ShAPO: Implicit Representations for Multi-Object Shape, Appearance and Pose Optimization

Muhammad Zubair Irshad Sergey Zakharov* Rares Ambrus Thomas Kollar Zsolt Kira Adrien Gaidon*

TOYOTA RESEARCH INSTITUTE

Georgia Tech

Motivation

- Real2Sim Asset Creation from a single-view RGB-D
- Object-centric holistic 3D scene understanding pipeline
- Recovers 3D shape, 6D pose and sizes and appearance of multiple novel objects
- No CAD models or explicit 3D input required
- Applications: Object Identification, Instance Tracking, Real2Sim Asset Creation,

★ **Input**

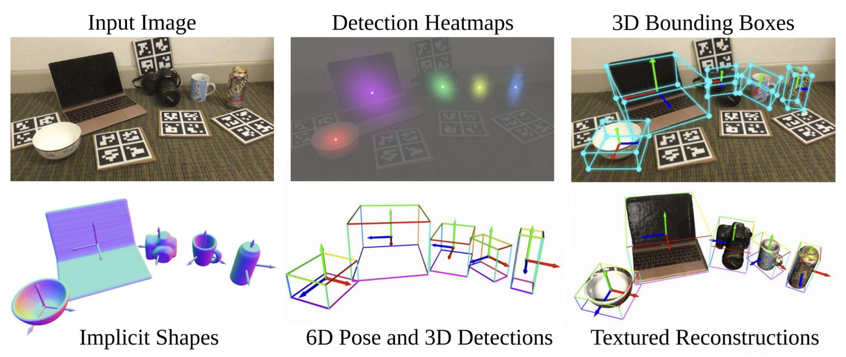
$I \in \mathbb{R}^{h_o \times w_o \times 3}, D \in \mathbb{R}^{h_o \times w_o}$

★ **Predict**

$\hat{\mathcal{P}} \in SE(3), \hat{\mathbf{s}} \in \mathbb{R}^3, \hat{\mathrm{M}} \in \mathbb{R}^{h_o \times w_o}$

$\boldsymbol{G}(x, z_{sdf}) = s : z_{sdf} \in \mathbb{R}^{64}$

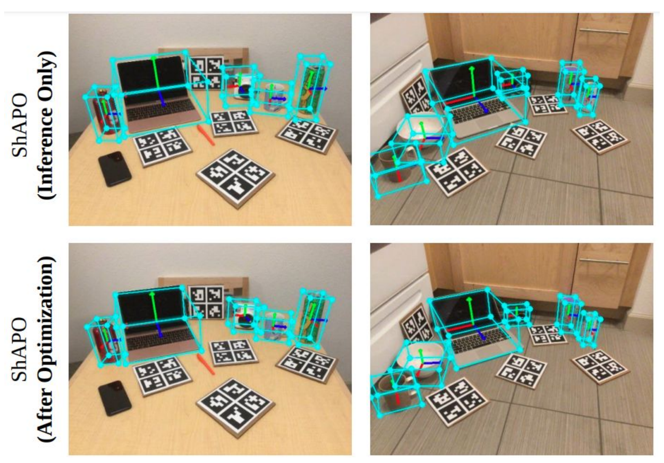
$\boldsymbol{t}_\theta(x, z_{sdf}, z_{tex}) = c, z_{tex} \in \mathbb{R}^{64}$



Overview

★ **Prior Works...**

- Not Scalable/Holistic
- Low performance in challenging scenarios
- Shape representation sample inefficient



★ **Contributions**

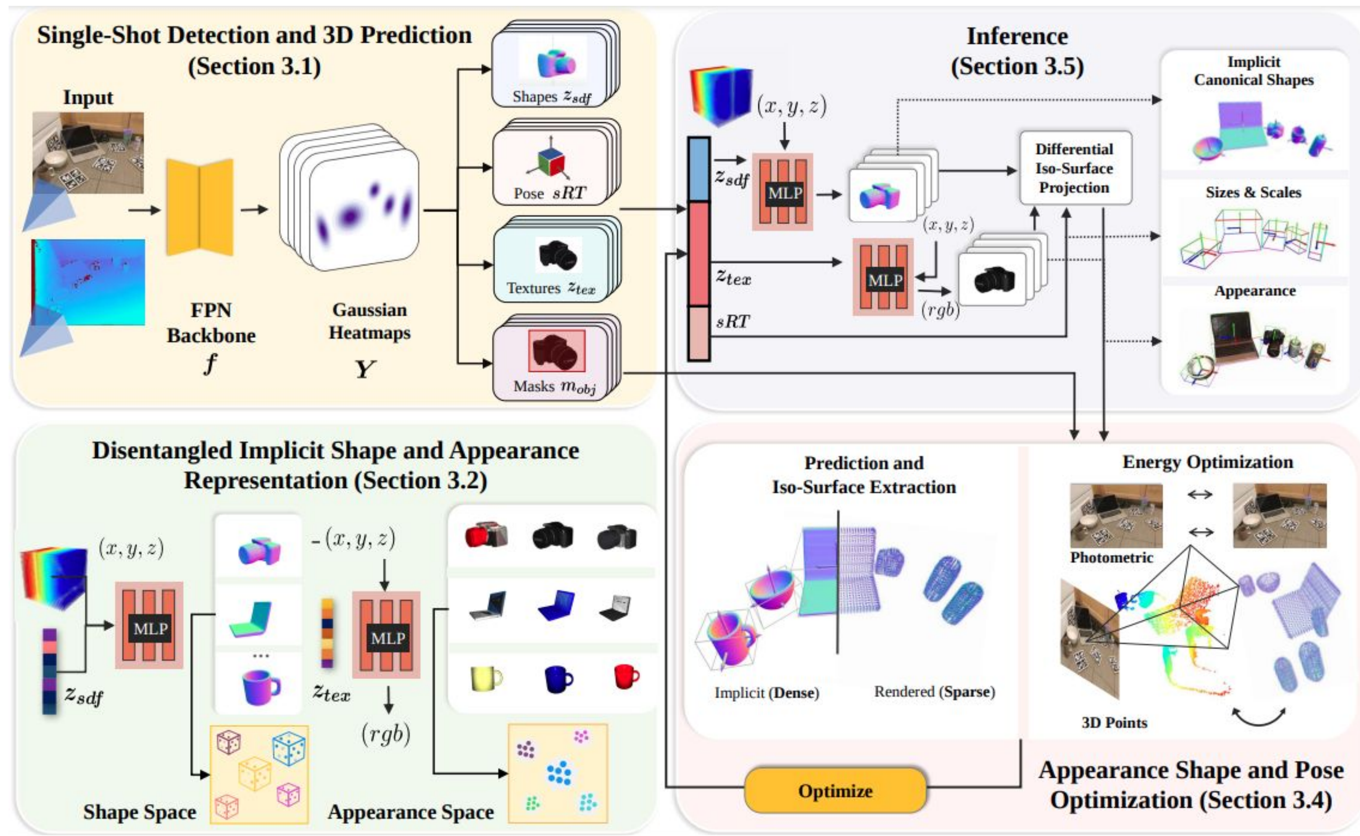
- Object-centric holistic scene-understanding
- Employ a joint **implicit textured shape-prior** to learn from a large collection of CAD models
- Fast **octree-based differentiable** optimization
- Over **8% improvement** in mAP for 6D pose
- 3D object understanding without needing 3D models

Single-Shot Prediction

★ **We employ an two-stage approach**

1. A single-shot network to predict 3D shape, pose and size codes along with segmentation masks in a per-pixel manner
2. Test-time optimization of joint shape, pose and size codes given a single-view RGB-D observation of a new instance

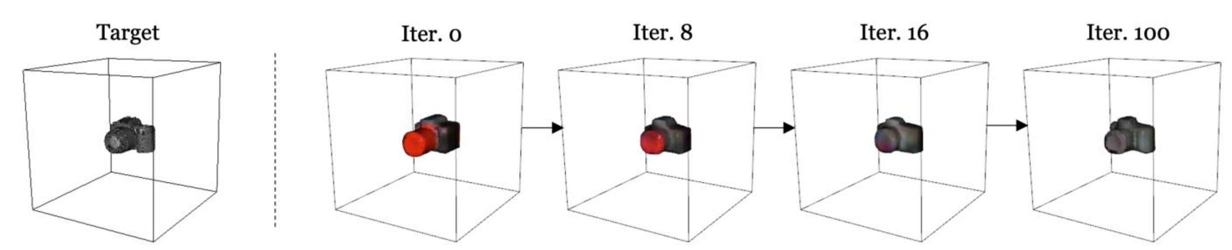
$$\mathcal{L} = \lambda_{inst}\mathcal{L}_{inst} + \lambda_{sdf}\mathcal{L}_{sdf} + \lambda_{tex}\mathcal{L}_{tex} + \lambda_M\mathcal{L}_M + \lambda_P\mathcal{L}_P$$



Optimization using Priors

★ **Shape, Pose, Size and Appearance Codes**

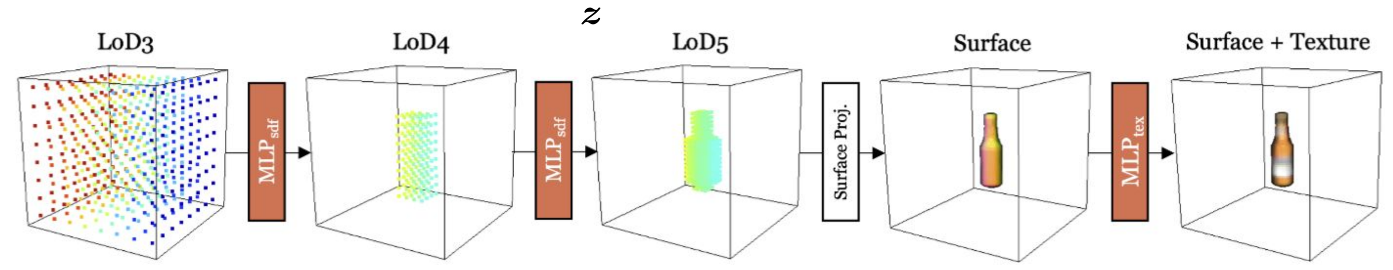
- Joint implicit textured representation
- Learn from a large variety of CAD models (~100 ShapeNet Models)
- Shape (SDF MLP), Texture (Siren MLP[2])



★ **Octree-based differentiable optimization**

- Octree-based Point Sampling (coarse-to-fine sampling)
- LOD 3 to LOD 8 (2-3x faster, 1.5x more memory efficient)
- Maximum-a-posteriori estimation to update latent codes

$$\boldsymbol{z}_{sdf} = \arg\min_{\boldsymbol{z}} (\mathcal{L}(D(G(\boldsymbol{z}, \boldsymbol{x})), P_d))$$

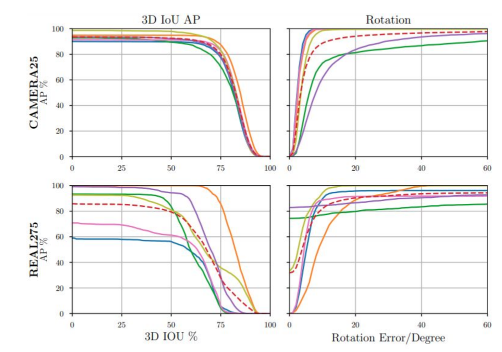


Evaluation

★ **Metrics:** $\boldsymbol{IOU}25, \boldsymbol{IOU}50, 5°5$ **cm**, $5°10$ **cm** and $10°10$ **cm**

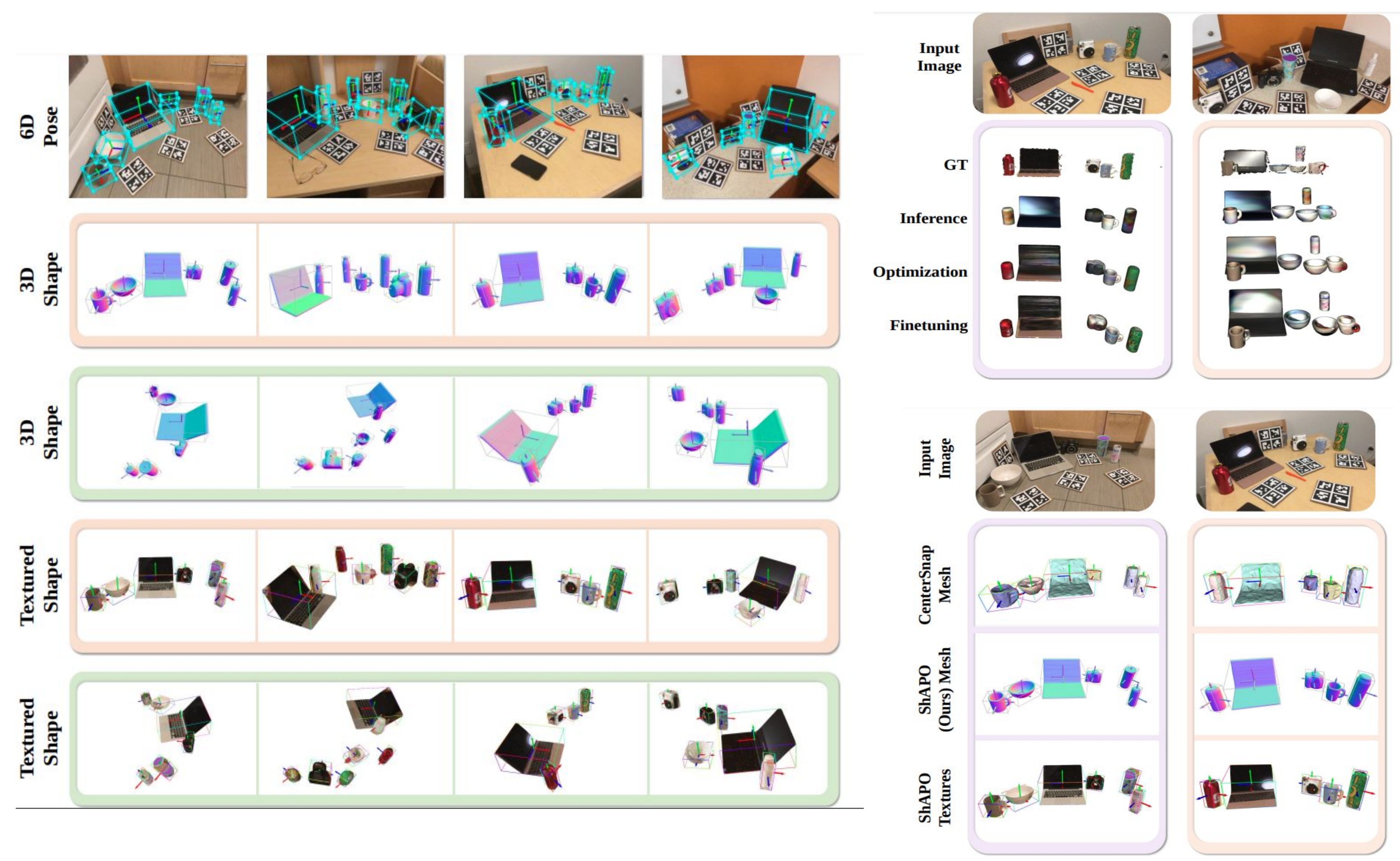
- Test on NOCS Real275, 6 novel scenes, 2750 images
- ShAPO demonstrates an absolute improvement of 1.8%, 25.4% and 7.1% on NOCS Real275 on state of the art baselines[1]
- Appearance optimization improves PSNR by ~78%

	CAMERA25						REAL275					
Method	IOU25	IOU50	5°5 cm	5°10 cm	10°5 cm	10°10 cm	IOU25	IOU50	5°5 cm	5°10 cm	10°5 cm	10°10 cm
1 NOCS [41]	91.1	83.9	40.9	38.6	64.6	65.1	84.8	78.0	10.0	9.8	25.2	25.8
2 Synthesis* [3]	-	-	-	-	-	-	-	-	0.9	1.4	2.4	5.5
3 Metric Scale [23]	93.8	90.7	20.2	28.2	55.4	58.9	81.6	68.1	5.3	5.5	24.7	26.5
4 ShapePrior [37]	81.6	72.4	59.0	59.6	81.0	81.3	81.2	77.3	21.4	21.4	54.1	54.1
5 CASS [2]	-	-	-	-	-	-	84.2	77.7	23.5	23.8	58.0	58.3
6 CenterSnap [15]	93.2	92.3	63.0	69.5	79.5	87.9	83.5	80.2	27.2	29.2	58.8	64.4
7 CenterSnap-R [15]	93.2	92.5	66.2	71.7	81.3	87.9	83.5	**80.2**	29.1	31.6	64.3	70.9
8 ShAPO (Ours)	**94.5**	**93.5**	**66.6**	**75.9**	**81.9**	**89.2**	**85.3**	79.0	**48.8**	**57.0**	**66.8**	**78.0**



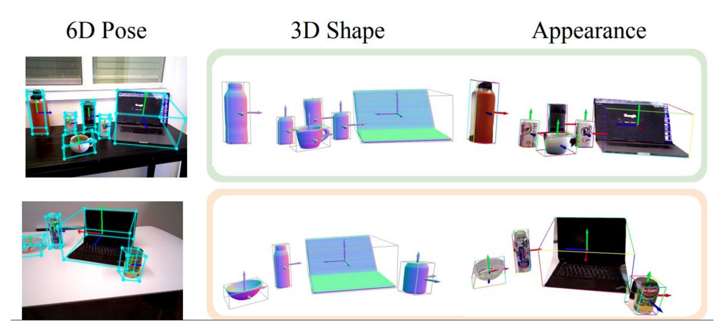
Qualitative Results

★ **Shape and Appearance Reconstruction & pose estimation**



★ **3D only Optimization**

Grid type	Resolution	Point Sampling		Efficiency (per object)		Reconstruction	
		Input	Output	Time (s)	Memory (MB)	Shape (CD)	Texture (PSNR)
Ordinary	40	64000	412	10.96	3994	0.30	10.08
	50	125000	835	18.78	5570	0.19	12.83
	60	216000	1400	30.51	7850	0.33	**19.52**
OctGrid	LoD5	1521	704	**5.53**	**2376**	0.19	9.27
	LoD6	5192	3228	6.88	2880	**0.18**	13.63
	LoD7	20246	13023	12.29	5848	0.24	16.14



ECCV TEL AVIV 2022

Available Material

Project Webpage: https://zubair-irshad.github.io/projects/ShAPO.html

Youtube: https://www.youtube.com/watch?v=LMg7NDcLDcA

References

[1] 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," ICRA, 202

[2] Sitzmann, V., Martel, J., Bergman, A., Lindell, D., & Wetzstein, G. (2020). "Implicit neural representations with periodic activation functions", Neurips, 2020