Learning Object-Centric Neural 3D Scene Representations

Muhammad Zubair Irshad

W. https://zubairirshad.com

E. muhammadzubairirshad@gmail.com

T. @mzubairirshad

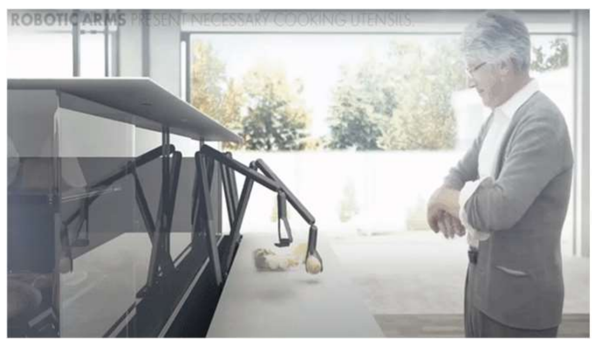
Learning Object-Centric Neural 3D Scene Representations

Goal: Build Generalizable 3D representation of objects useful for a variety of downstream applications

Approach: Learning with Structured Inductive Bias and Priors

Real-World Robotics

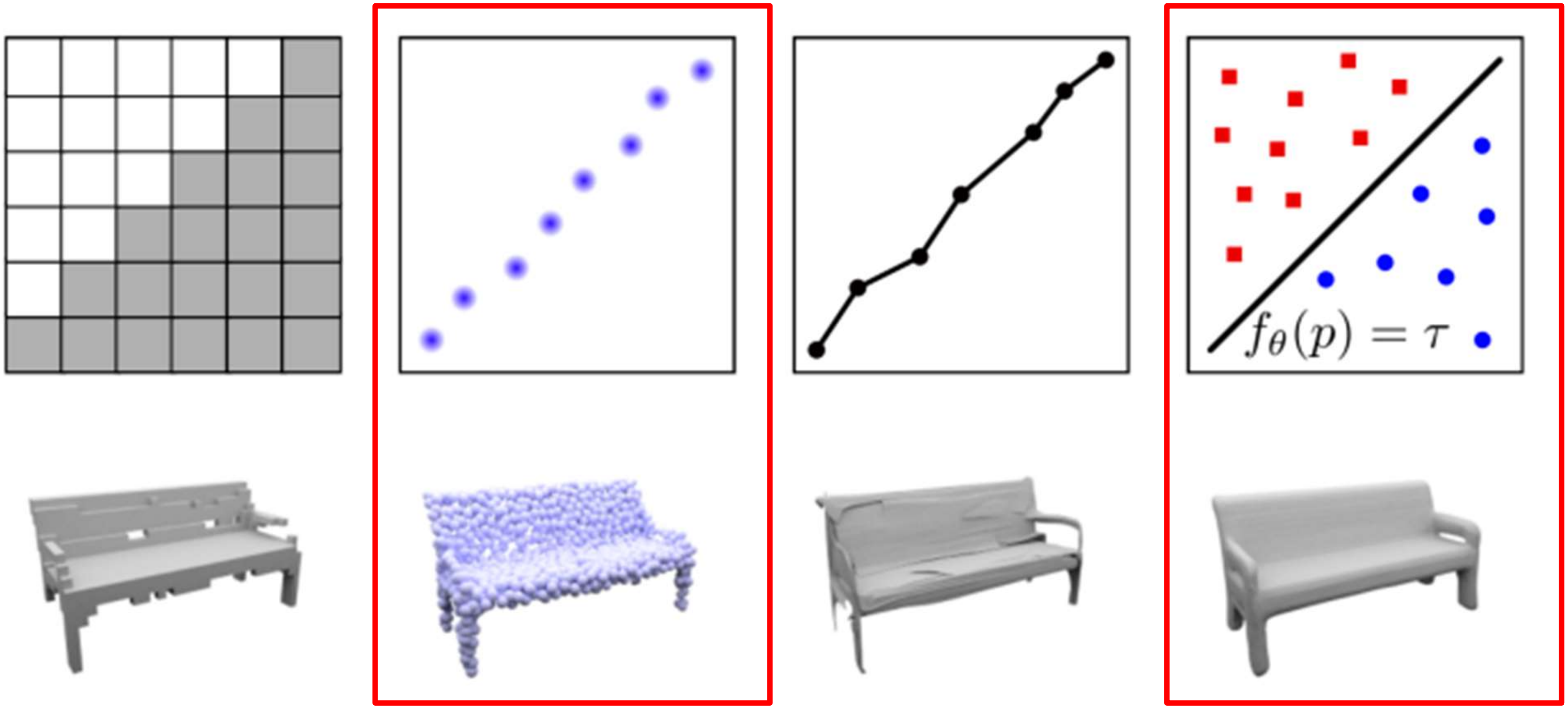
Generalizable Autonomy



Fleet Learning

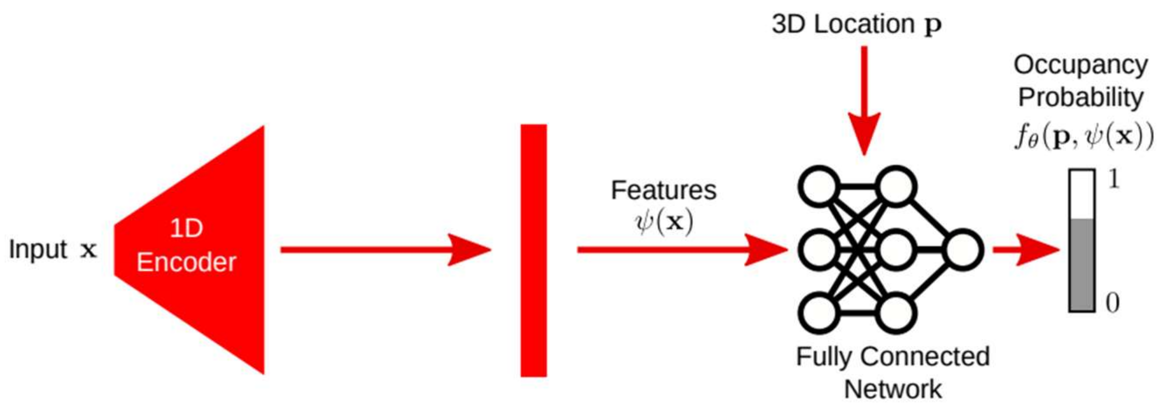
Credits: Sony AI Cooking, Netflix

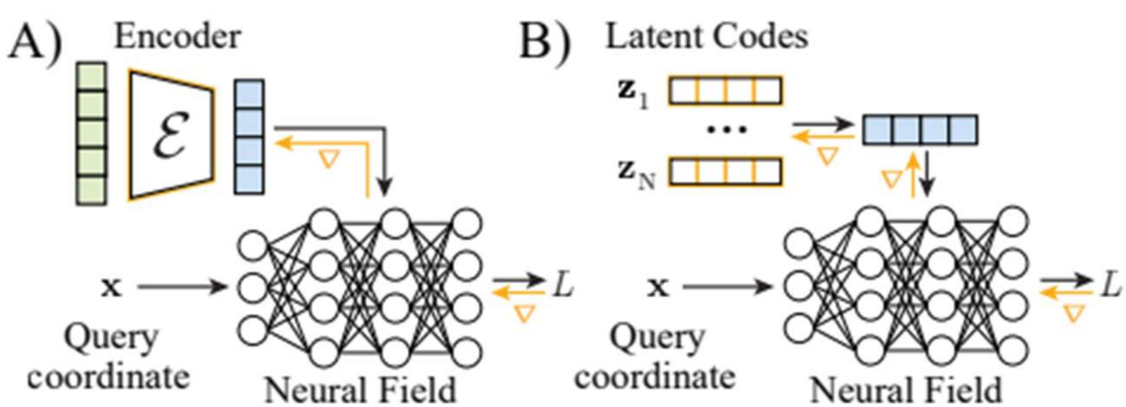
Perception for 3D Object Understanding: Shape Representations



Andrieas Giger Implicit Neural Representations: From Objects to 3D Scene, June 2020

Perception for 3D Object Understanding: Shape Representations

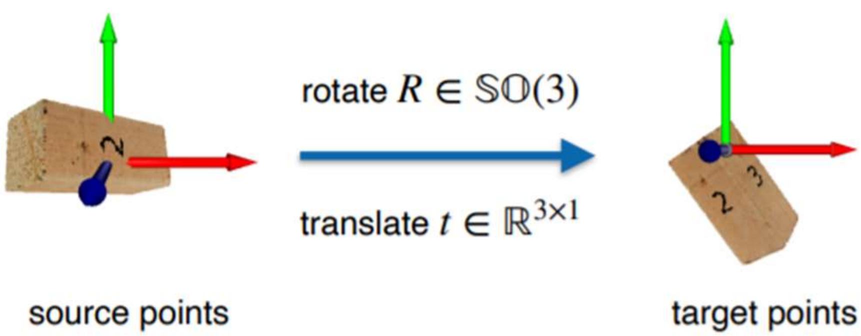




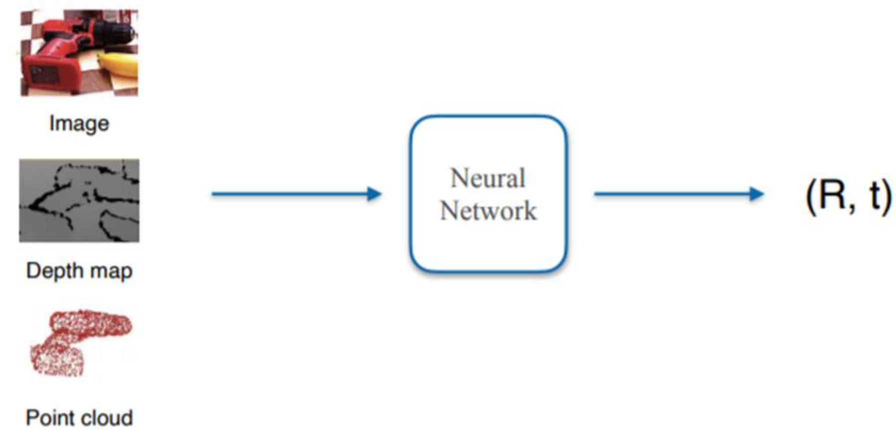
Andrieas Giger Implicit Neural Representations: From Objects to 3D Scene, June 2020
Neural Fields in Visual Computing and Beyond, Eurographics 2022

Perception for 3D Object Understanding: 6D Object Pose Estimation

6D Pose Estimation

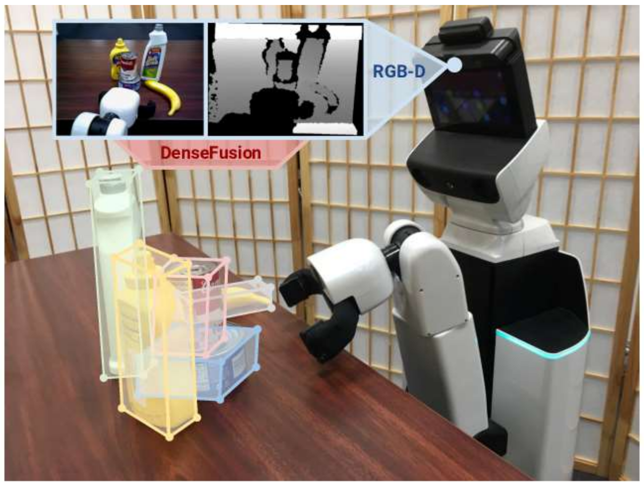


Learning 6D Pose



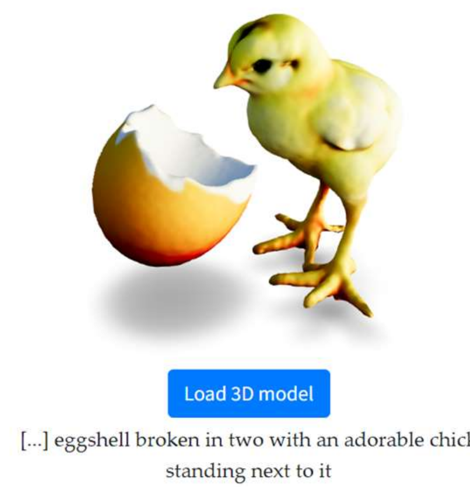
[Machine Learning Meets Geometry, Winter 2021, He Wang 2019 CVPR]

Perception for 3D Object Understanding: Applications



Object Grasping

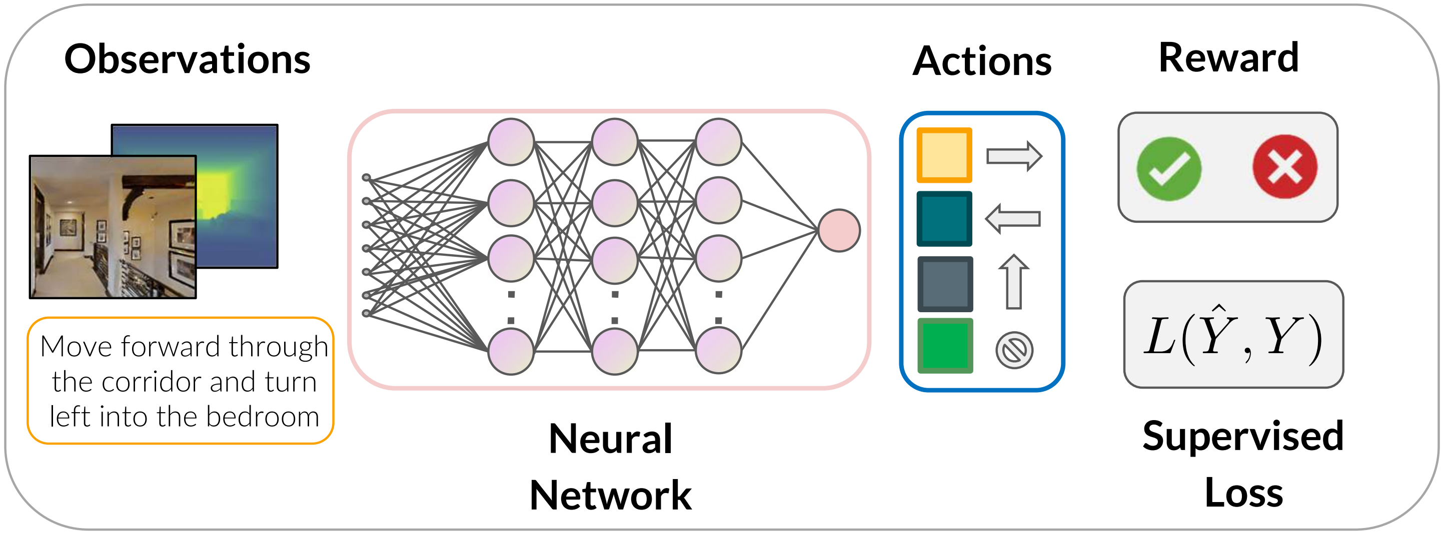
AR/VR Augmentations



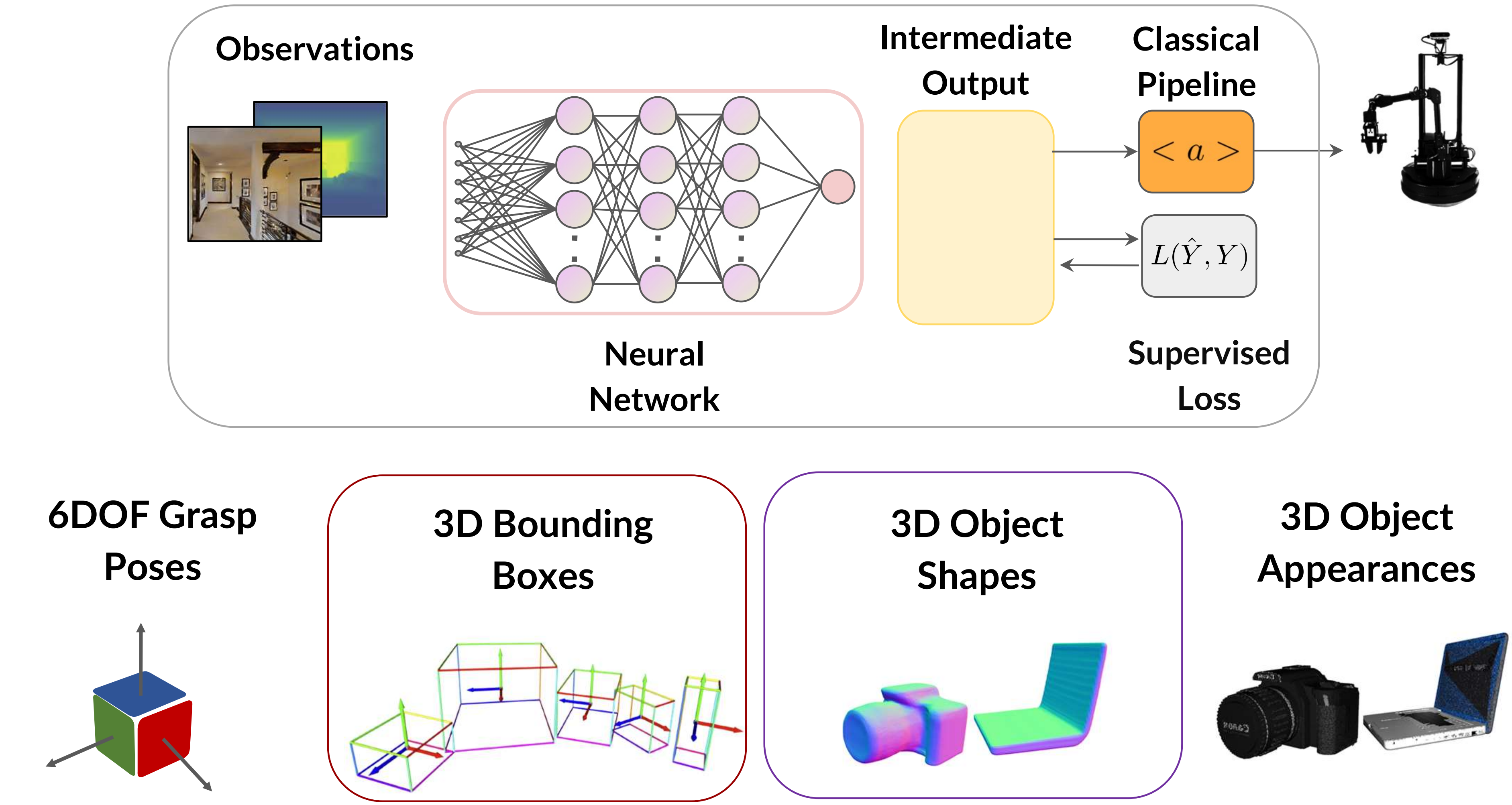
Text-to-3D

DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion, CVPR'2019
Towards Monocular 6D Object Pose Estimation, Thesis 2019, Fabian Manhardt
Dreamfusion, CVPR 2022

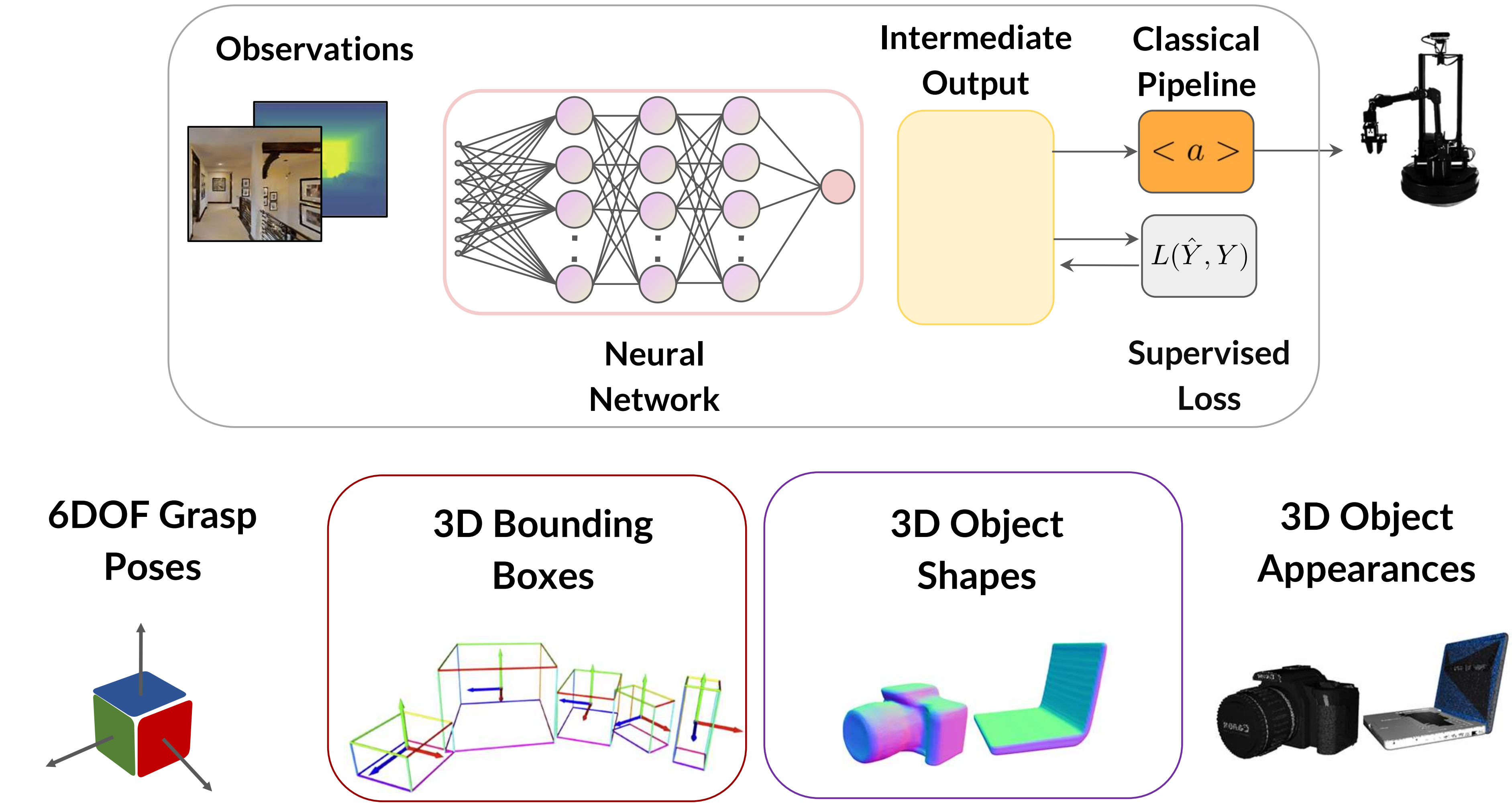
Perception for 3D Object Understanding: Current Paradigm



Perception for 3D Object Understanding: Current Paradigm

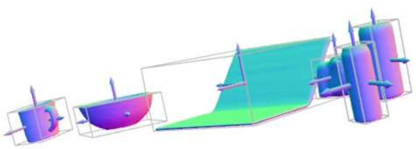


Perception for 3D Object Understanding: Current Paradigm

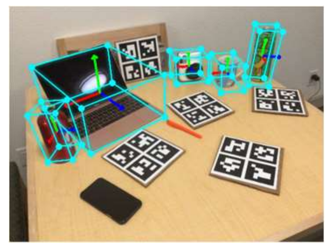


Perception for 3D Object Understanding: Proposed Work

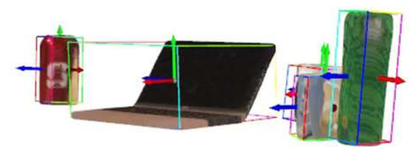
Input



3D Shape



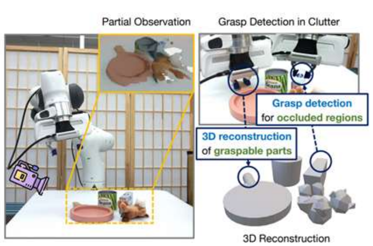
6D pose and size



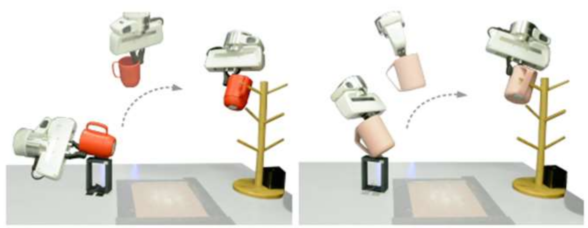
Appearance

holistic category-level
3D object understanding

Applications



Robotics Grasping [GIGA RSS'21]



Category-level Manipulation [NDF'21]

Asset Creation [HHBD ICCV'19]

Perception for 3D Object Understanding: Current Paradigm

Key highlights (Our proposed):

- \+ Anchor-free
- \+ Joint shape reconstruction and object-centric scene context
- \+ Fast (Real-time) reconstruction
- \+ Category agnostic reconstruction and 6D pose and size estimation
- \+ Single-forward pass for entire network
- \+ All heads share the same level of expertise i.e., gradient sharing

(b) Disjoint Shape Reconstruction and Pose and Size Estimation

Key highlights (Prior Methods):

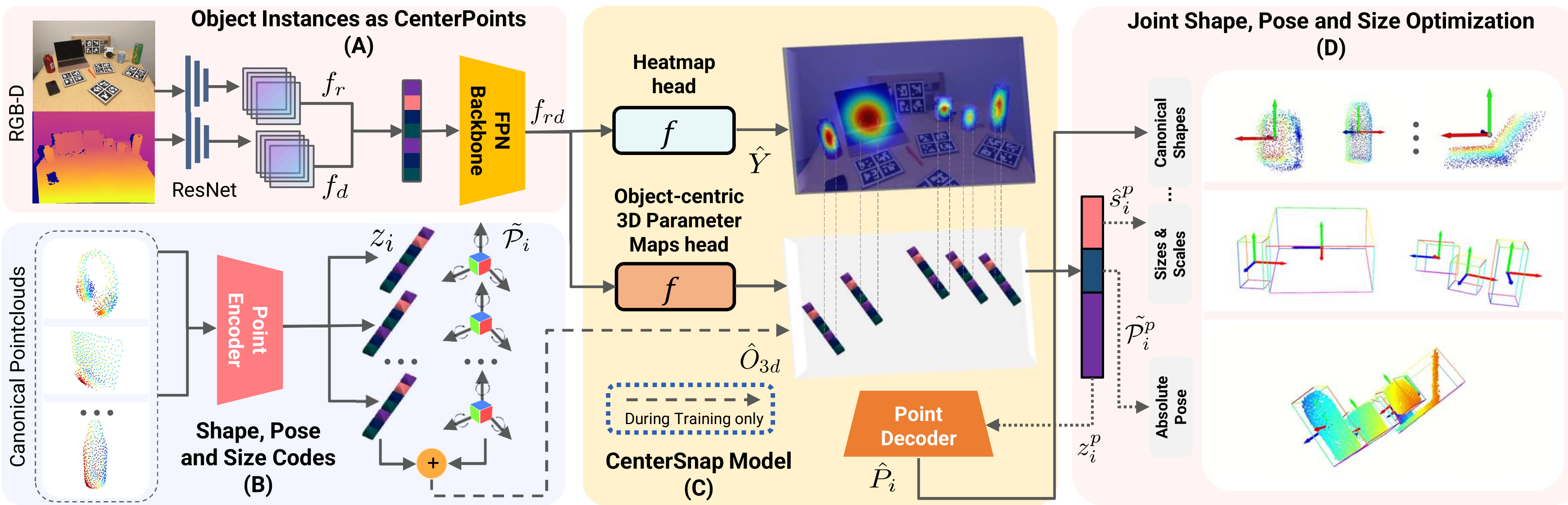
- \- Anchor-Based
- \- Disjoint shape reconstruction and object-centric scene context
- \- Slow reconstruction
- \- Category-specific reconstruction and 6D pose and size estimation
- \- Multiple forward passes for each task
- \- Heads can be at different level of expertise

(c) Joint Detection, Reconstruction and Pose Estimation

Perception for 3D Object Understanding: Our Approach

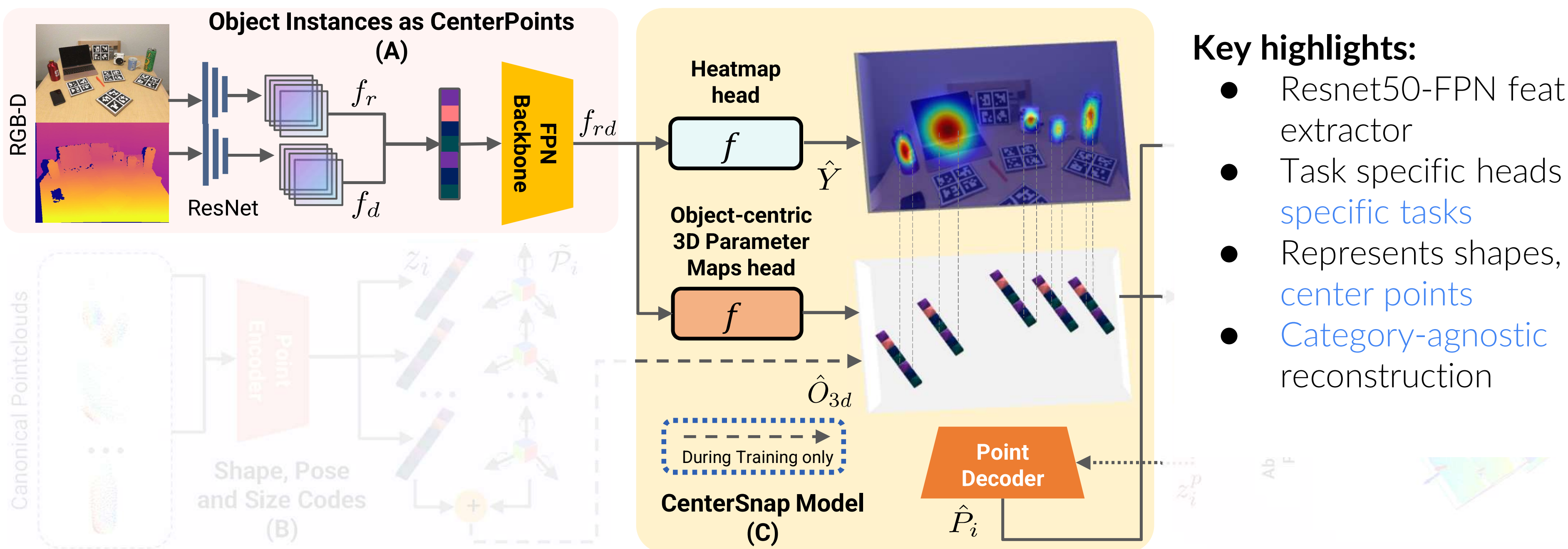
*"..Train **intelligent** perception system capable of utilizing **geometry prior** for **efficient (real-time)** shape reconstruction and 6D pose estimation of multiple objects"*

CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation



[Ref] **M.Z.Irshad**, T.Kollar, M.Laskey, K.Stone, Z.Kira, " CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and Categorical 6D Pose and Size Estimation, **ICRA 2022**

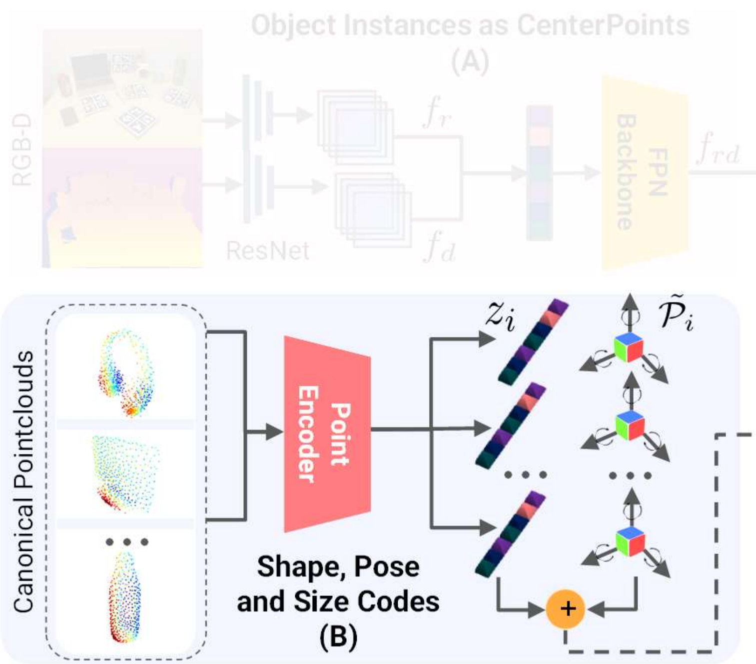
CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation



Key highlights:

- Resnet50-FPN feature extractor
- Task specific heads for specific tasks
- Represents shapes, poses as center points
- Category-agnostic reconstruction

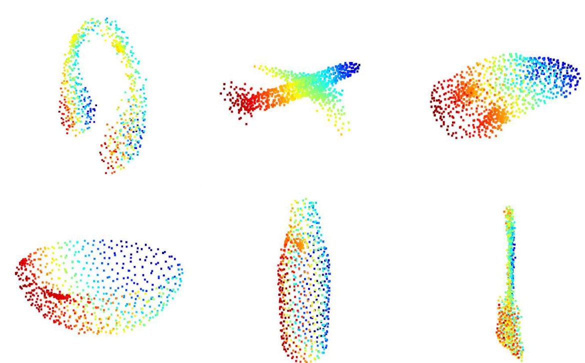
CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation

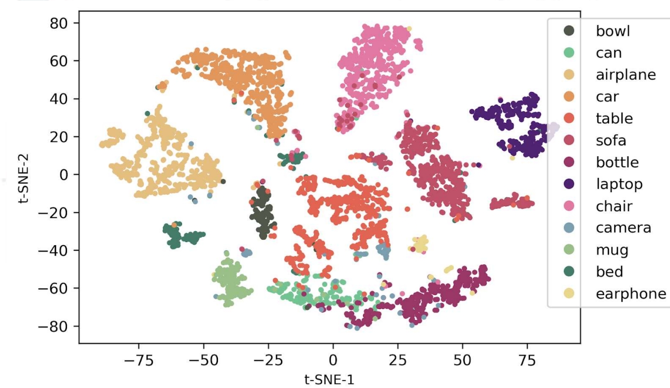


Key highlights:

- Unique shape code for each object
- Strong geometry prior from shape net 3D models
- Conv De-conv Neural Network as Auto-Encoder
- Category-agnostic reconstruction

$$D_{cd}(\mathbf{P}_i, \hat{\mathbf{P}}_i) = \frac{1}{|\mathbf{P}_i|} \sum_{x \in \mathbf{P}_i} \min_{y \in \hat{\mathbf{P}}_i} \|x - y\|_2^2 + \frac{1}{|\hat{\mathbf{P}}_i|} \sum_{\mathbf{y} \in \hat{\mathbf{P}}_i} \min_{x \in \mathbf{P}_i} \|x - y\|_2^2$$





CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation

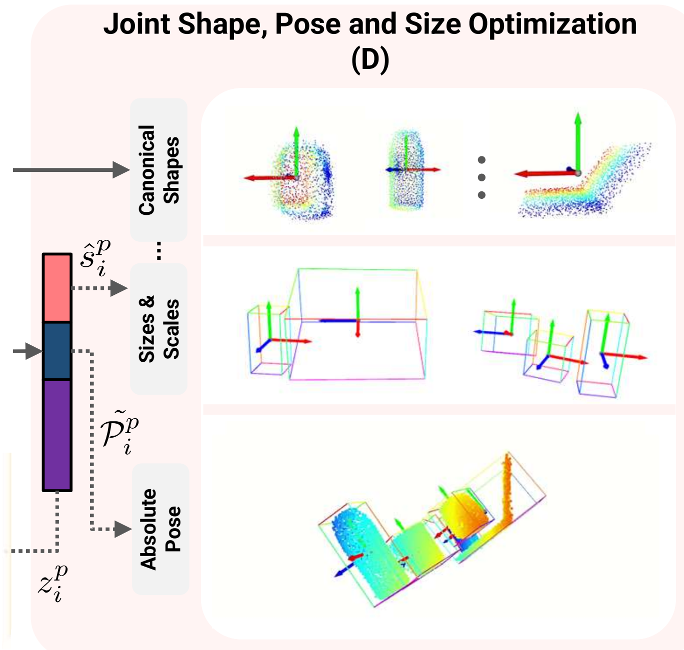
Key highlights:

- Single-forward pass inference
- Optimized jointly
- Maksed L1 Loss for Object Parameter Map
- Huber Loss for Heatmap
- Symmetry consideration for symmetric objects
- Artifact free-depth prediction to improve sim2real transfer

$$\mathcal{L} = \lambda_l \mathcal{L}_{inst} + \lambda_{O_{3d}} \mathcal{L}_{O_{3d}} + \lambda_d \mathcal{L}_D$$

$$\mathcal{L}_{\text{inst}} = \sum_{xyg} \left(\hat{Y} - Y\right)^2$$

$$\mathcal{L}_{3D}(O_{3d}, \hat{O}_{3d}) = \begin{cases} \frac{1}{2}(O_{3d} - \hat{O}_{3d})^2 & if|(O_{3d} - \hat{O}_{3d})| < \delta \\ \delta\left((O_{3d} - \hat{O}_{3d}) - \frac{1}{2}\delta\right) & \text{otherwise} \end{cases}$$



CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation

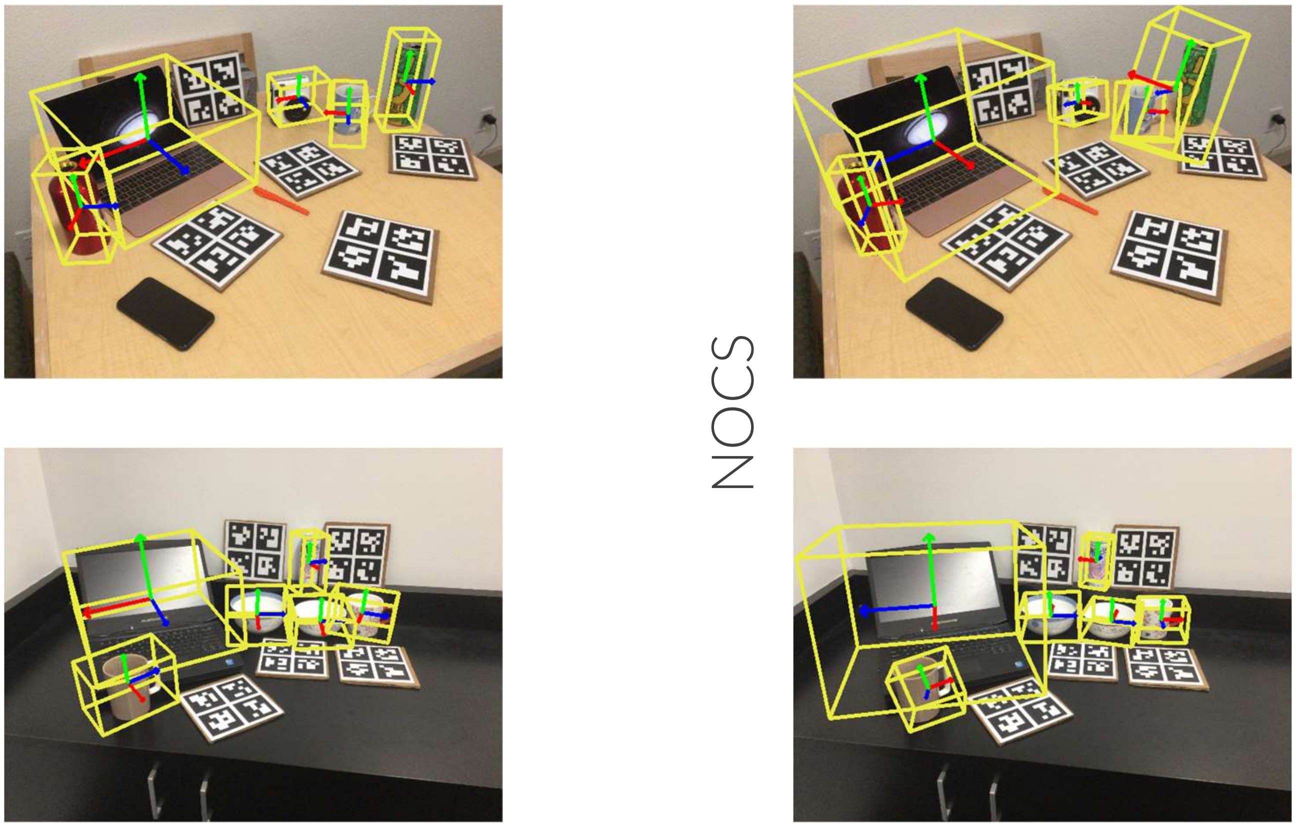
Task Setup/Dataset

- **Dataset:**
  - NOCS Synthetic and Real275 Dataset
- **Objective:**
  - For novel instances, reconstruct their shapes and infer 6D pose and sizes
- **Metrics:**
  - 3D Detection
    - Mean Average Precision (**IOU25, IOU50, IOU75**)
  - 6D pose and size
    - 5° 5cm, 10° 5cm, 10° 10cm
  - 3D shape reconstruction
    - Chamfer Distance (**CD**)

CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation

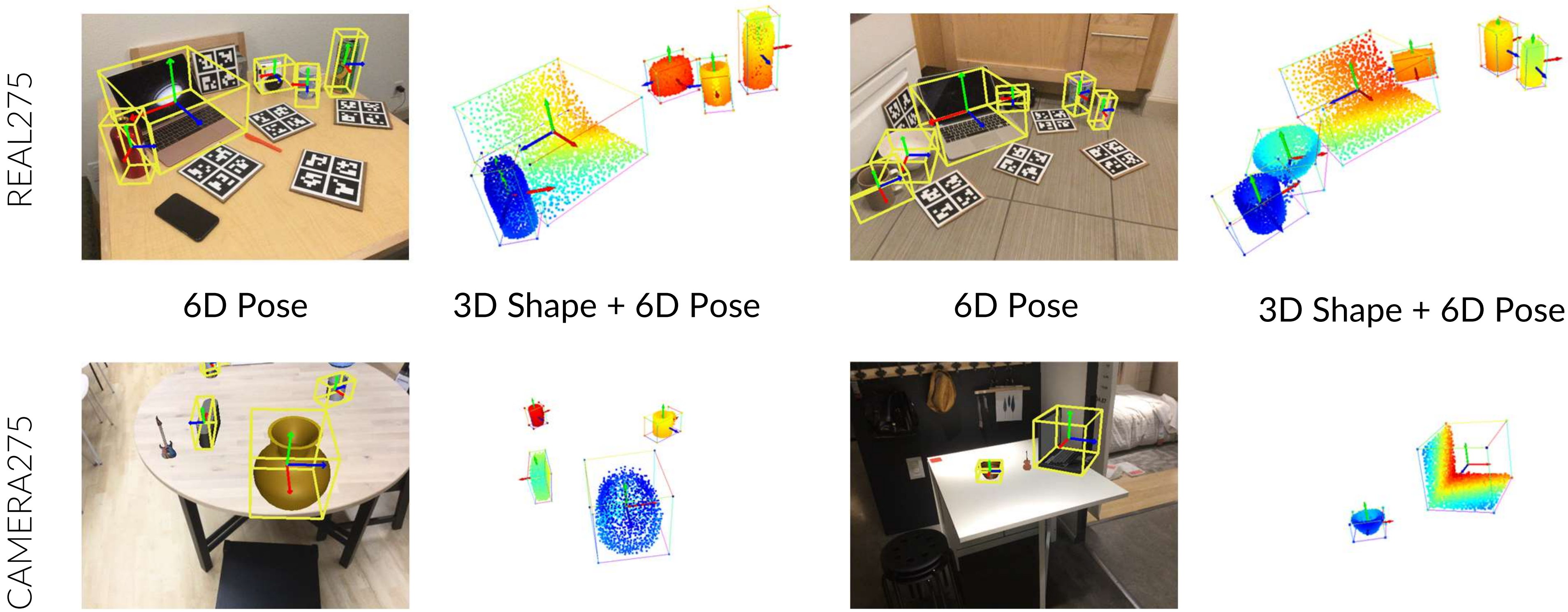
CenterSnap (**Ours**)

NOCS



Qualitative Pose Estimation Results on NOCS-Real275 Dataset

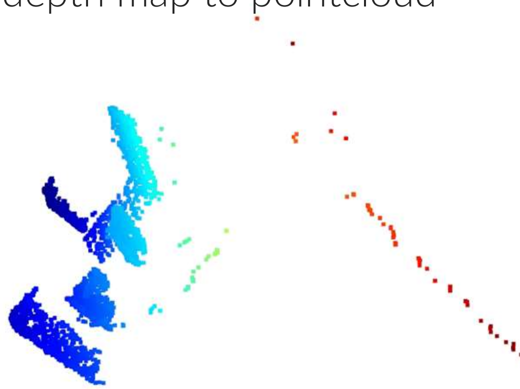
CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation



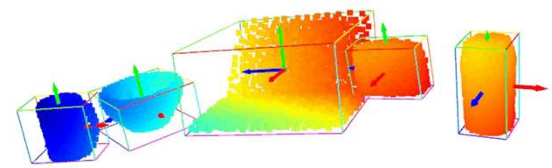
Qualitative Pose Estimation and Shape Reconstruction on NOCS-Real275 Dataset

CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation

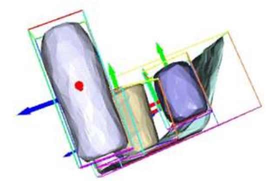
Masked depth map to pointcloud



CenterSnap (Ours) - Pointcloud



Ours - Meshes



Ours - Textures

Comparison to depth-map reconstruction on NOCS-Real275 Dataset

CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation

TABLE I: **Quantitative comparison of 3D object detection and 6D pose estimation on NOCS** [22]: Comparison with strong baselines. Best results are highlighted in **bold**. $*$ denotes the method does not evaluate size and scale hence does not report IOU metric. For a fair comparison with other approaches, we report the per-class metrics using nocs-level class predictions. Note that the comparison results are either fair re-evaluations from the author's provided best checkpoints or reported from the original paper.

		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 [22]	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* [59]	-	-	-	-	-	-	-	-	0.9	1.4	2.4	5.5
3	Metric Scale [60]	**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 [21]	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 [44]	-	-	-	-	-	-	84.2	77.7	23.5	23.8	58.0	58.3
6	**CenterSnap (Ours)**	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 (Ours)**	93.2	**92.5**	**66.2**	**71.7**	**81.3**	**87.9**	83.5	**80.2**	**29.1**	**31.6**	**64.3**	**70.9**

TABLE II: **Quantitative comparison of 3D shape reconstruction on NOCS** [22]: Evaluated with **CD** metric (10^{-2}). Lower is better.

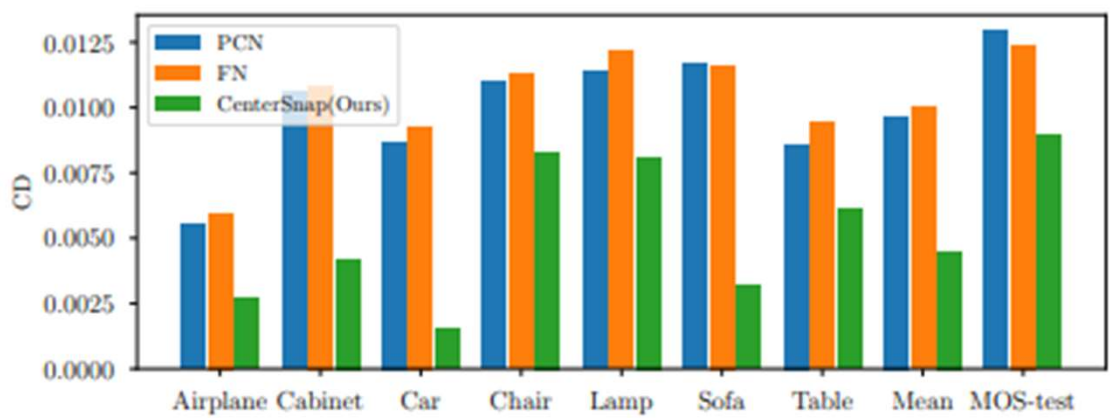
		CAMERA25							**REAL275**						
	Method	**Bottle**	**Bowl**	**Camera**	**Can**	**Laptop**	**Mug**	**Mean**	**Bottle**	**Bowl**	**Camera**	**Can**	**Laptop**	**Mug**	**Mean**
1	Reconstruction [21]	0.18	0.16	0.40	0.097	0.20	0.14	0.20	0.34	0.12	0.89	0.15	0.29	0.10	0.32
2	ShapePrior [21]	0.34	0.22	0.90	0.22	0.33	0.21	0.37	0.50	0.12	0.99	0.24	0.71	0.097	0.44
3	**CenterSnap (Ours)**	0.11	0.10	0.29	0.13	0.07	0.12	0.14	0.13	0.10	0.43	0.09	0.07	0.06	0.15

CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation

Ablation and Shape Reconstruction

- **Effect of :**
  - Input Modailty, (i.e. RGB, Depth or RGB-D), Shape, Training-regime and Depth-Auxiliary loss
- **Conclusions:**
  - Mono-RGB sensors give lowest performance (Depth helps!)
  - Shape prediction network helps boost network's performance (#3 vs #8)
  - Depth auxiliary loss helps Sim2Real Transfer
- **Shape Reconstruction:**
  - Outperforms state-of-the-art supervised shape completion baseline on CD metric

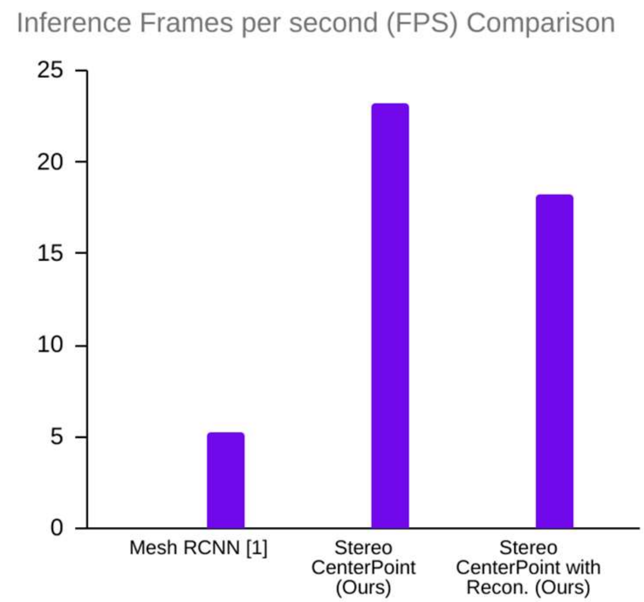
					Metrics				
					3D Shape	6D Pose			
#	Input	Shape	TR	D-Aux	CD ↓	IOU25 ↑	IOU50 ↑	5°10 cm ↑	10°10 cm ↑
1	RGB-D	✓	C	✓	0.19	28.4	27.0	14.2	48.2
2	RGB-D	✓	C+R	✓	0.19	41.5	40.1	27.1	58.2
3	RGB-D*		C+RF	✓	—	—	—	13.8	50.2
4	RGB	✓	C+RF	✓	0.20	63.7	31.5	8.30	30.1
5	Depth	✓	C+RF	✓	0.15	74.2	66.7	30.2	63.2
6	RGB-D	✓	C+RF		0.17	82.3	78.3	30.8	68.3
7	RGB-D	✓	C+RF	✓	0.15	83.5	80.2	31.6	70.9



CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and 6D Pose and Size Estimation for Robust Manipulation

Timing Comparison

- **Result:**
  - Our technique runs at 40 FPS on Nvidia Quadro RTX 5000 GPU
- **Conclusions:**
  - Outperforms MeshRCNN, state-of-the art mesh reconstruction approach by ~4x speed up
- **Shape Reconstruction:**
  - MeshRCNN bottlenecked by 2-stage approach i.e. detection and shape reconstruction
  - Ours is a single-shot with sharable parameters
  - One side note: Less error-compounding since no head is smarter than the others



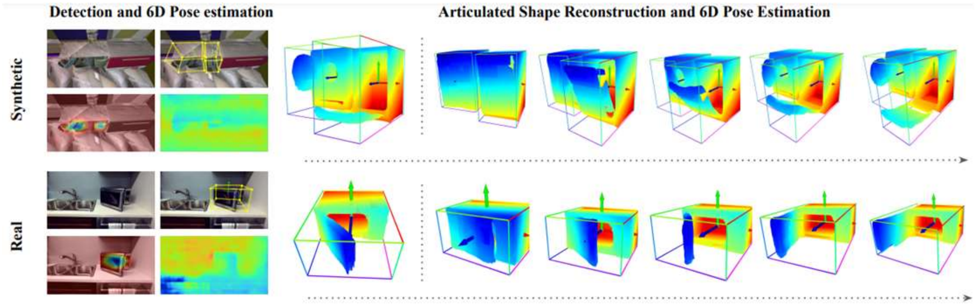
Follow-up work

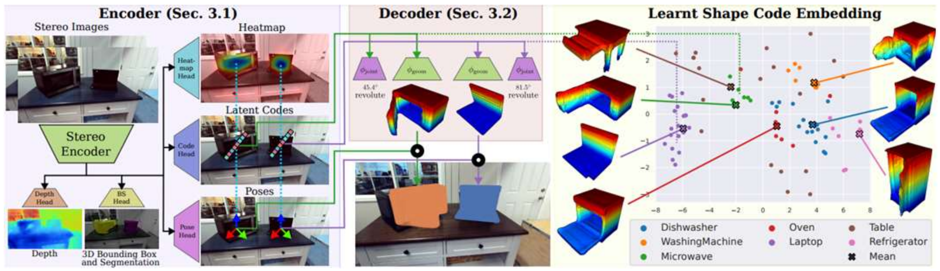
(*Not part of thesis)

CARTO: Category and Joint Agnostic Reconstruction of Articulated Objects

Key highlights:

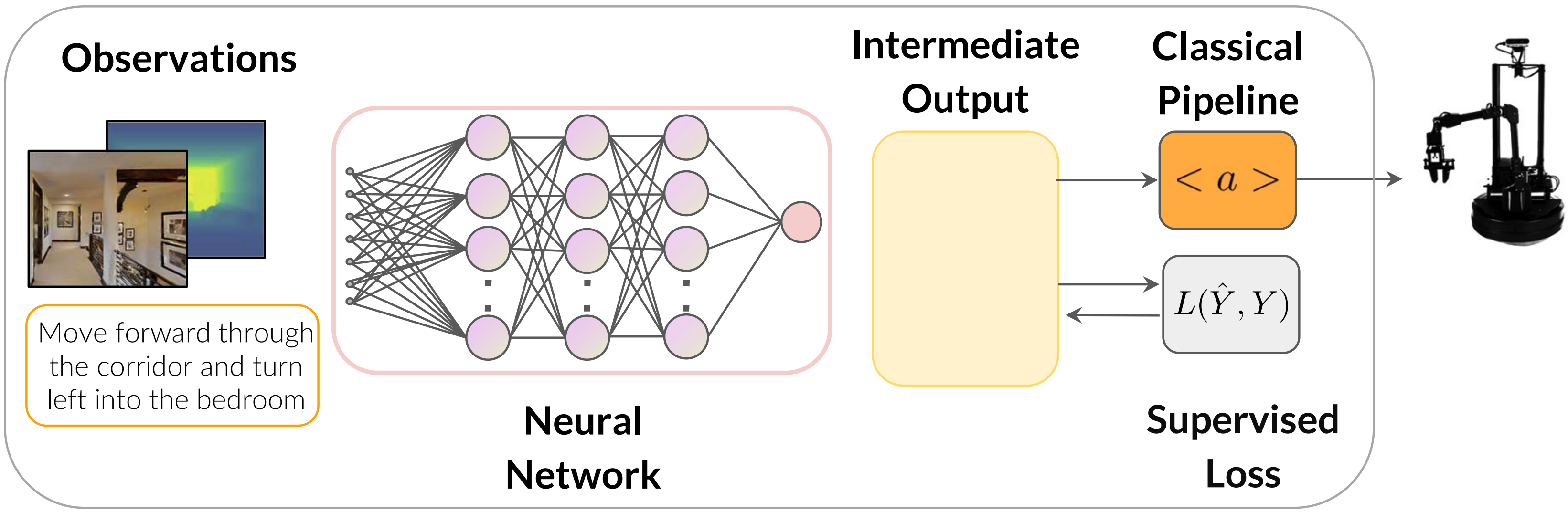
- Extends CenterSnap to Articulated Objects
- Joint-agnostic reconstruction
- Learn a per-category shape and articulation prior
- Fast (~1s) per image articulated reconstruction
- Trained fully in sim, transfers to real-world without re-training or finetuning

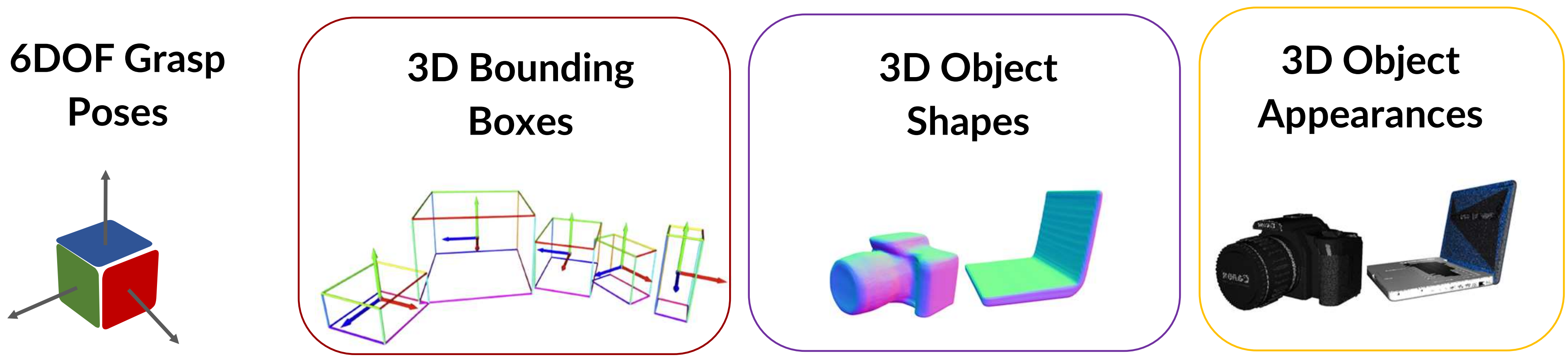




[Ref] N.Heppert, **M.Z.Irshad**, S. Zakharov, K.Liu, R.Ambrus, J.Bohg, A.Valada, T.Kollar, " CARTO: Category and Joint Agnostic Reconstruction of ARTiculated Objects", **CVPR 2023**

Perception for 3D Object Understanding: Current Paradigm

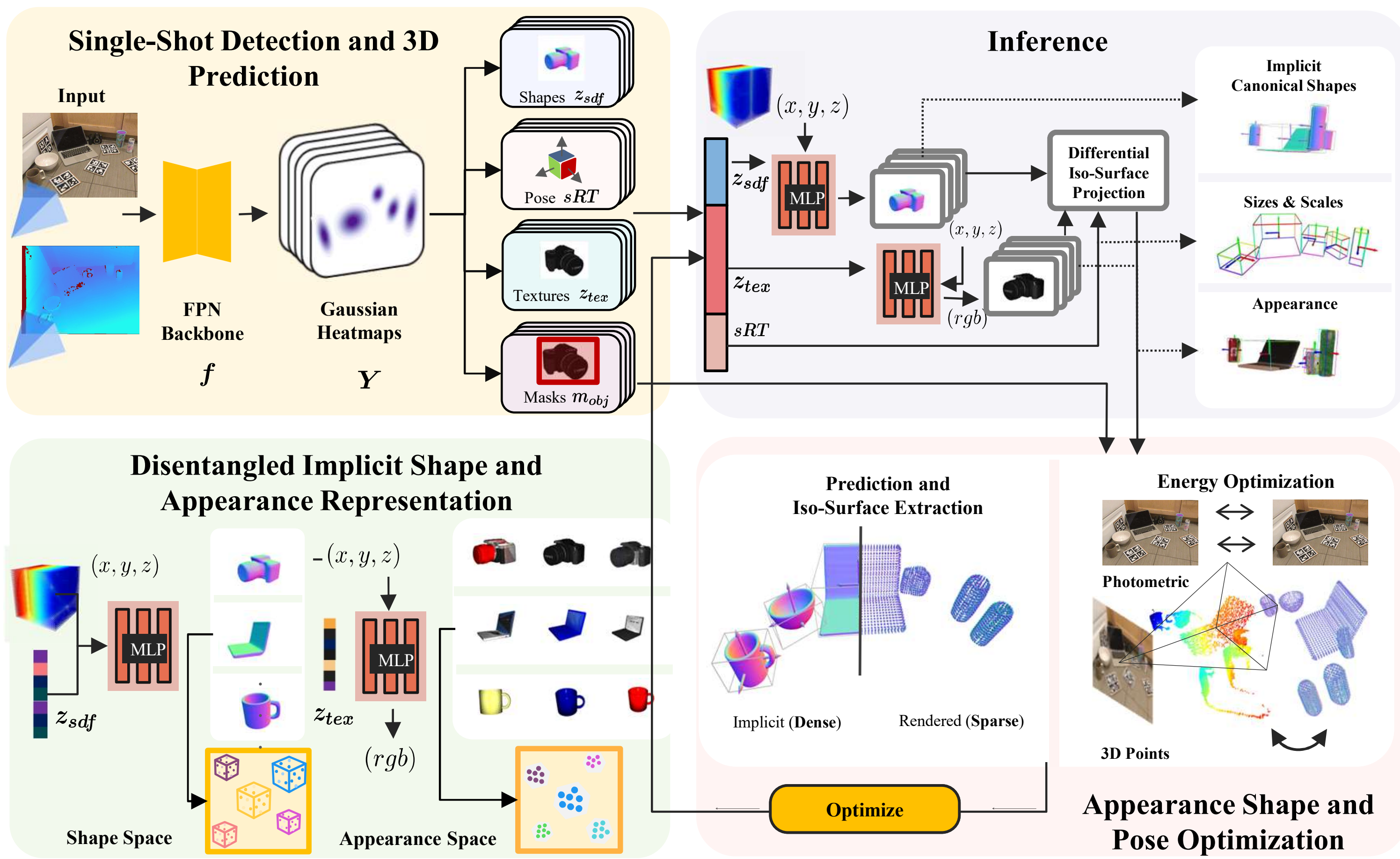




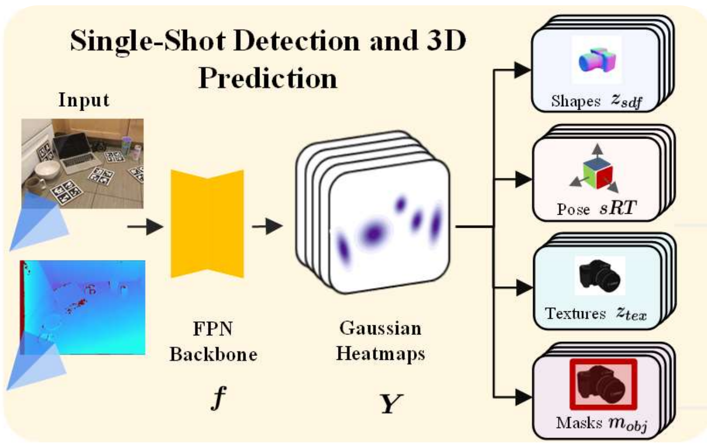
ShAPO : Implicit Representations for Multi Object Shape Appearance and Pose Optimization

*"..Train **intelligent** perception system capable of utilizing **geometry and appearance prior** for **generalizable** shape and appearance reconstruction as well as incorporate object-centric scene context"*

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



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

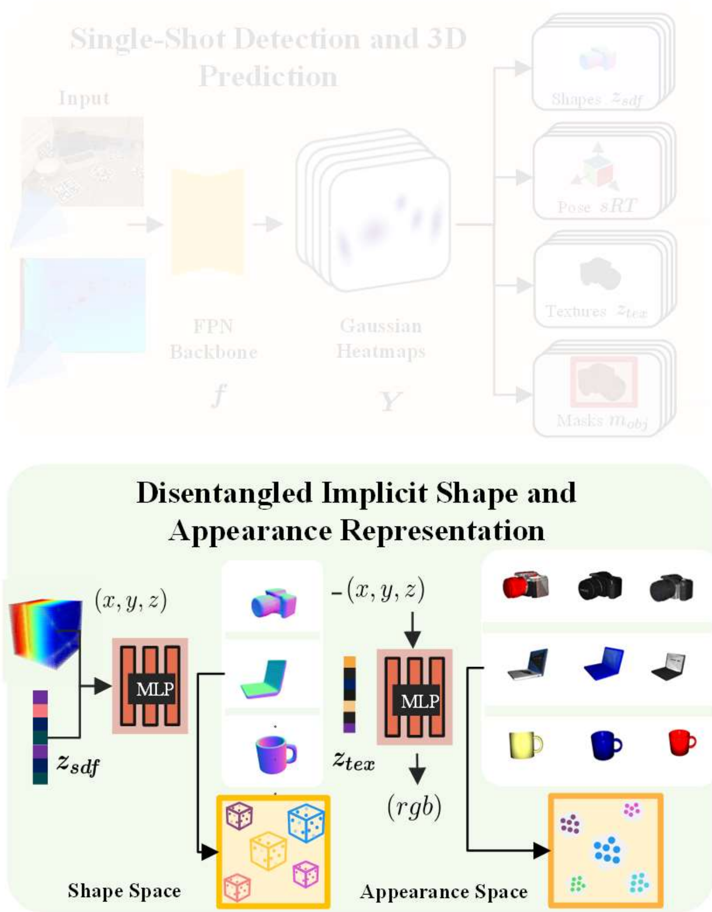


Key highlights:

- Extends CenterSnap to include appearance and segmentation masks
- Single-forward pass for efficiency
- Conv De-conv multi-headed architecture with parameter sharing
- Trained using supervised learning objective

$$\mathcal{L} = \lambda_{inst}\mathcal{L}_{inst} + \lambda_{sdf}\mathcal{L}_{sdf} + \lambda_{tex}\mathcal{L}_{tex} + \lambda_{M}\mathcal{L}_{M} + \lambda_{P}\mathcal{L}_{P}$$

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



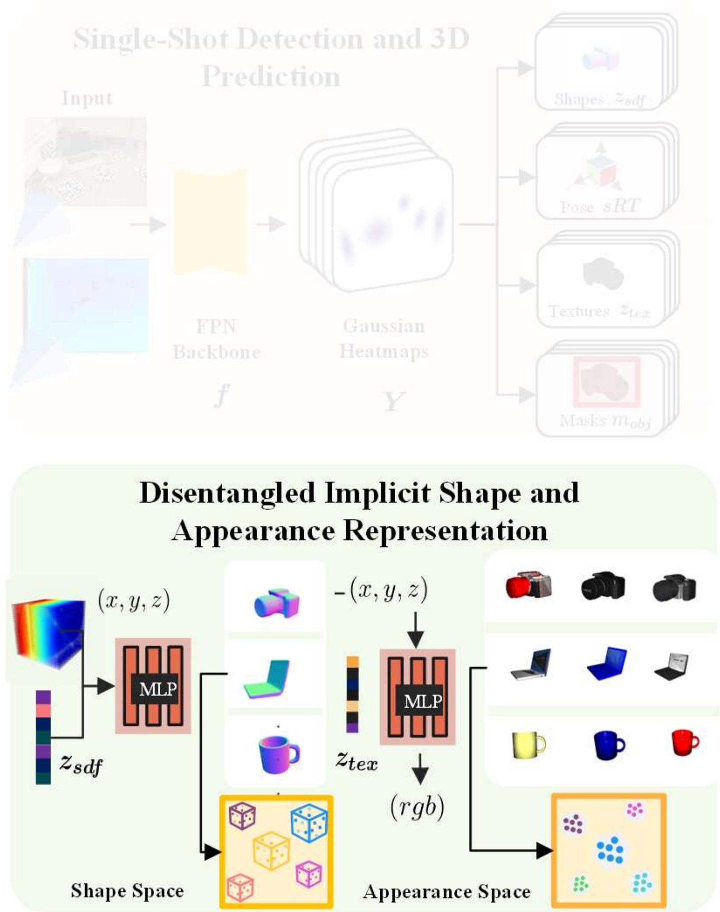
Key highlights:

- Represents geometry as continuous SDF
  - $G(\boldsymbol{x}, \boldsymbol{z}_{sdf}) = s : z_{sdf} \in \mathbb{R}^{64}, s \in \mathbb{R}$
- Represents appearance as Texture Field
  - $t_\theta : \mathbb{R}^3 \rightarrow \mathbb{R}^3$
- Architecture: Single MLP each
- Trained using supervised learning objective
- Dataset: Shapenet synthetic dataset
  - 6 Categories, 1k+ textured models

$$L_{SDF} = |\text{clamp}\,(G(\boldsymbol{x}, \boldsymbol{z}_{sdf}), \delta) - \text{clamp}\,(\boldsymbol{s}_{gt}, \delta)| + L_{contrastive}(\boldsymbol{z}_{sdf})$$

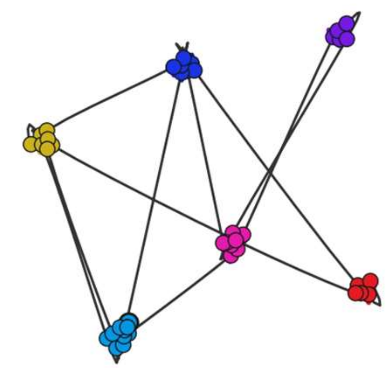
$$L_{RGB} = \sum_{n=1}^{N} \|\boldsymbol{c}_{gt} - t_\theta(\boldsymbol{x}, \boldsymbol{z}_{sdf}, \boldsymbol{z}_{tex})\|_2^2$$

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

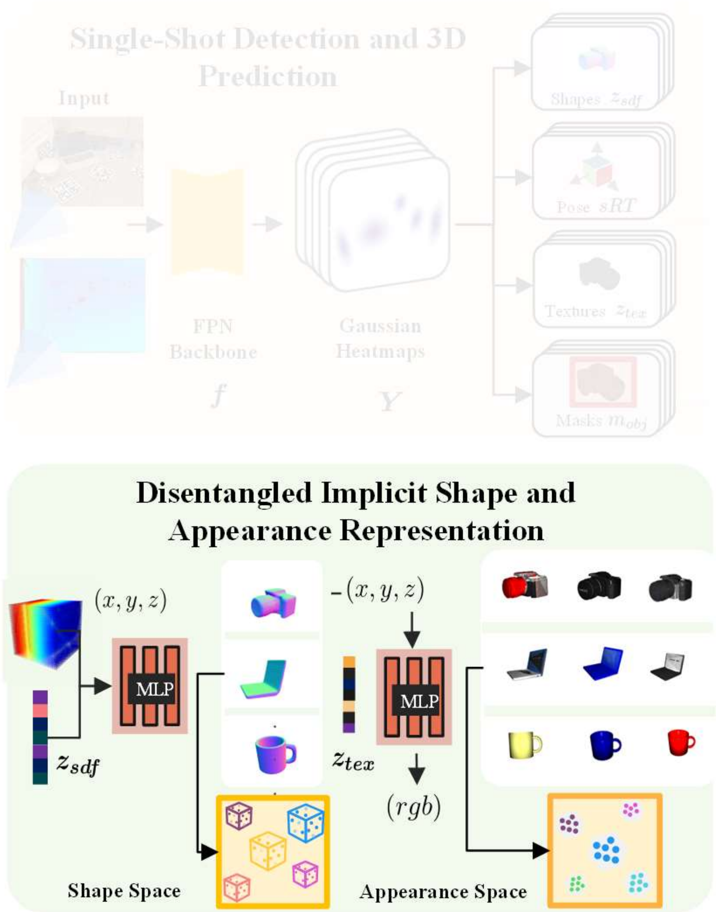


Key highlights:

- Represents geometry as continuous SDF
  - $G(\boldsymbol{x}, \boldsymbol{z}_{sdf}) = s : z_{sdf} \in \mathbb{R}^{64}, s \in \mathbb{R}$
- Represents geometry as continuous SDF
  - $t_\theta : \mathbb{R}^3 \rightarrow \mathbb{R}^3$

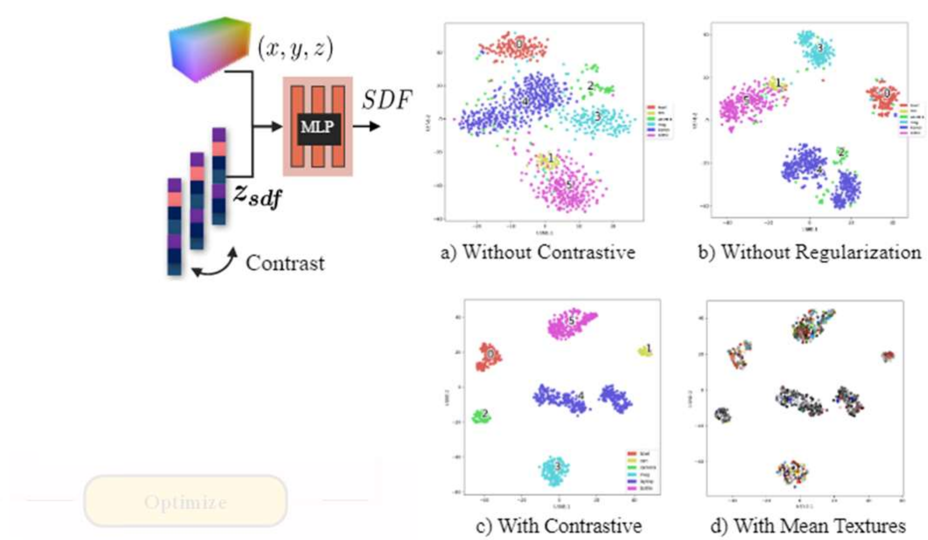


ShAPO : Implicit Representations for Multi Object Shape Appearance and Pose Optimization



Key highlights:

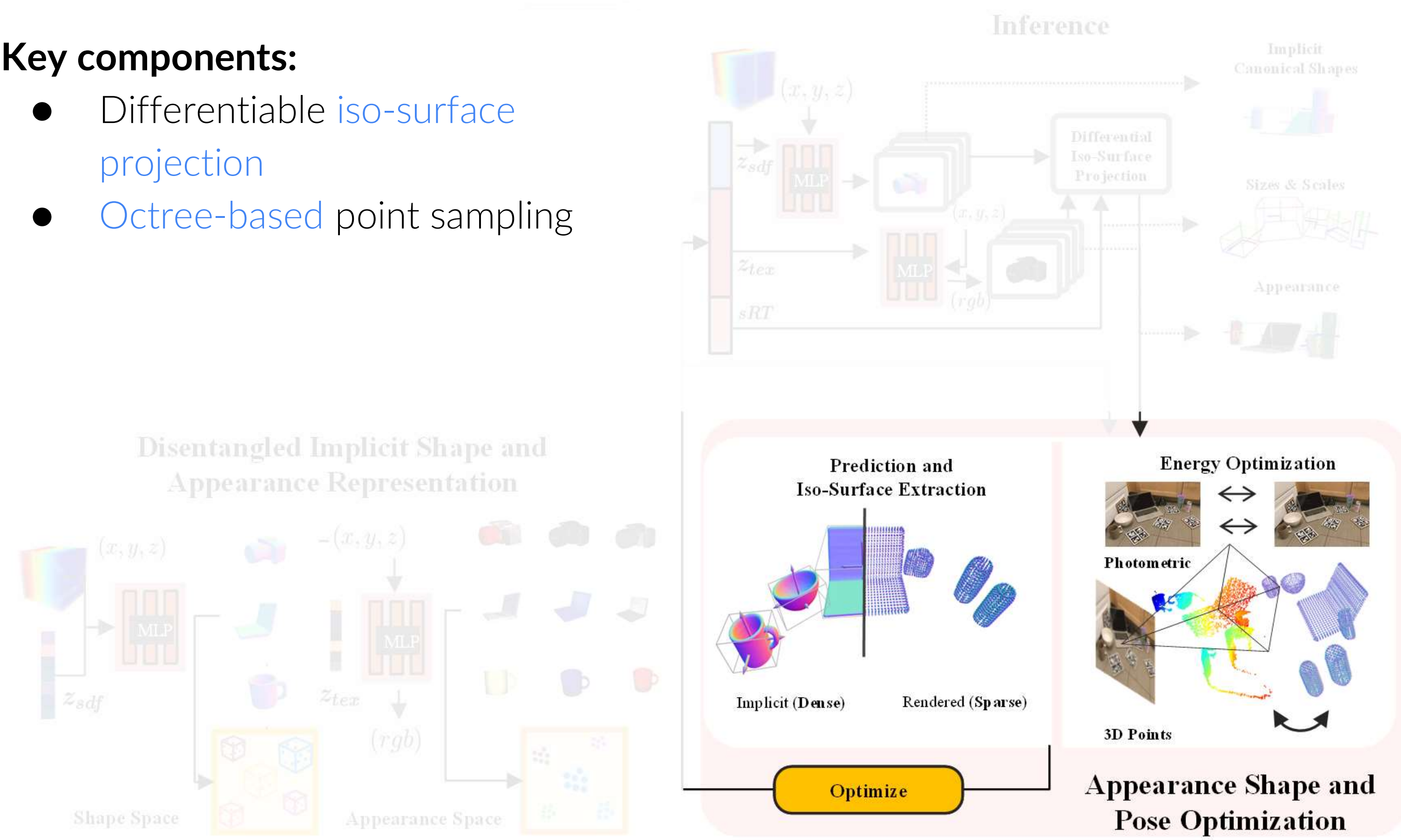
- Represents geometry as continuous SDF
  - $G(\boldsymbol{x}, \boldsymbol{z}_{sdf}) = s : z_{sdf} \in \mathbb{R}^{64}, s \in \mathbb{R}$
- Represents geometry as continuous SDF
  - $t_\theta : \mathbb{R}^3 \rightarrow \mathbb{R}^3$



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

Key components:

- Differentiable iso-surface projection
- Octree-based point sampling



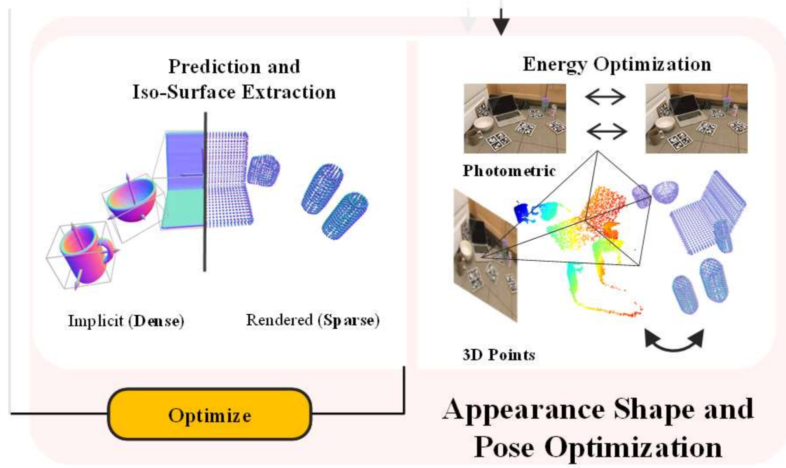
ShAPO : Implicit Representations for Multi Object Shape Appearance and Pose Optimization

Differentiable iso-surface projection:

- Trivial Solution: Threshold the points based on SDF value, Non-Differentiable
- Alternate solution: Utilize gradients and normal values (Ours)

$$n_i = \frac{\partial G(x_i; \mathbf{z}_{sdf})}{\partial x_i}$$

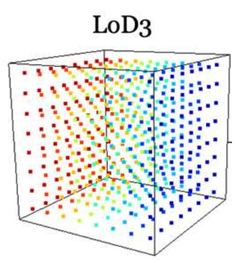
$$p_i = x_i - \frac{\partial G(x_i; \mathbf{z}_{sdf})}{\partial x_i} G(x_i; \mathbf{z}_{sdf})$$



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

Octree-based point sampling:

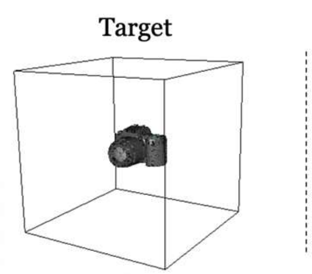
- Brute Force Solution: Extremely inefficient
- 603 points = 216000 ~= 1600 surface points (0.7%)
- Solution: Coarse-to-fine sampling
- LoD3 to LoD7



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

Octree-based point sampling:

- Brute Force Solution: Extremely inefficient
- 603 points = 216000 ~= 1600 surface points (0.7%)
- Solution: Coarse-to-fine sampling
- LoD3 to LoD7



ShAPO : Experiments

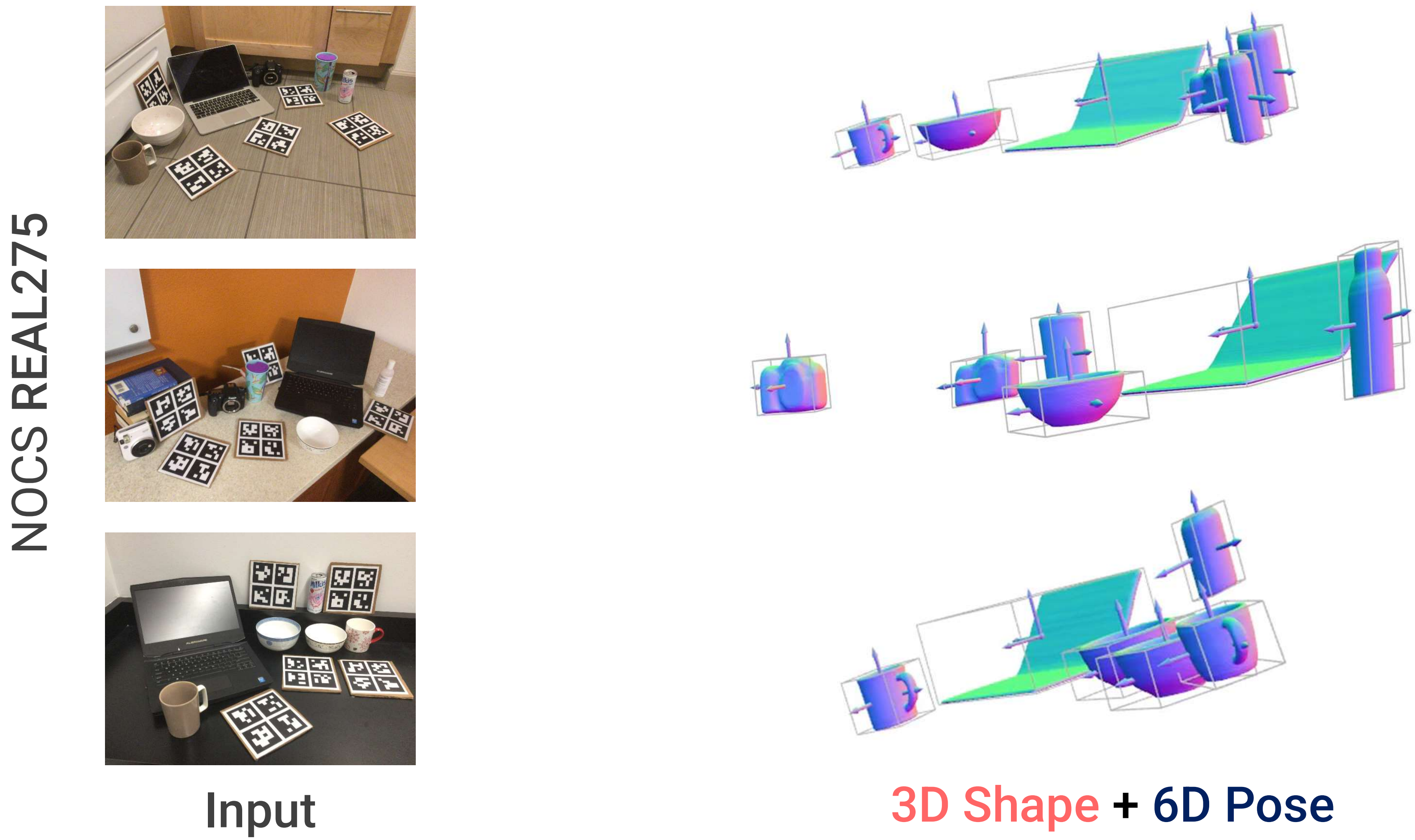
How well does ShAPO **recover pose and sizes** of novel objects?

How well does ShAPO perform in terms of **reconstructing geometry and appearance** of multiple objects from a single-view RGB-D observation?

How well does our **differentiable iterative improvement and multi-level optimization** impact shape, appearance, pose and size?

ShAPO : Qualitative Results

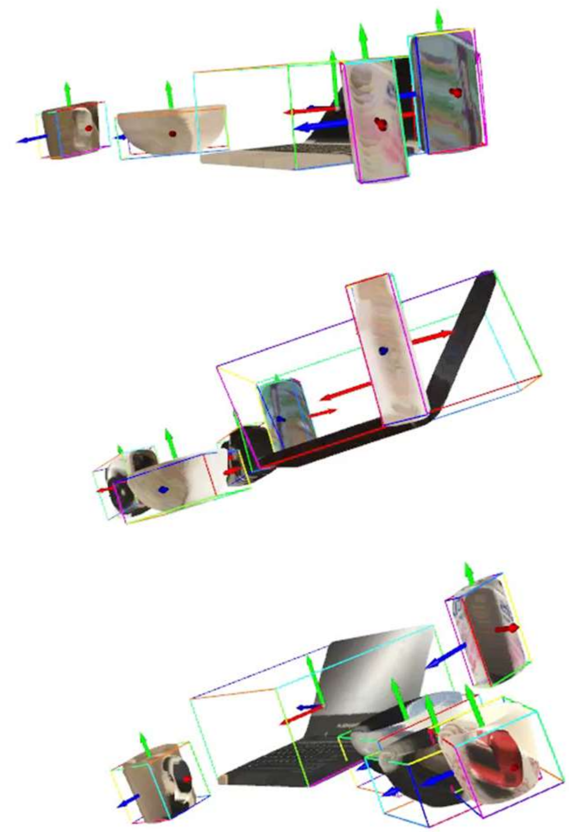
Our qualitative results show complete and accurate shape reconstruction with **fine-grained geometric detail**



Our qualitative results show complete and accurate texture reconstruction with **fine-grained geometric detail**

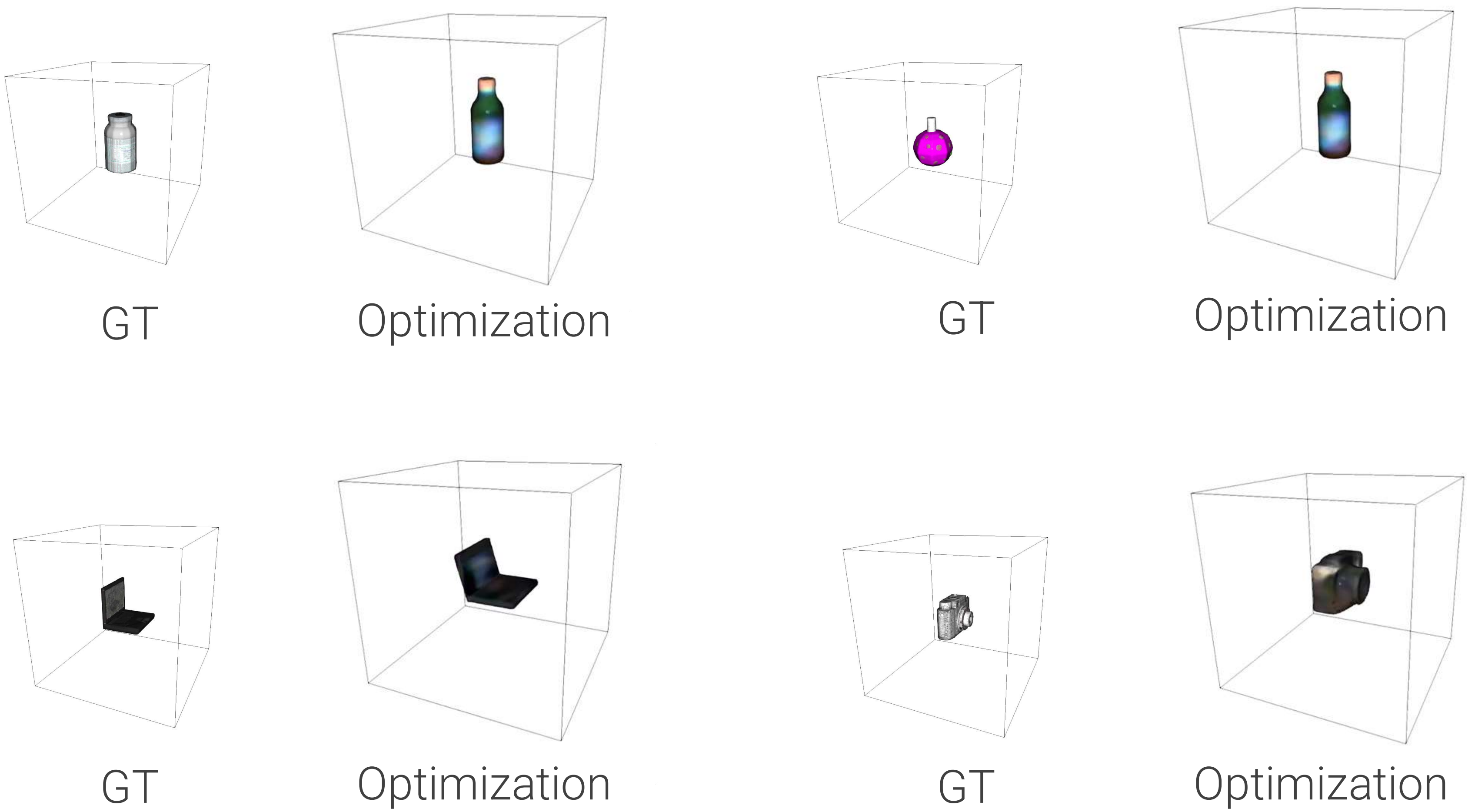
NOCS REAL275

Input

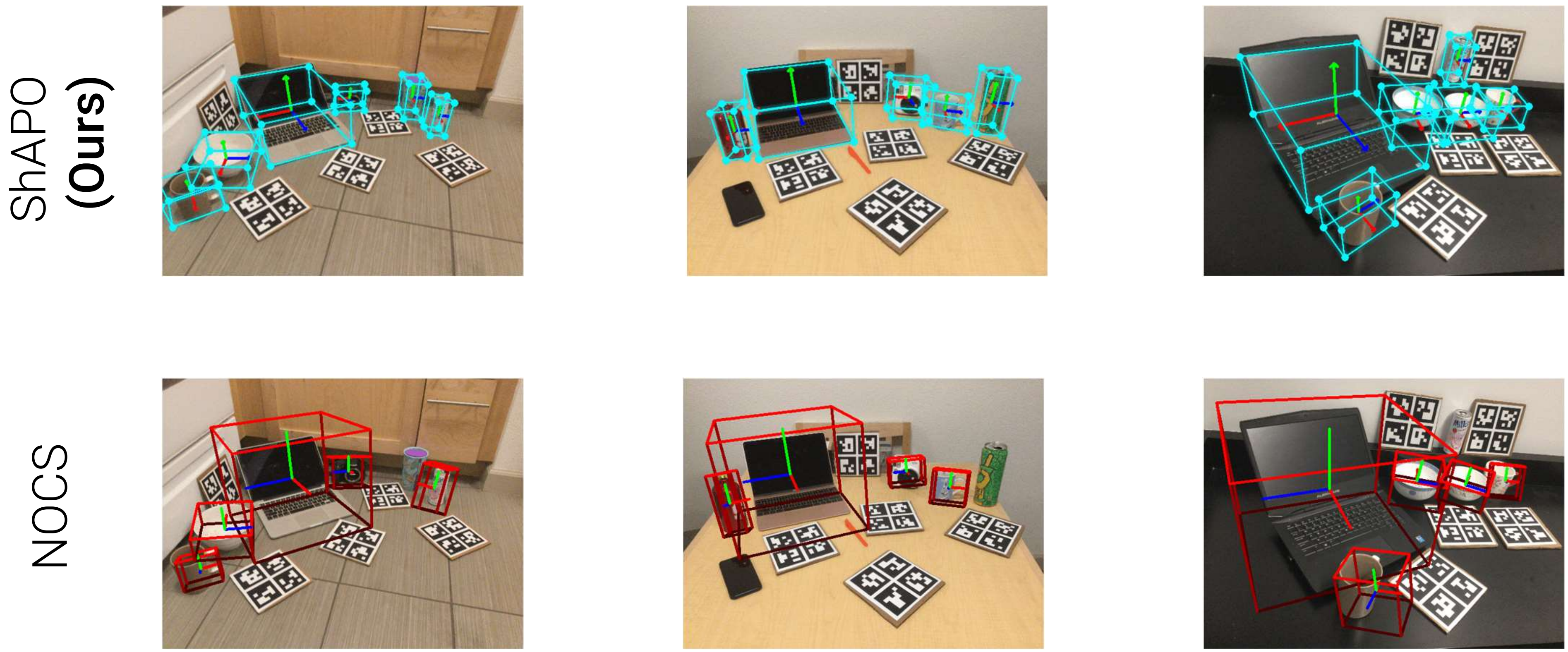


3D Shape + 6D Pose + Appearance

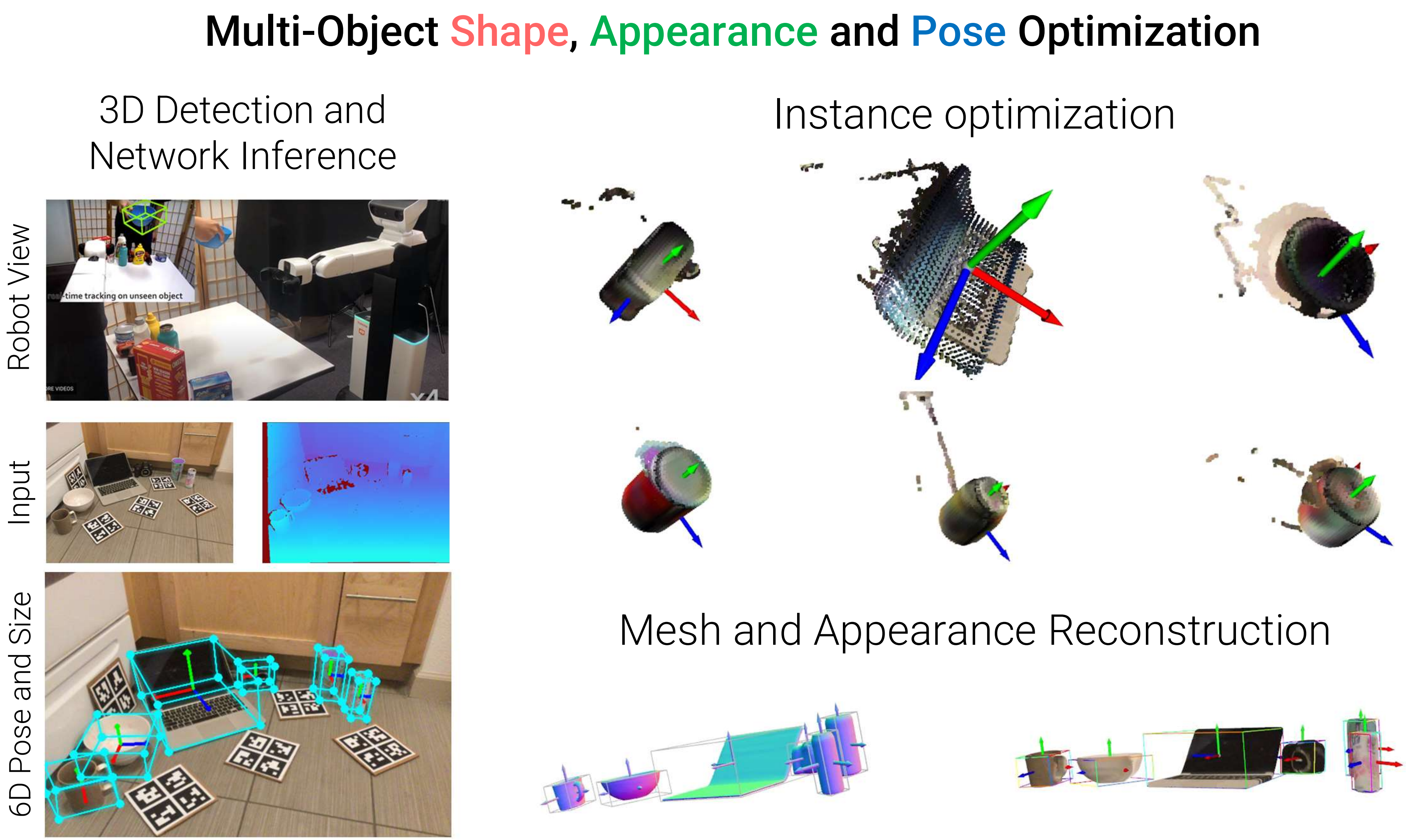
Our novel implicit textured representation learns to **embed objects** in a concise space for **downstream optimization**



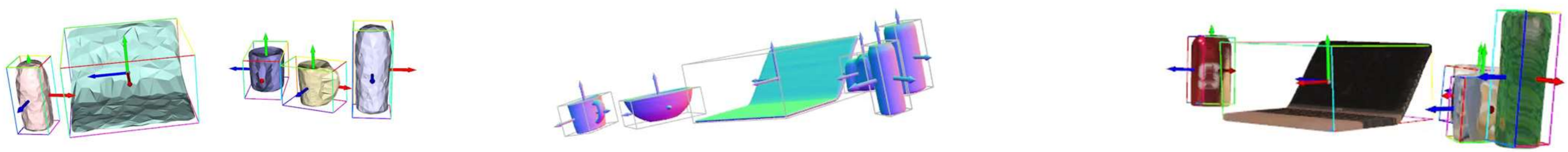
Our inference-time optimization allows us to perform **accurate 6D pose and size estimation**



Testing Results on **NOCS-Real275** Dataset


Multi-Object Shape, Appearance and Pose Optimization
3D Detection and Network Inference
Robot View
Input
6D Pose and Size
Instance optimization
Mesh and Appearance Reconstruction

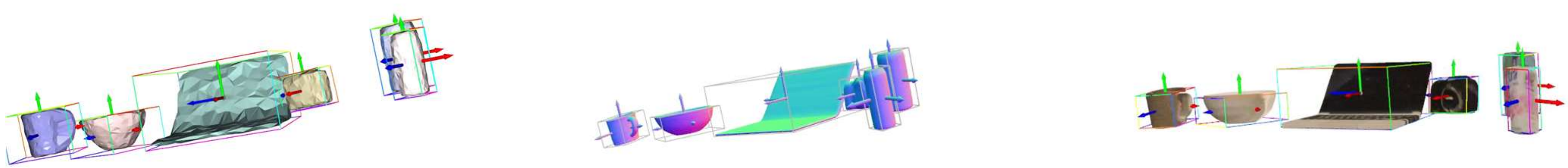
Our superior **shape** and **appearance** reconstruction in comparison to strong baseline *CenterSnap*



CenterSnap Mesh

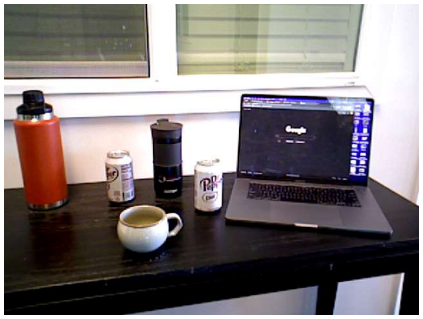
ShAPO **(Ours)** Mesh

ShAPO **(Ours)** Appearance

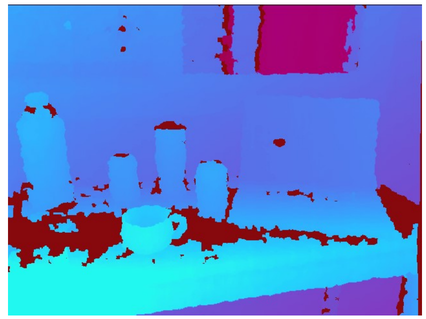


Testing Results on **NOCS-Real275** Dataset

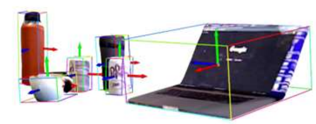
Our results on real-world single-view RGBD
captured on an HSR Robot Camera



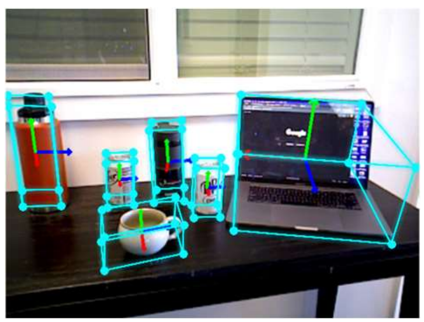
RGB



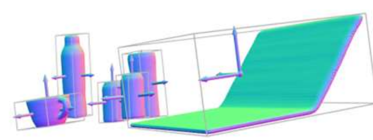
Depth



Appearance
Reconstruction



6D pose and size



3D Shape

Testing Results on **Xtion Pro Live** Camera on HSR Robot

ShAPO : Quantitative Results

- Compared against 7 baseline variations:

  1. NOCS 2. Synthesis 3. Metric Scale 4. Shape Prior 5. CASS 6. CenterSnap

- Outperform baselines on 6D pose and size, 3D shape

Table 2: **Quantitative comparison of 6D pose estimation and 3D object detection on NOCS** [41]: Comparison with strong baselines. Best results are highlighted in **bold**. $*$ denotes the method does not report IOU metrics since size and scale is not evaluated. We report metrics using nocs-level class predictions for a fair comparison with all baselines.

	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**

Table 3: **Quantitative comparison of 3D shape reconstruction on NOCS** [41]: Evaluated with **CD** metric (10^{-2}). Lower is better.

	CAMERA25							REAL275						
Method	Bottle	Bowl	Camera	Can	Laptop	Mug	Mean	Bottle	Bowl	Camera	Can	Laptop	Mug	Mean
1 Reconstruction [37]	0.18	0.16	0.40	0.097	0.20	0.14	0.20	0.34	0.12	0.89	0.15	0.29	0.10	0.32
2 ShapePrior [37]	0.34	0.22	0.90	0.22	0.33	0.21	0.37	0.50	0.12	0.99	0.24	0.71	0.097	0.44
3 CenterSnap	0.11	0.10	0.29	0.13	0.07	0.12	0.14	0.13	0.10	0.43	0.09	0.07	0.06	0.15
3 ShAPO (Ours)	0.14	0.08	0.2	0.14	0.07	0.11	0.16	0.1	0.08	0.4	0.07	0.08	0.06	0.13

ShAPO : Quantitative Results

- Compared CD, PSNR and Sample Efficiency of different level of details (LoDs)
- LoD7 has the higher accuracy while LoD6 gives the best speed/accuracy trad-off
- PSNR for novel real-world scenes after inference, optimization and fine-tuning

Table 4: **Generalizable Implicit Representation Ablation**: We evaluate the efficiency (point sampling/time(s)/memory(MB)) and generalization (shape(CD) and texture(PSNR) reconstruction) capabilities of our implicit object representation as well as its sampling efficiency for different levels of detail (LoDs) and compare it to the ordinary grid sampling. All ablations were executed on NVIDIA RTX A6000 GPU.

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

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

Collaborators


Zsolt Kira


Rares Ambrus


Sergey Zakharov


Adrien Gaidon


Thomas Kollar


Michael Laskey

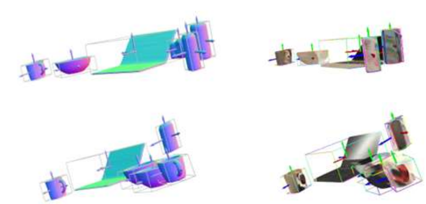

Kevin Stone

Thank you!

Question?

[MTMKZ ICRA-22]

CenterSnap: 3D geometry prior for fast, multi-object 3D object-centric learning



[MSRTAZ ECCV-22]

ShAPO: 3D shape and appearance prior for accurate object-centric scene reconstruction