step. As shown in the Table 2, MeshSDF's implementation doesn't scale well to higher resolutions, whereas our technique does, achieving accurate fine-grained reconstruction with minimal runtime. Additionally, our method also supports appearance optimization.

Occupancy Networks [5]: We extract the object's surface using a **differentiable 0-isosurface projection** which is a crucial component that allows us to perform shape/pose/appearance optimization. Conversely, the procedure from Occupancy Networks [5] is applied once to extract the object's surface using non-differentiable Marching Cubes.

Texture Fields (TF) [6] trains a single network per category making it difficult to model a large number of categories. We model multiple categories through our novel texture code ($\mathbf{z}_{tex}$) unique to each object in our database of shape and texture priors using a single network (cf. supplementary video). Second, TF does not consider test-time optimization, whereas we propose a novel test-time warping of textures by updating $\mathbf{z}_{tex}$ to fit unseen object appearances. Lastly, TF reconstructs one model per image whereas we infer multiple objects from a single-view RGBD.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
2. Irshad, M.Z., Kollar, T., Laskey, M., Stone, K., Kira, Z.: Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. In: IEEE International Conference on Robotics and Automation (ICRA) (2022), https://arxiv.org/abs/2203.01929
3. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6399–6408 (2019)
4. Kirillov, A., He, K., Girshick, R., Rother, C., Dollar, P.: Panoptic segmentation. In: CVPR (June 2019)
5. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: CVPR (2019)
6. Oechsle, M., Mescheder, L., Niemeyer, M., Strauss, T., Geiger, A.: Texture fields: Learning texture representations in function space. In: Proceedings IEEE International Conf. on Computer Vision (ICCV) (2019)
7. Remelli, E., Lukoianov, A., Richter, S., Guillard, B., Bagautdinov, T., Baque, P., Fua, P.: Meshsdf: Differentiable iso-surface extraction. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 22468–22478. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/file/fe40fb944ee700392ed51bfe84dd4e3d-Paper.pdf
8. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. NeurIPS (2020)
9. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: CVPR (2019)