DeepRob

[Student] Lecture 15

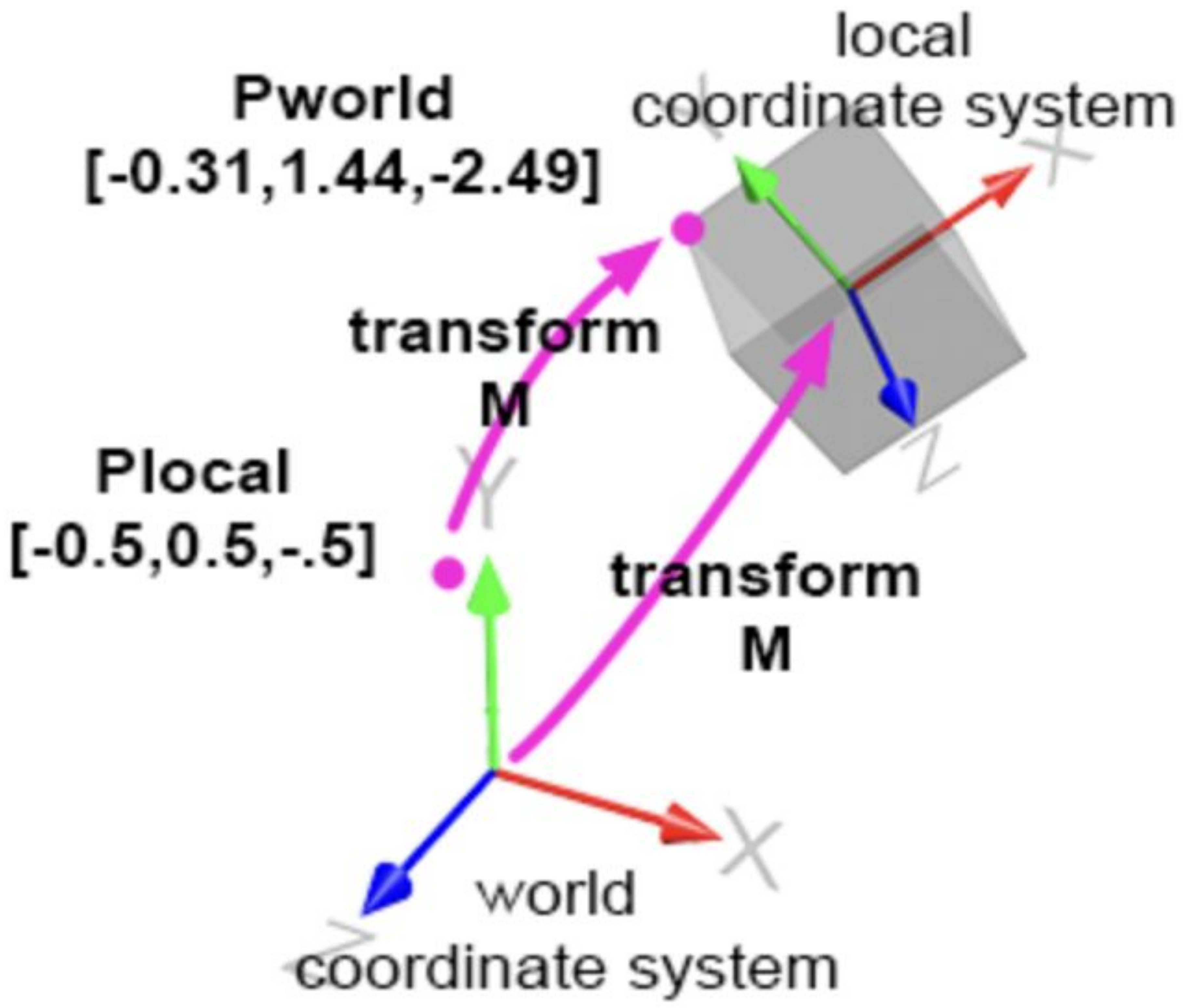
by Bharath Sivaram, Sahith Reddy, Prakadeeswaran Manivanan

Rigid Object Perception, Dense Descriptors, Category-level Object Pose Estimation

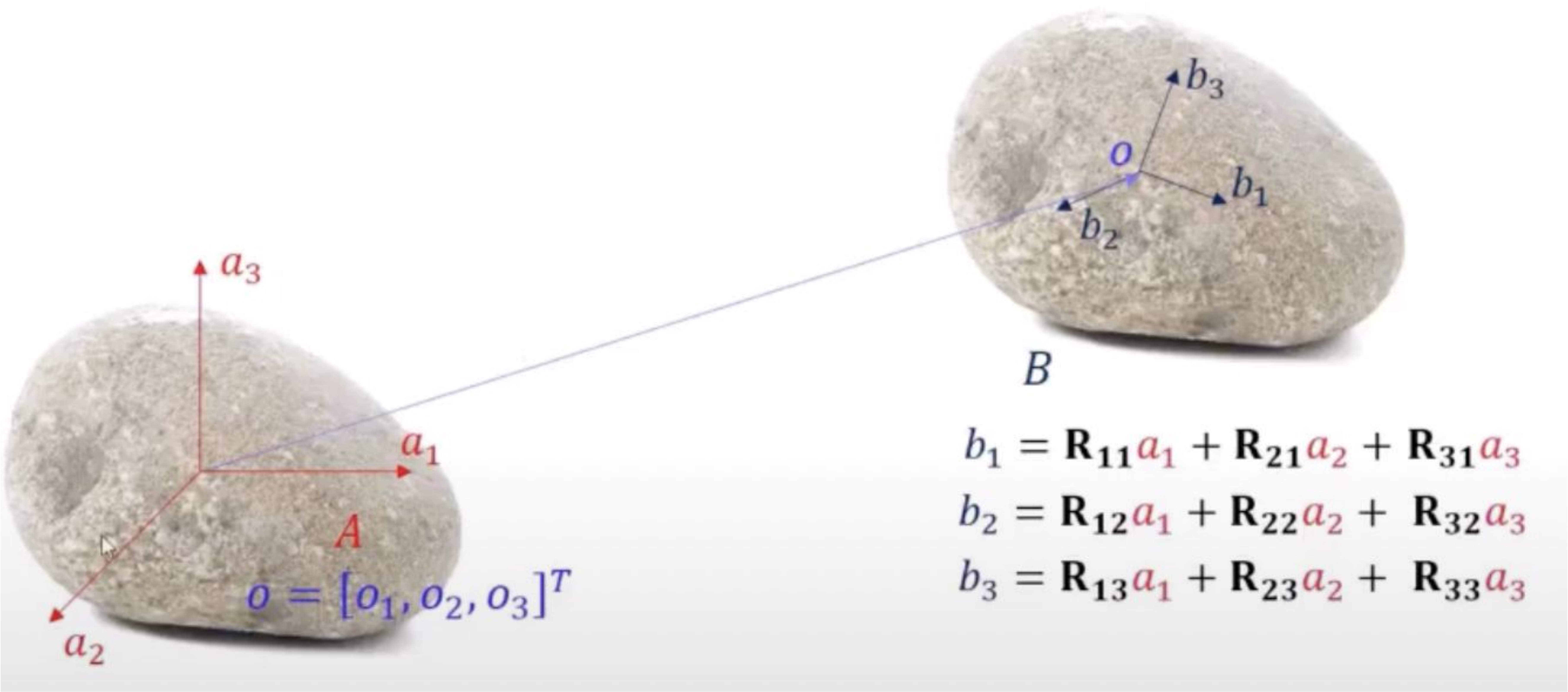
University of Michigan and University of Minnesota

What is a point transform?

- A mathematical operation that changes the position and orientation of a point in space
- Involves rotation and translation

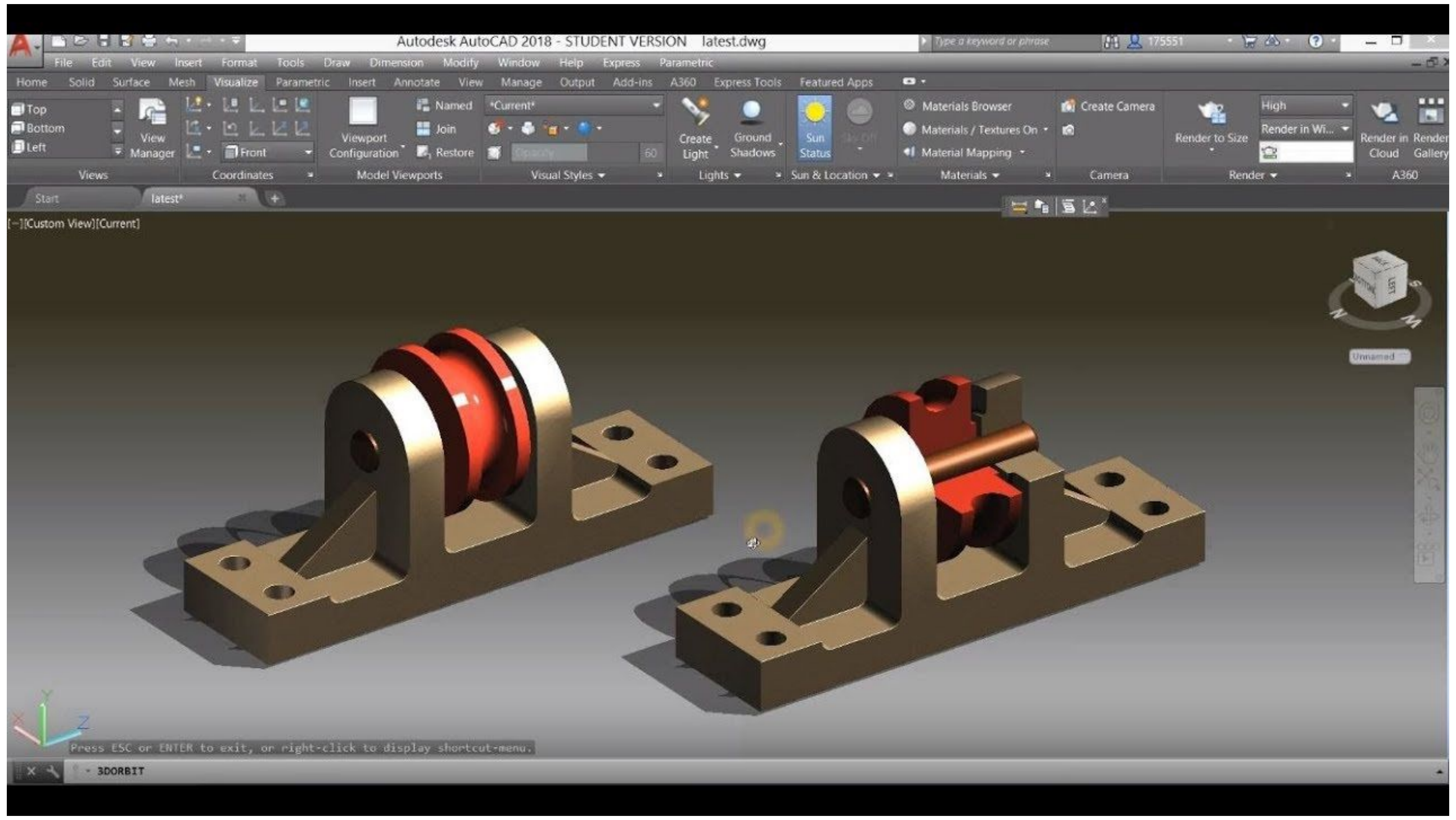


Pose as an Object



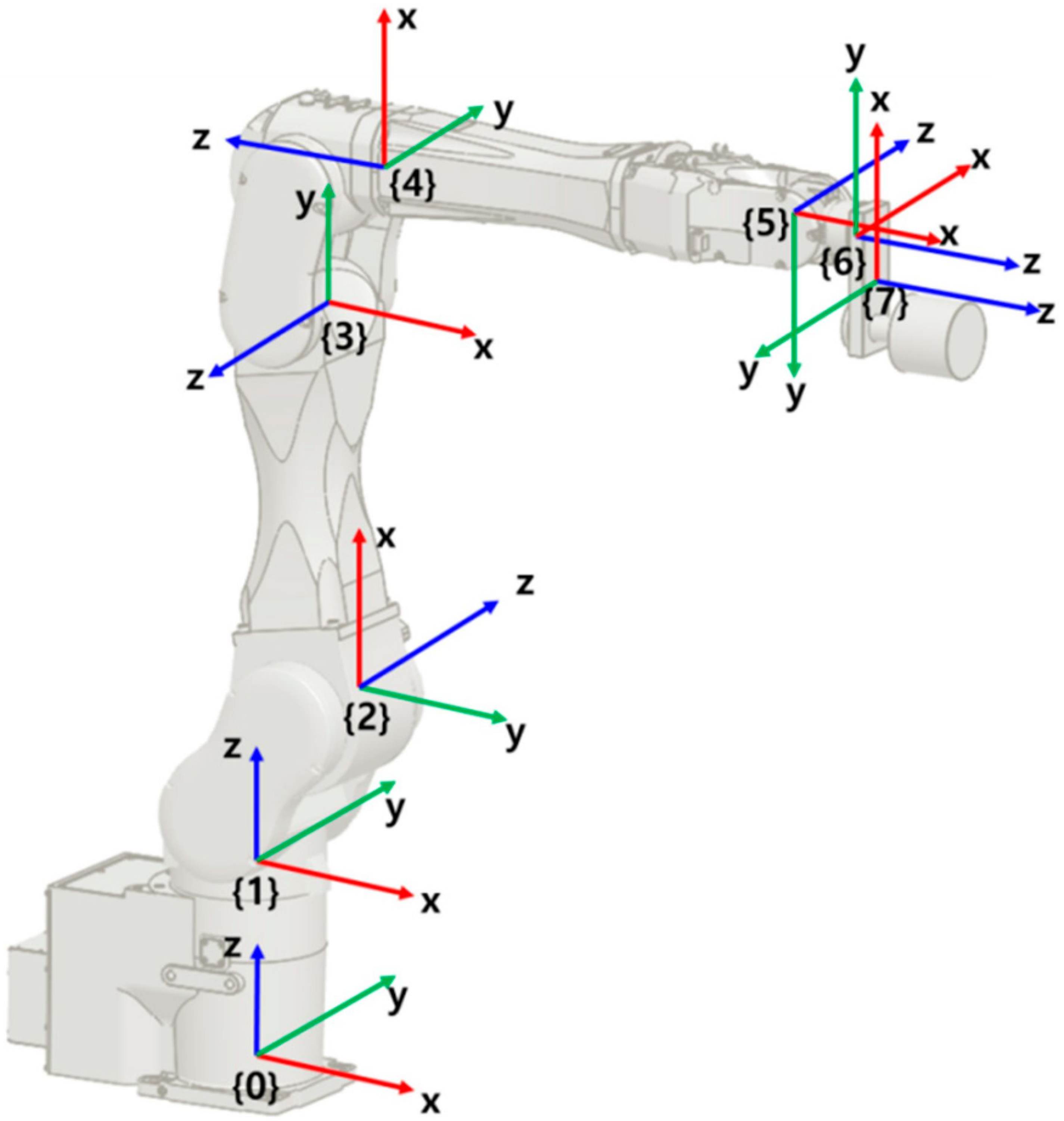
Source: https://www.youtube.com/watch?v=Jc5PlahEfU4

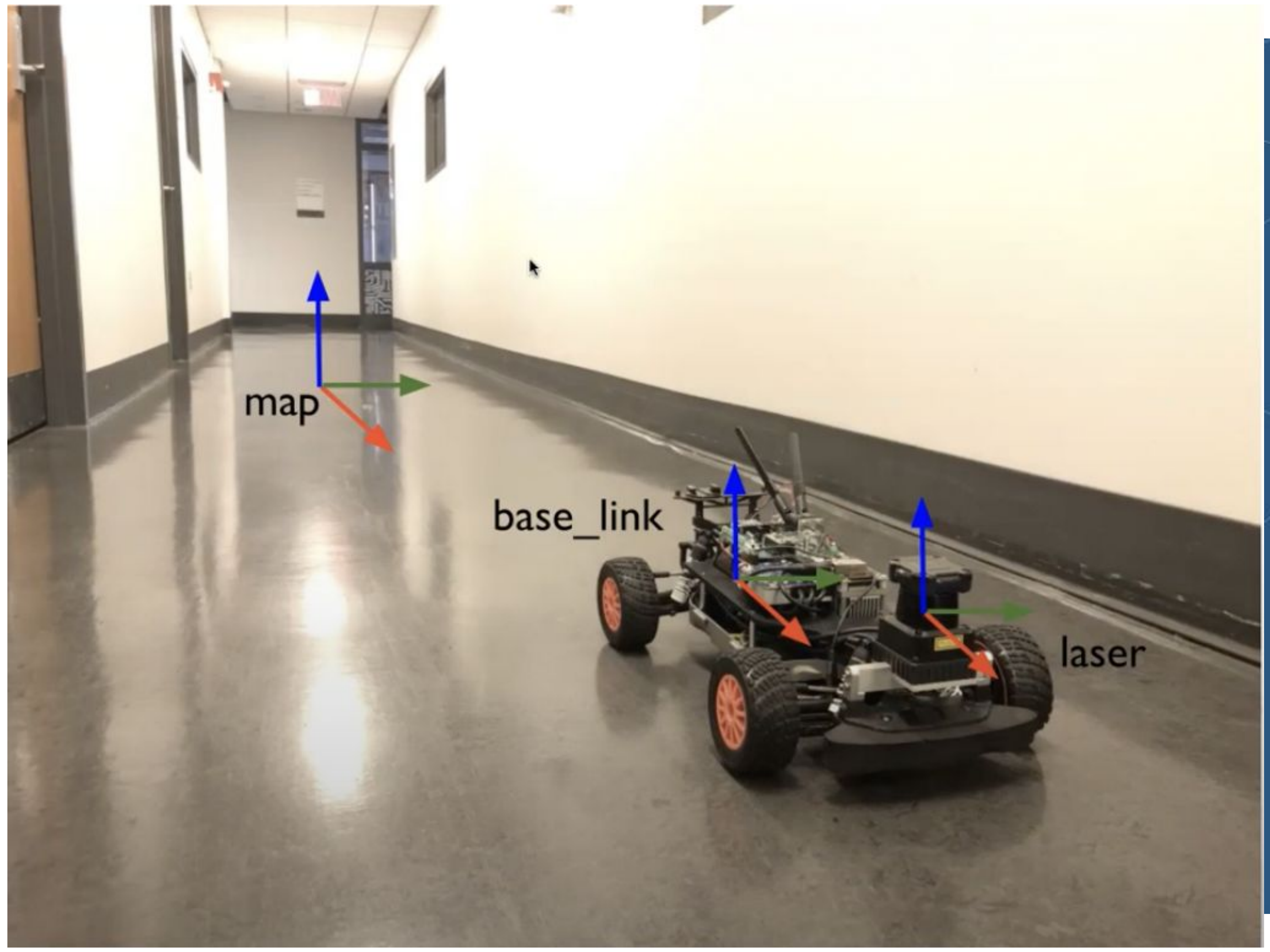
Pose in engineering

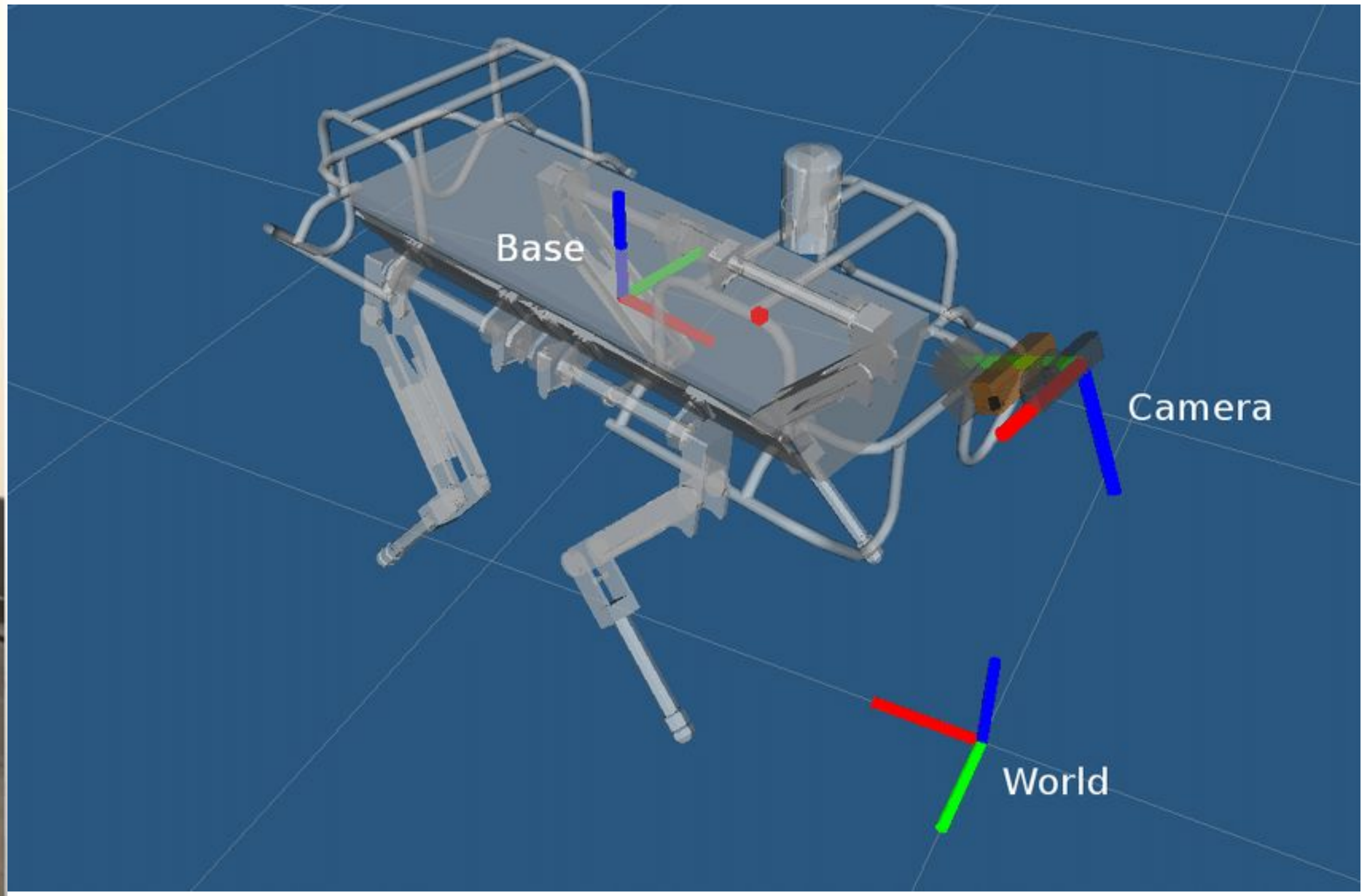


Design and Build of Objects

Pose for robotics





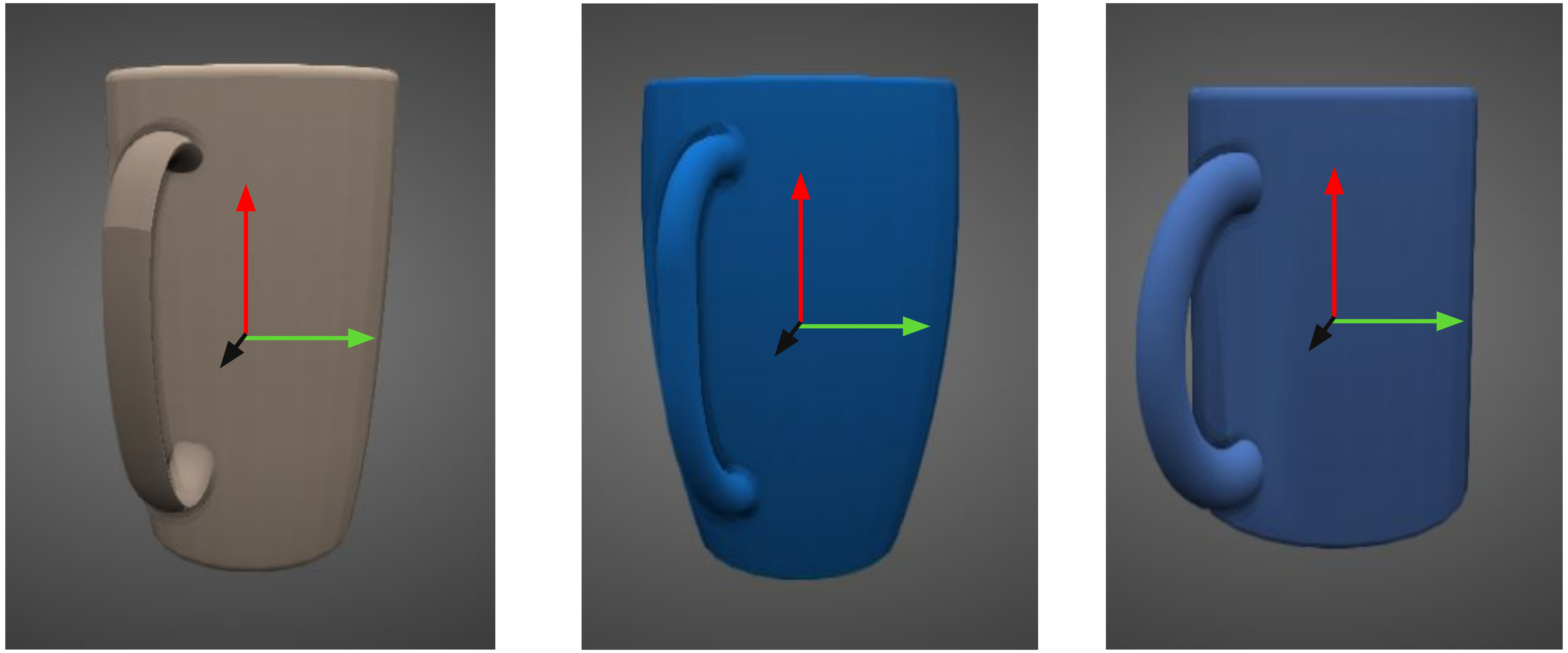


How to define a pose for an object?

Local Reference Frame for Manipulation

- Local frame of reference is subjective
- Must be assigned carefully by designer
- Common Orientation for object, also showcasing features (handle on mug for ex)

ShapeNetCore



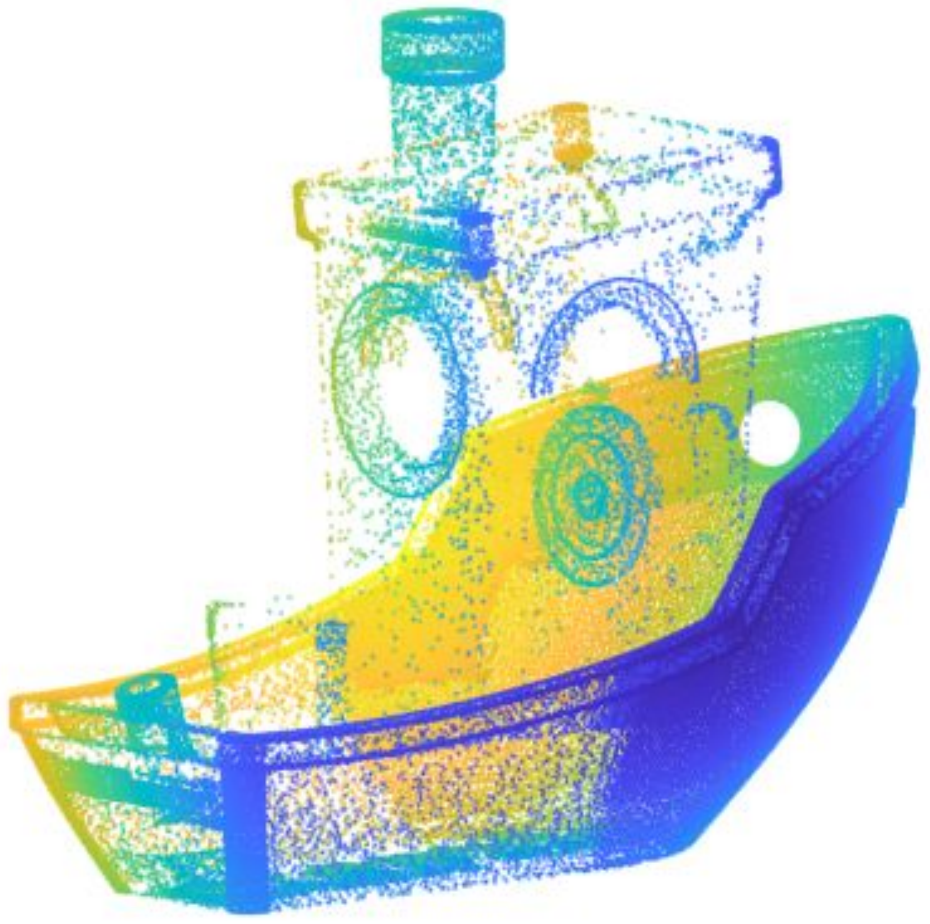
- Upright Orientation, usually from CAD model
- Front orientation which usually aligns with an axis of CAD model

Model Capture and Format

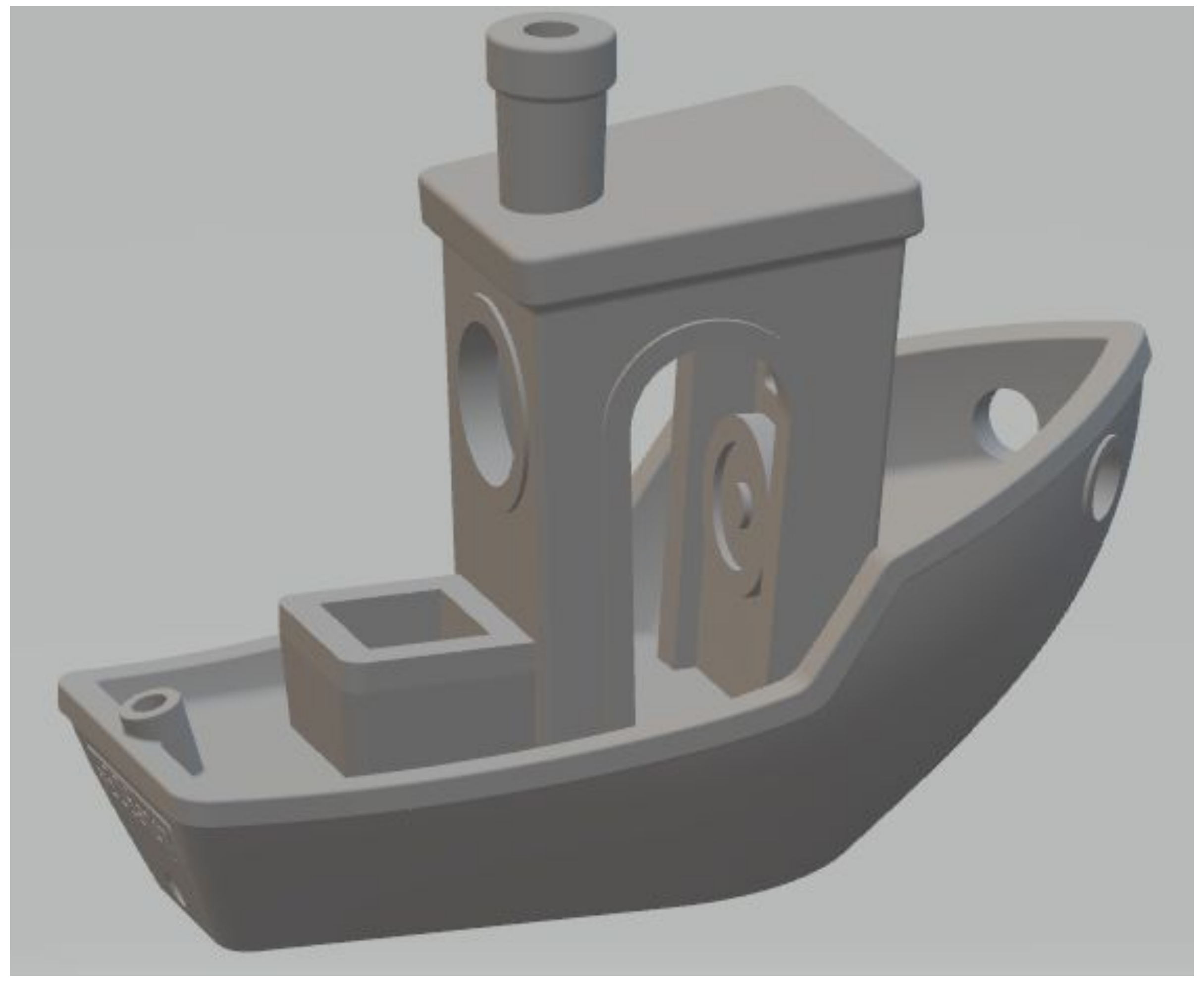
Azure Kinect

Intel RealSense

Structure Sensor

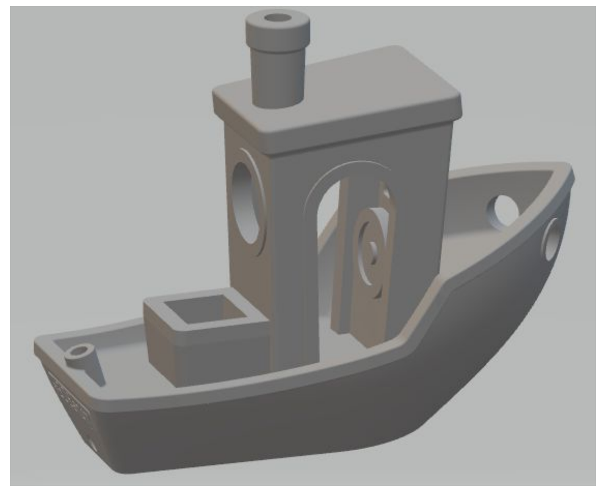


Point Cloud



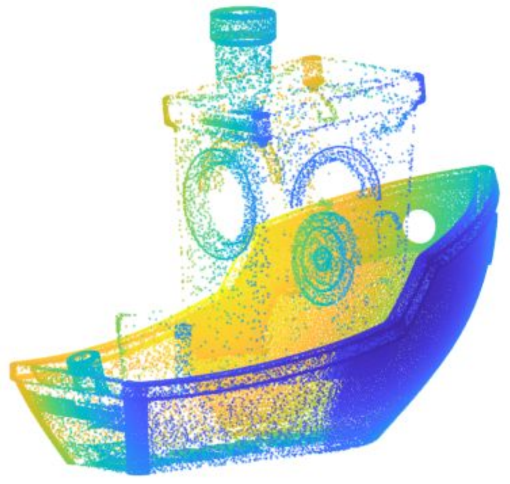
CAD Model

Pose Estimation Problem



Object Model
(CAD etc.)

Sensor Data
of Object
(Point Cloud)



Matching Algorithm

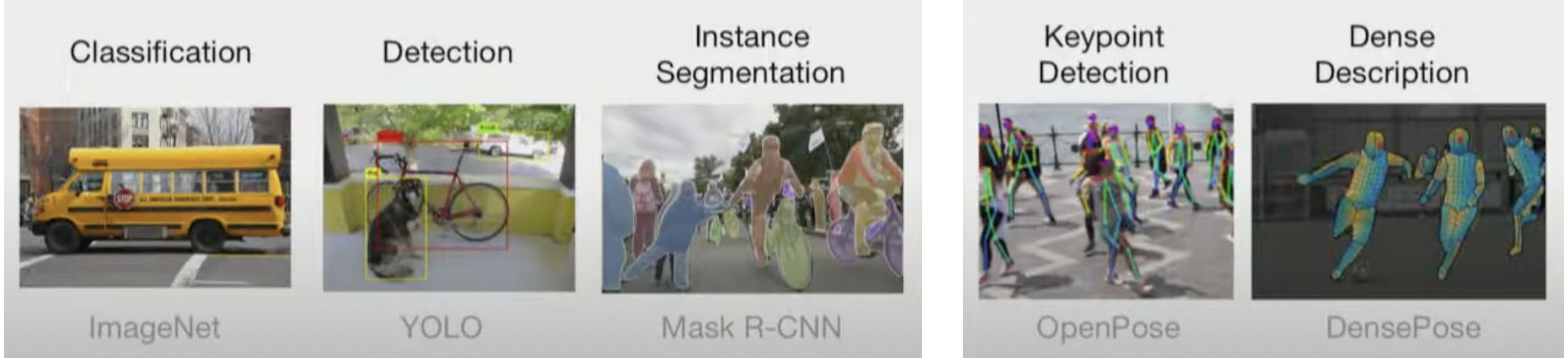
3D Translation
(x,y,z)

3D Rotation
(x,y,z)

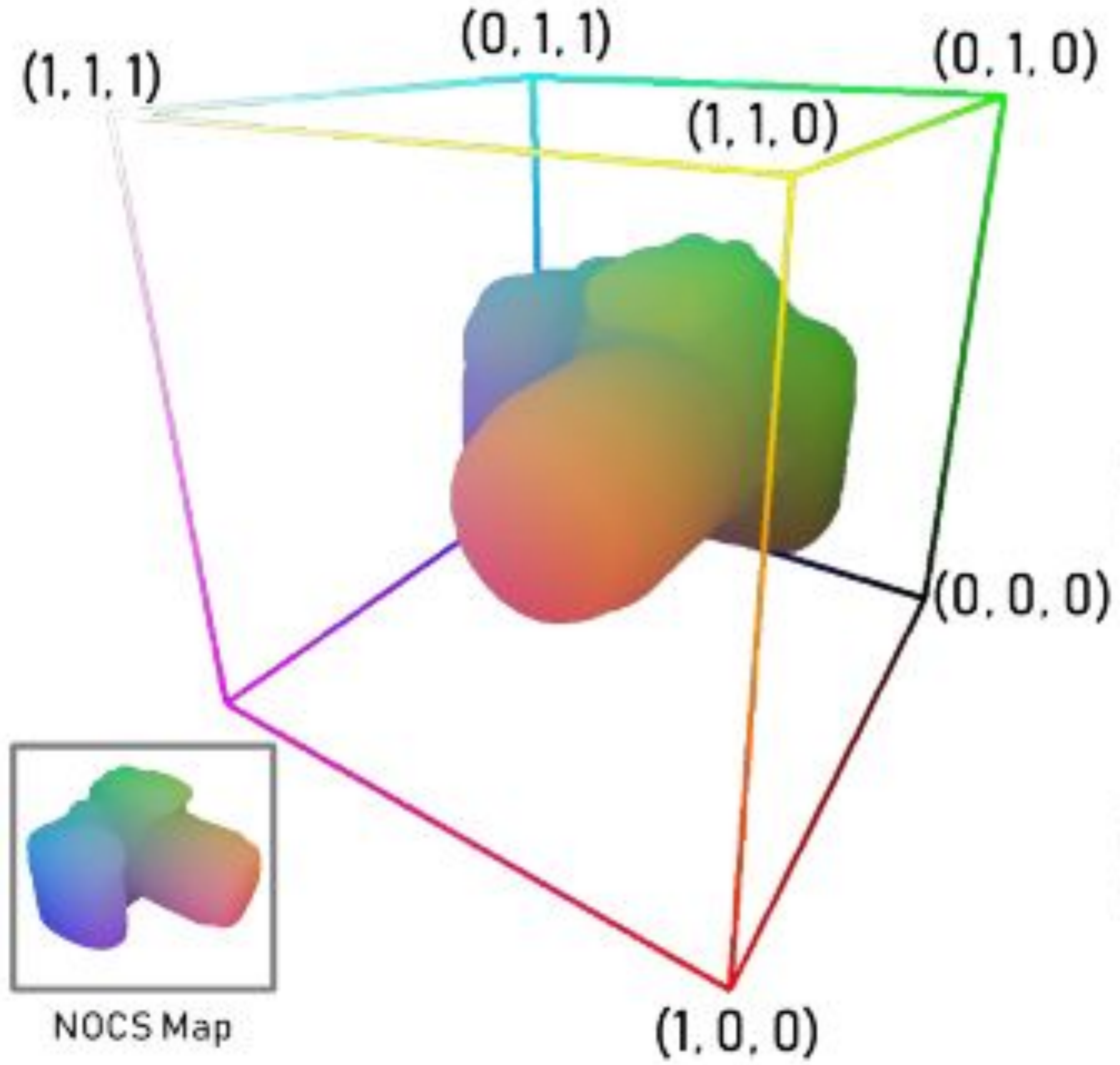
DR

What is a good alternative for the CAD model?

Object Descriptor

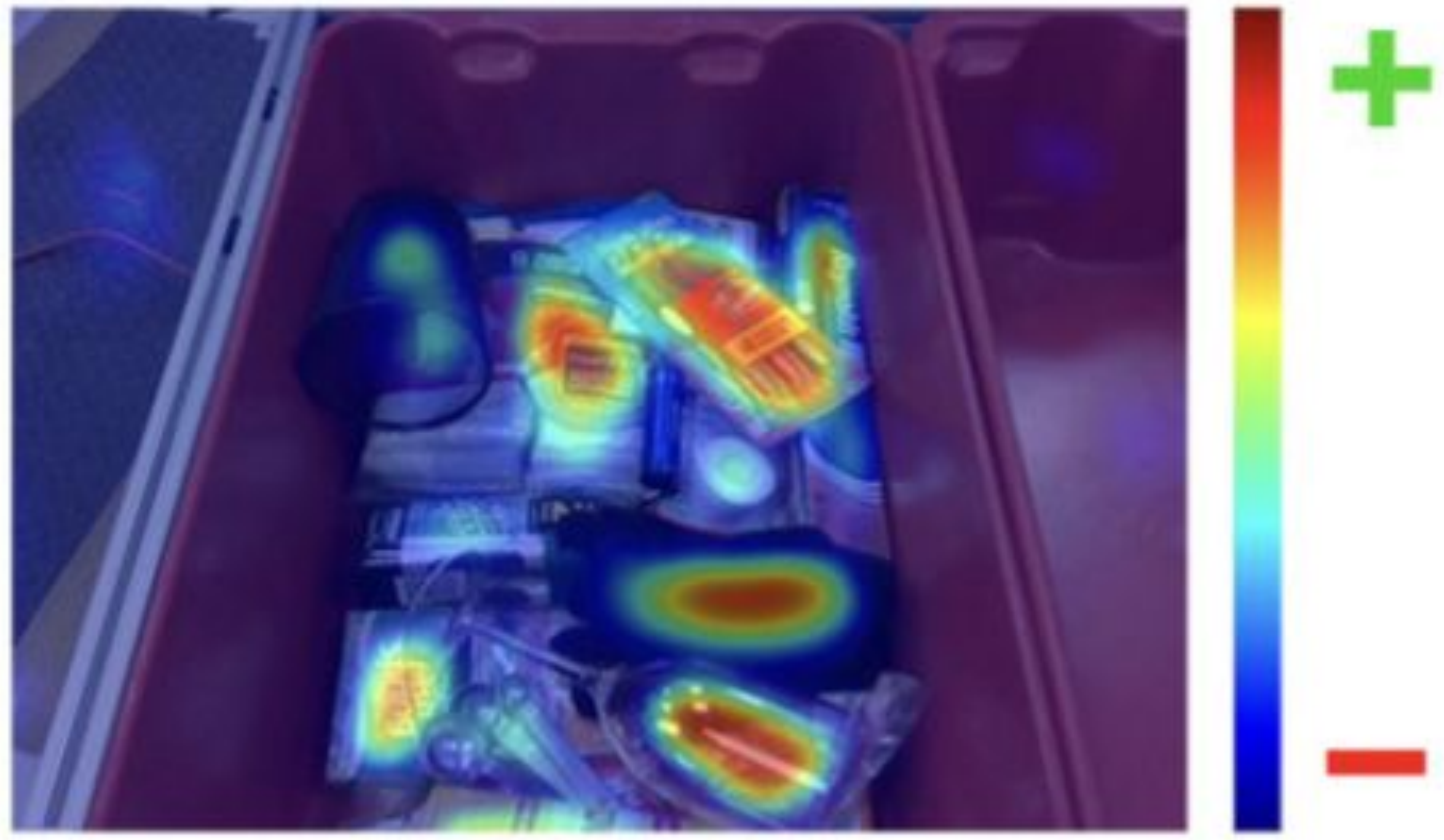


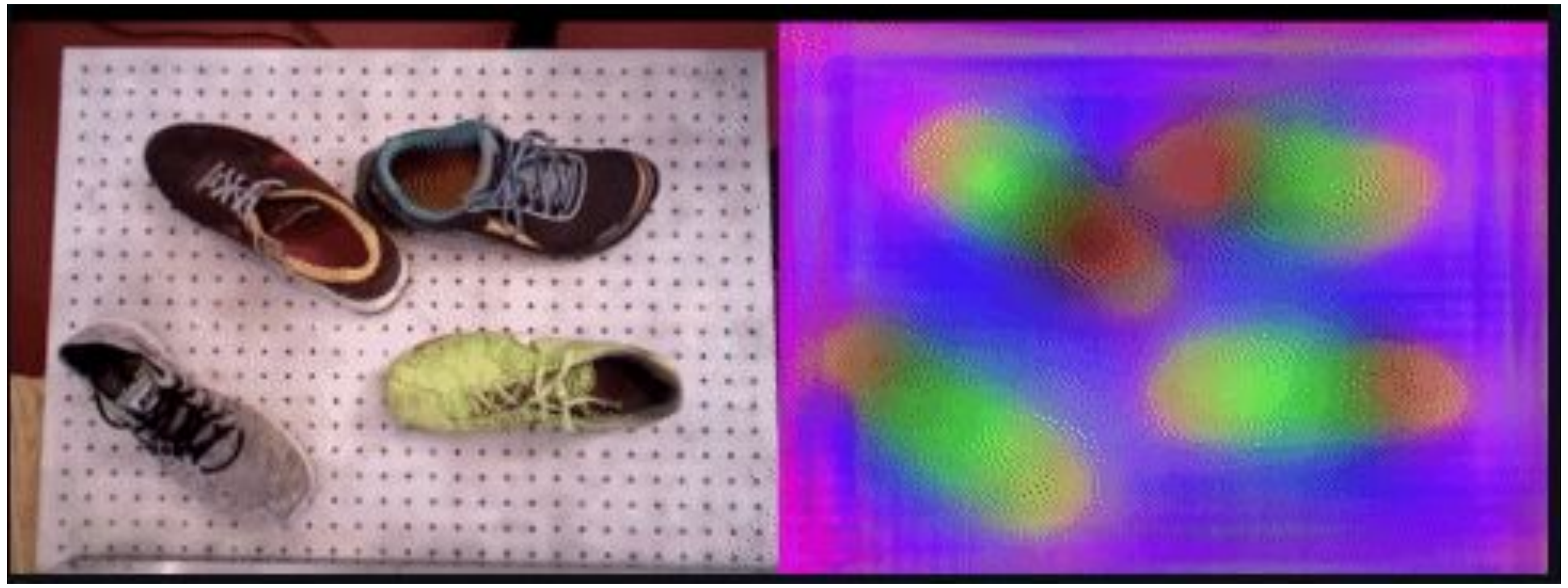
- Dense object descriptors is a normalised way of describing the pose assignment to an object in a category



Object Descriptor

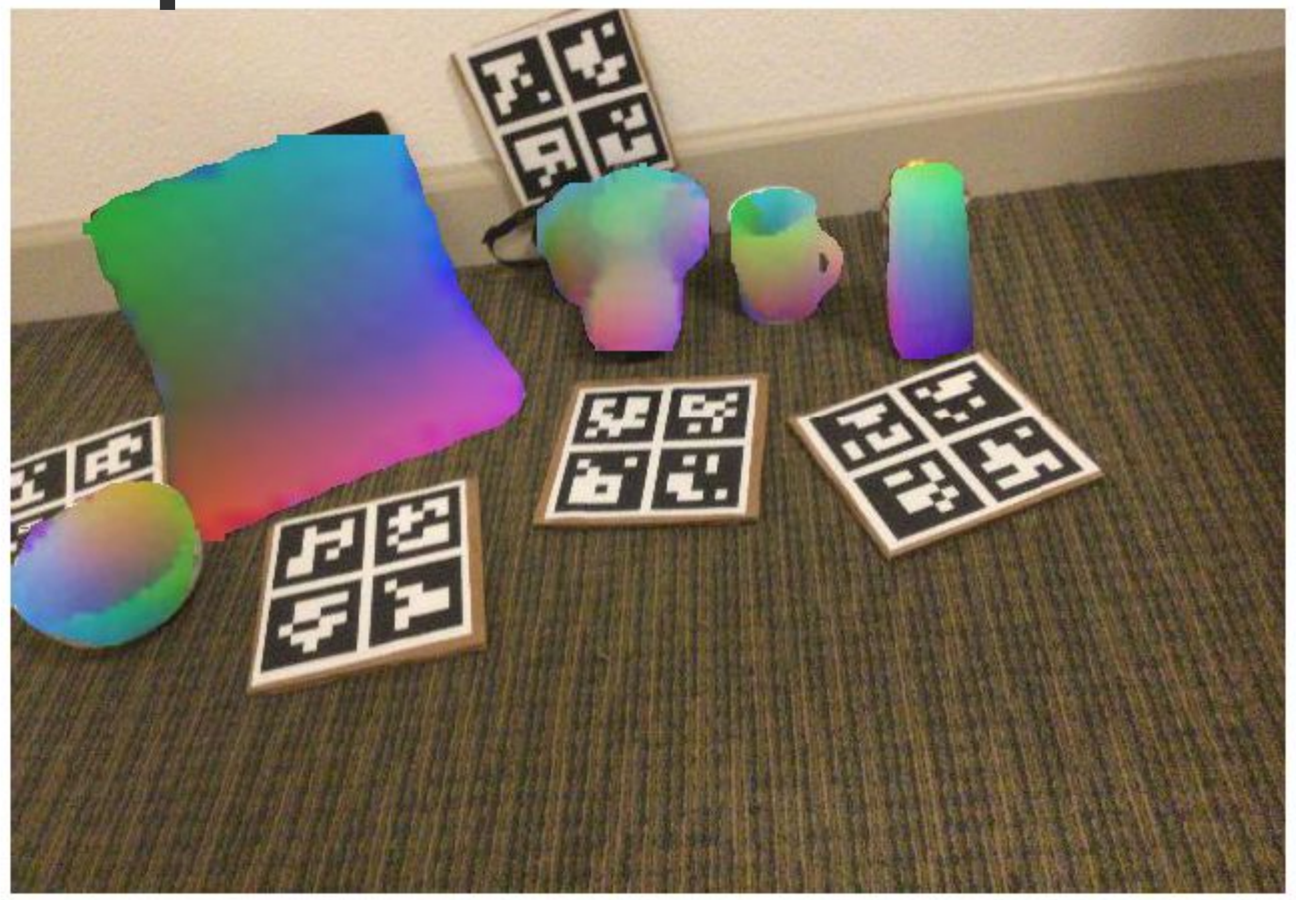
For Grasping



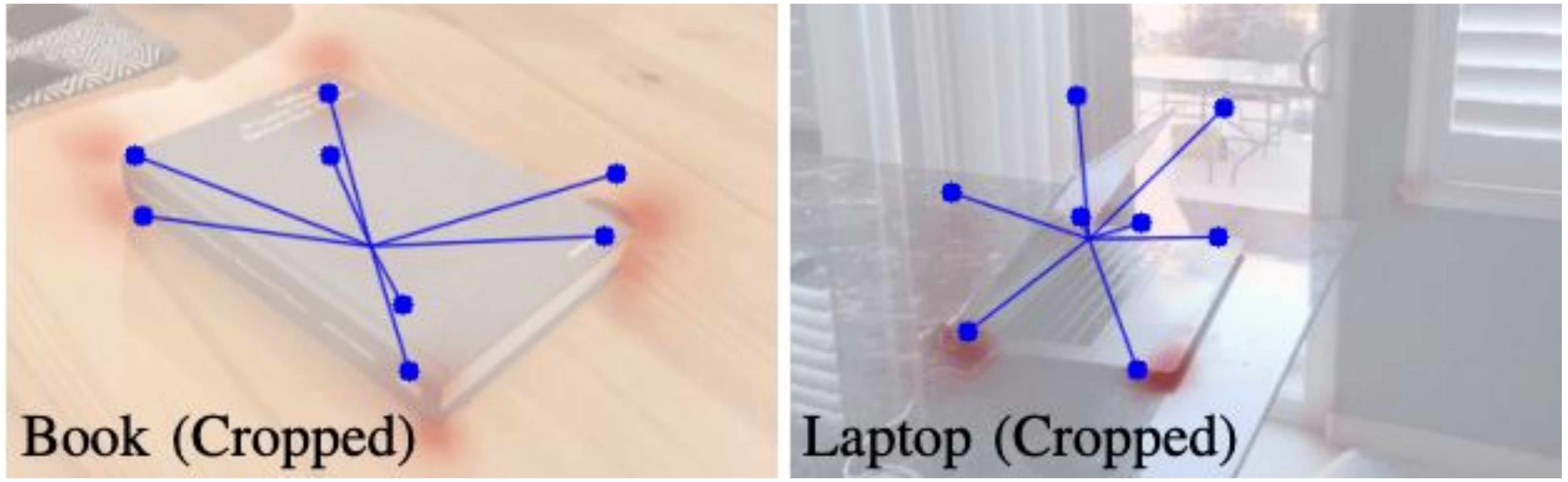


Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation

For pose estimation

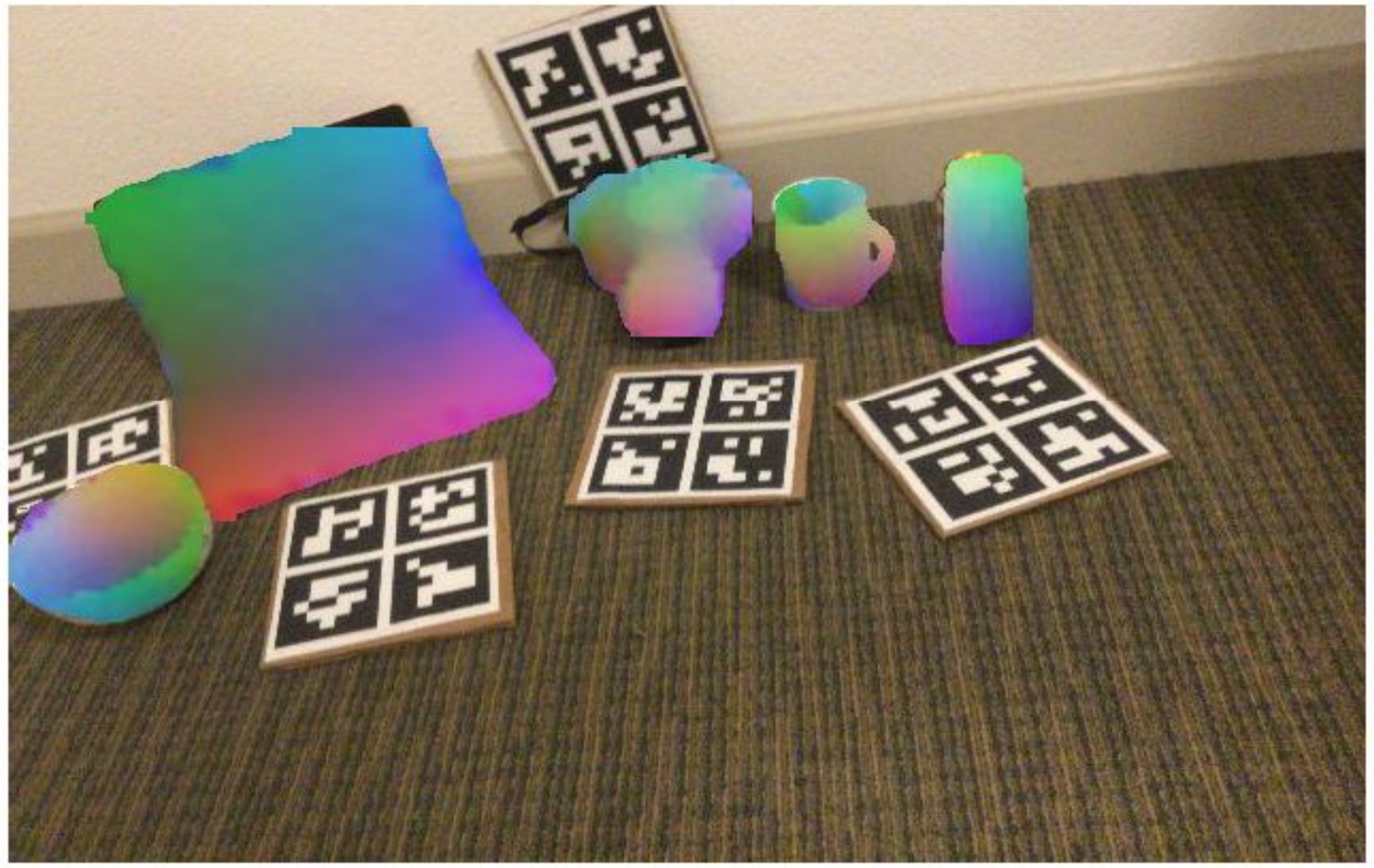


Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation

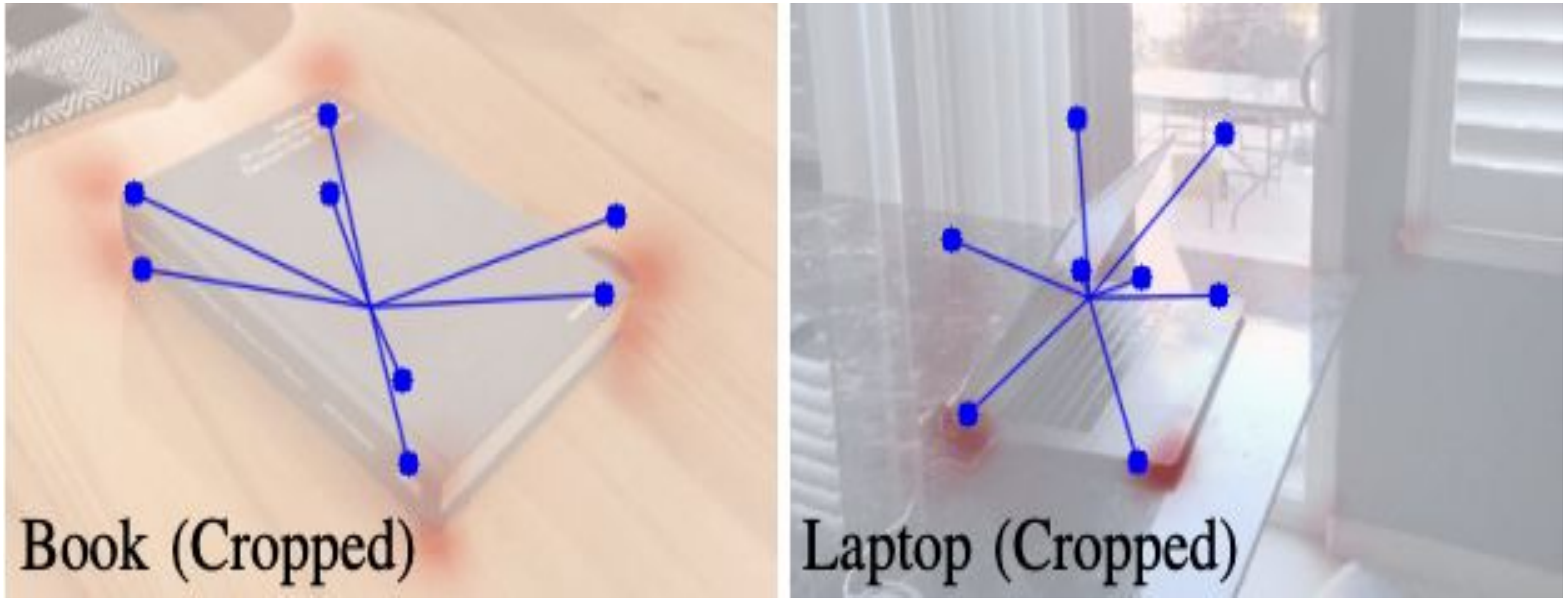


Single-Stage Keypoint-Based Category-Level Object Pose Estimation from an RGB Image

Object Descriptor



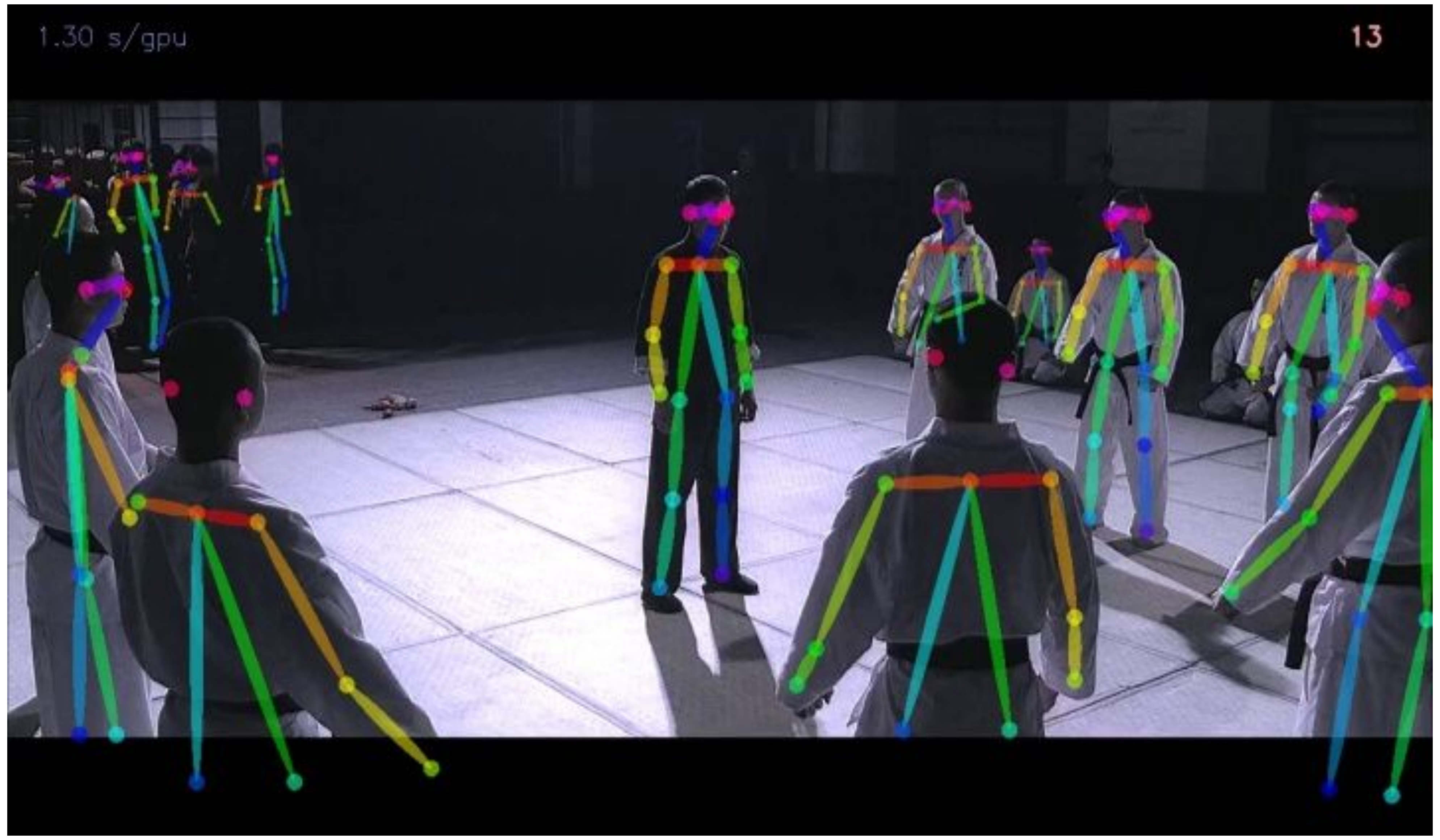
Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation



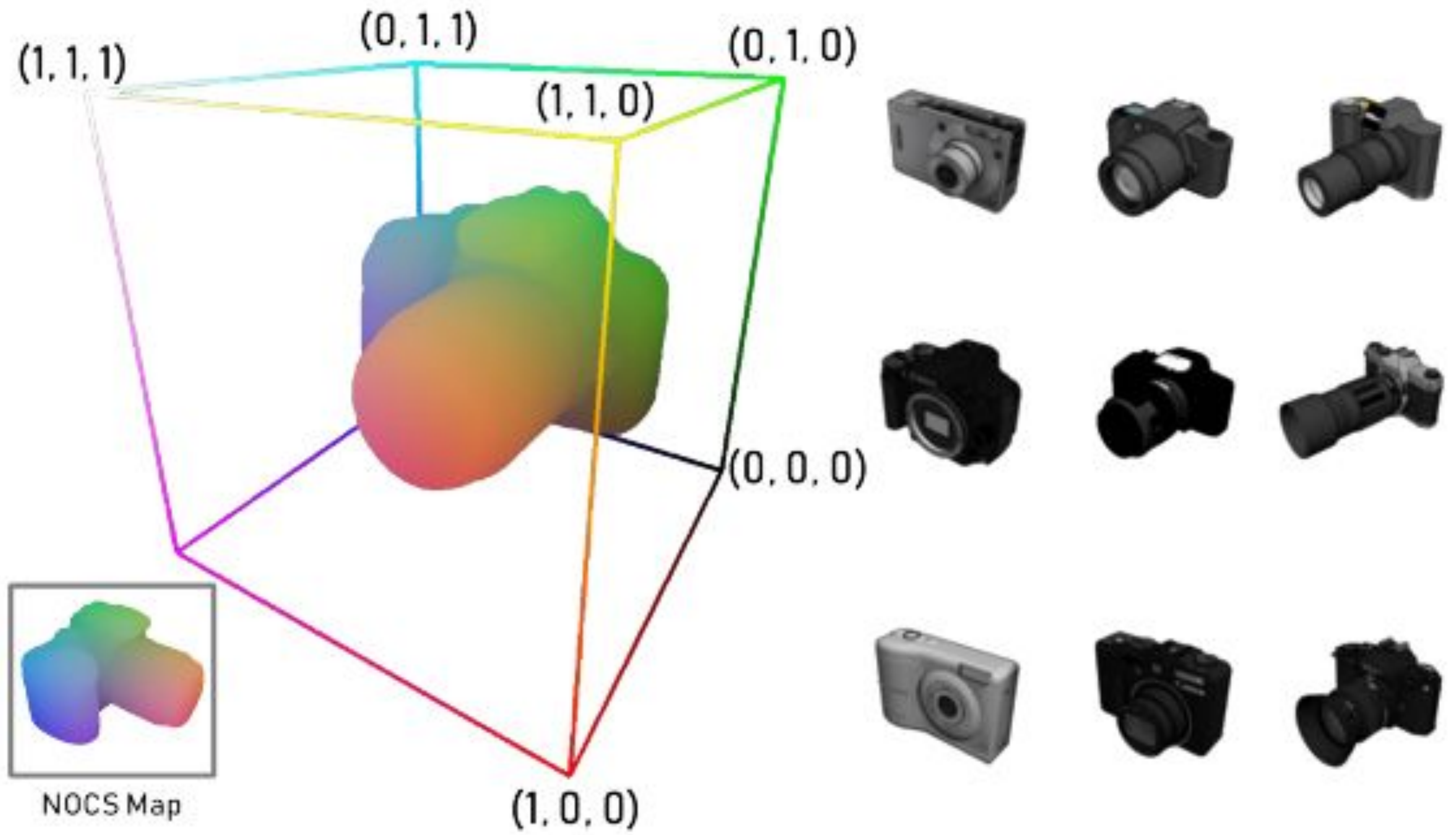
Single-Stage Keypoint-Based Category-Level Object Pose Estimation from an RGB Image

Constraints for Object Descriptors

- Consistent across viewpoints
- Consistent across Object configurations
- Consistent across the object class.



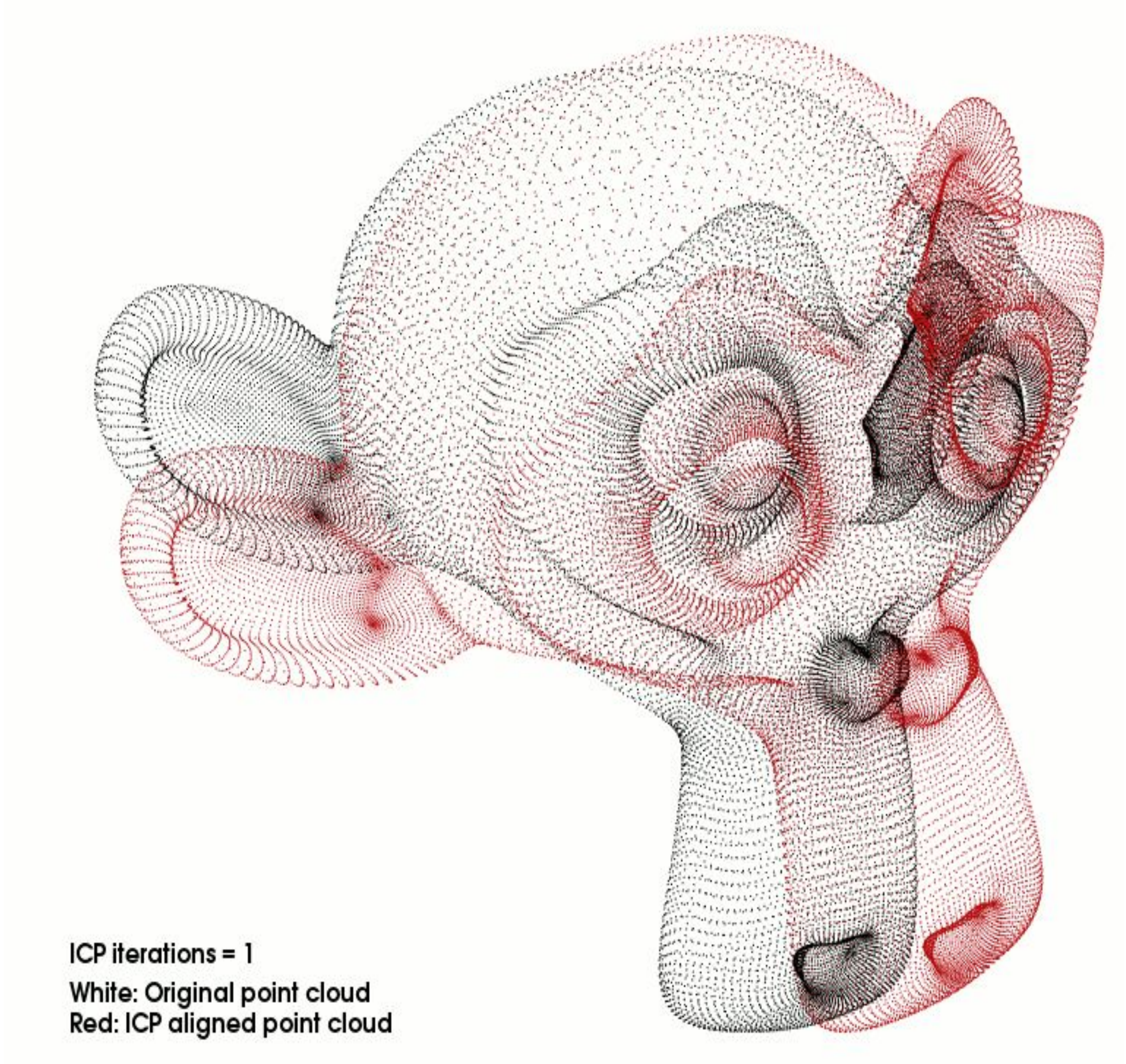
OpenPose



Normalised Object Coordinate Space

Traditional Methods for Pose Estimation

Iterative Closest Point (ICP)



Reference Point Cloud (PC)

Test Point Cloud (PC)

Step 1: Match each point on TestPC to nearest point on Reference PC

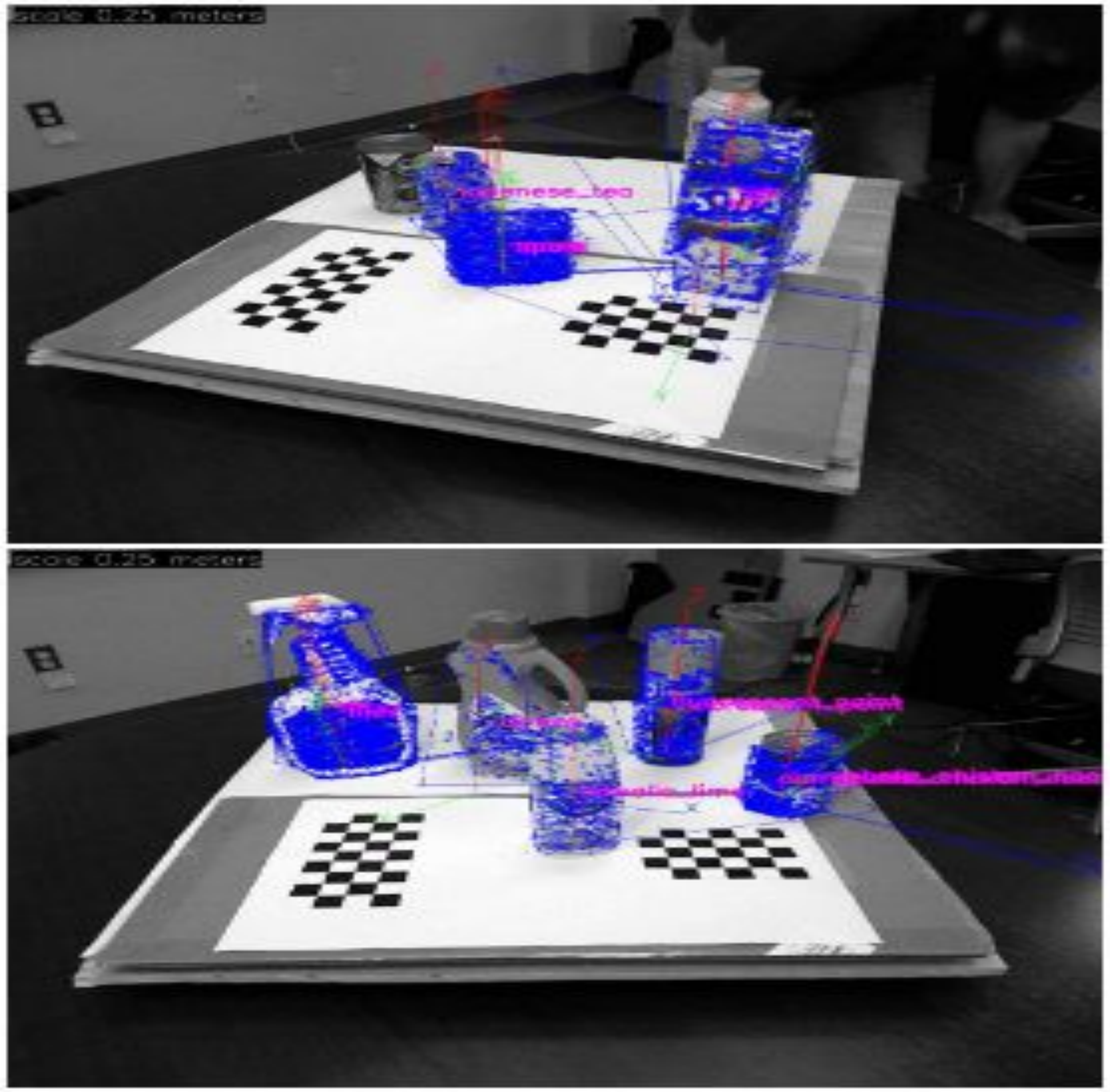
Step 2: Translate TestPC by aligning center of masses, then rotate using SVD

Step 3: If total distance between points < threshold, done. Else repeat

Full Rotation and Translation Matrix (6D Pose)

Point Cloud Library. "Interactive ICP". PCL Tutorials. 2021, https://pcl.readthedocs.io/projects/tutorials/en/latest/interactive_icp.html.

ORB based Pose Estimation



Steps Involved

1) Match features and descriptors to a database of 49 household objects captured under various views using a 2D camera and a Kinect device.

2) In order to establish a match, it is necessary to not only match the descriptors, but also to compute a pose.

3) To obtain an estimate of the pose, we apply the Progressive Sample Consensus and Efficient Perspective-n-Point algorithms.

Rublee, Ethan, Vincent Rabaud, Kurt Konolige, and Gary Bradski. “ORB: An Efficient Alternative to SIFT or SURF.” Proceedings of the IEEE International Conference on Computer Vision, vol. 2, no. 3, 2011, pp. 2564-2571, doi: 10.1109/ICCV.2011.6126544.

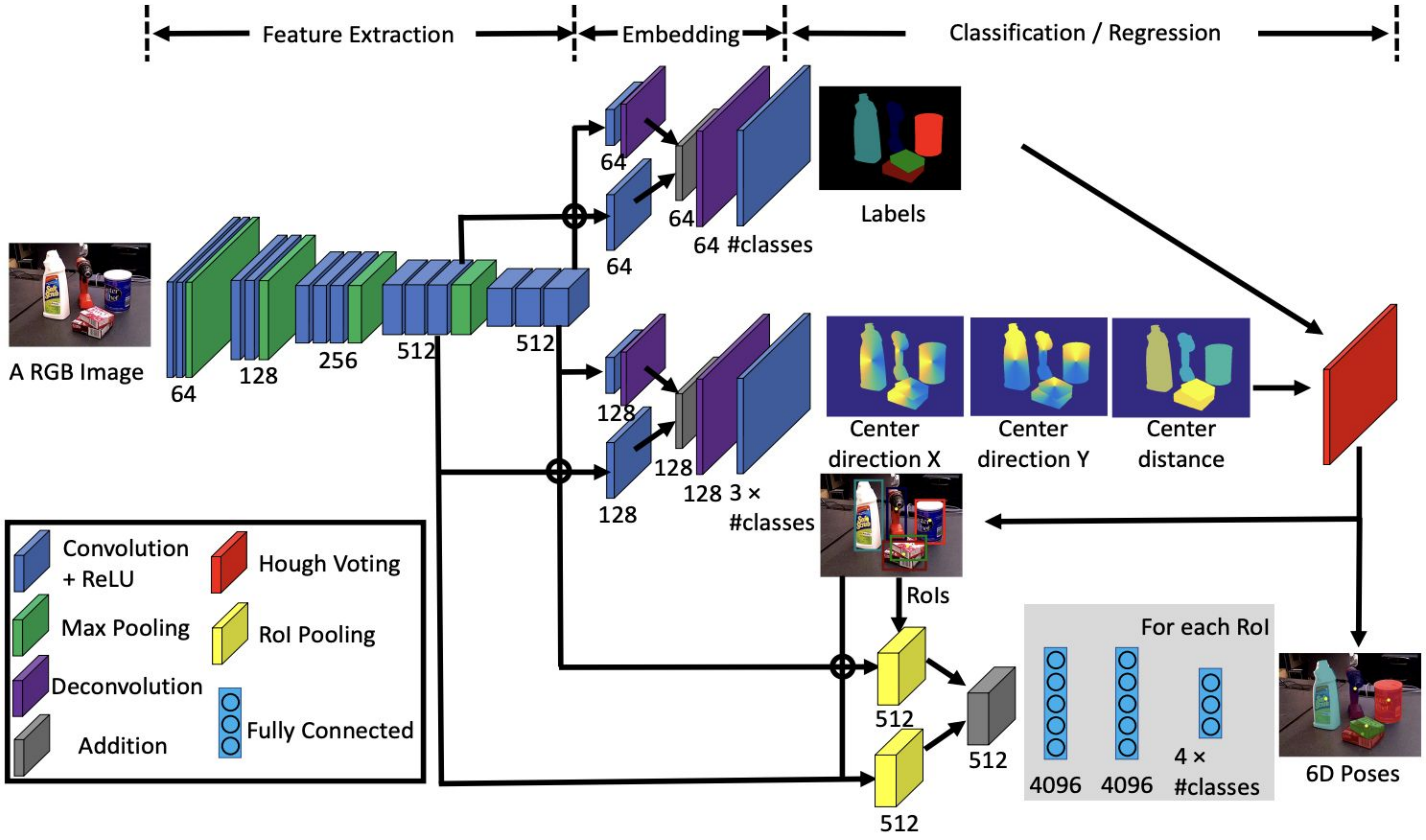
The MOPED framework: Object Recognition and Pose Estimation for Manipulation

Steps Involved

1. **Feature extraction**: SIFT features
2. **Feature matching:** ANN algorithm
3. **Image space clustering**: Mean Shift algorithm
4. **Estimation #1**: RANSAC algorithm and Levenberg-Marquardt optimization
5. **Cluster clustering:** Mean Shift clustering
6. **Estimation #2**: RANSAC and Levenberg-Marquardt optimization
7. **Pose recombination:** Mean Shift and Levenberg-Marquardt optimization

Collet, A., Martinez, M., & Srinivasa, S. S. (2011). The MOPED framework: Object recognition and pose estimation for manipulation. In 2011 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4967-4974). IEEE.

PoseCNN



Xiang, Y., Schmidt, T., Narayanan, V., & Fox, D. (2018). PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018-Jun, 1-10. DOI: 10.1109/CVPR.2018.00016

Datasets

BOP Challenge

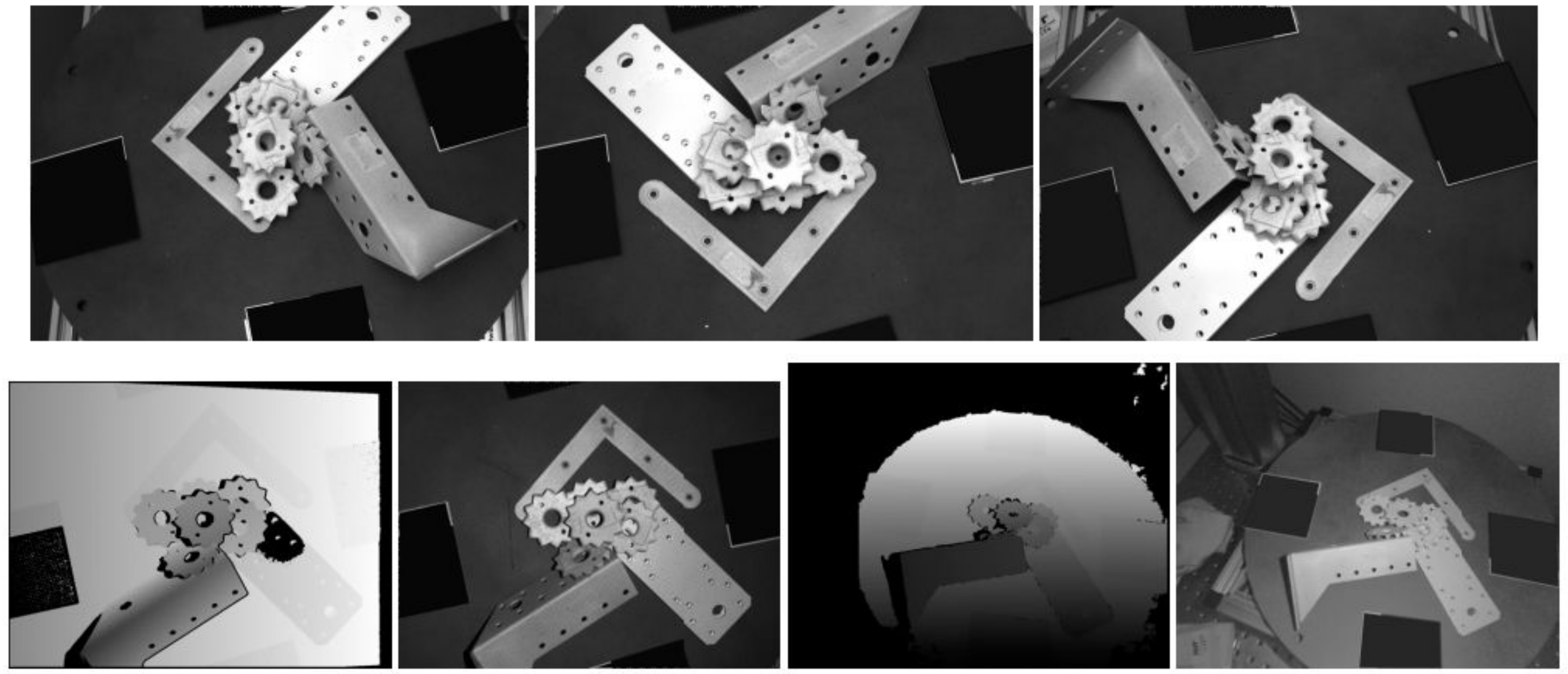
- BOP: Benchmark for 6D Object Pose Estimation
- Small items, focus on manipulation

Dataset Name	Application	Year
Linemod	Texture-less 3D objects, cluttered	2012
HOPE	Household objects	2020
ITODD	Industrial setting objects	2017
RU-APC	Warehouse setting objects	2016
TYO-L (Toyota Light)	Lighting condition variation	2018

https://bop.felk.cvut.cz/home/

BOP Challenge

HOPE



I-TODD

RU-APC

BOP Challenge

Pose estimation (BOP 2019-2022) – Core datasets

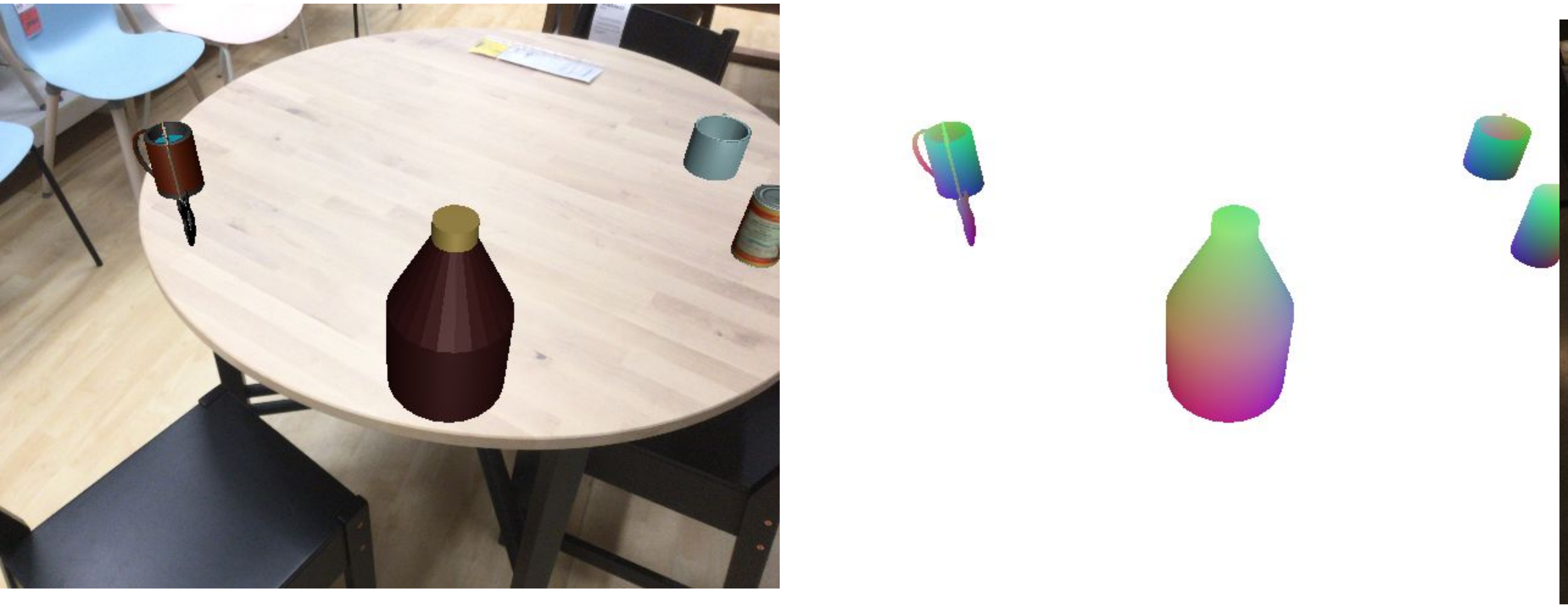
This leaderbord shows the overall ranking on the core datasets (LM-O, T-LESS, TUD-L, IC-BIN, ITODD, HB, YCB-V). For each method, the date of the latest considered submission is reported. If more submissions of a method are available for a dataset, the submission with the highest AR_{Core} score is considered. The performance scores are defined in the BOP Challenge 2019 description. The reported time is the average image processing time averaged over the core datasets.

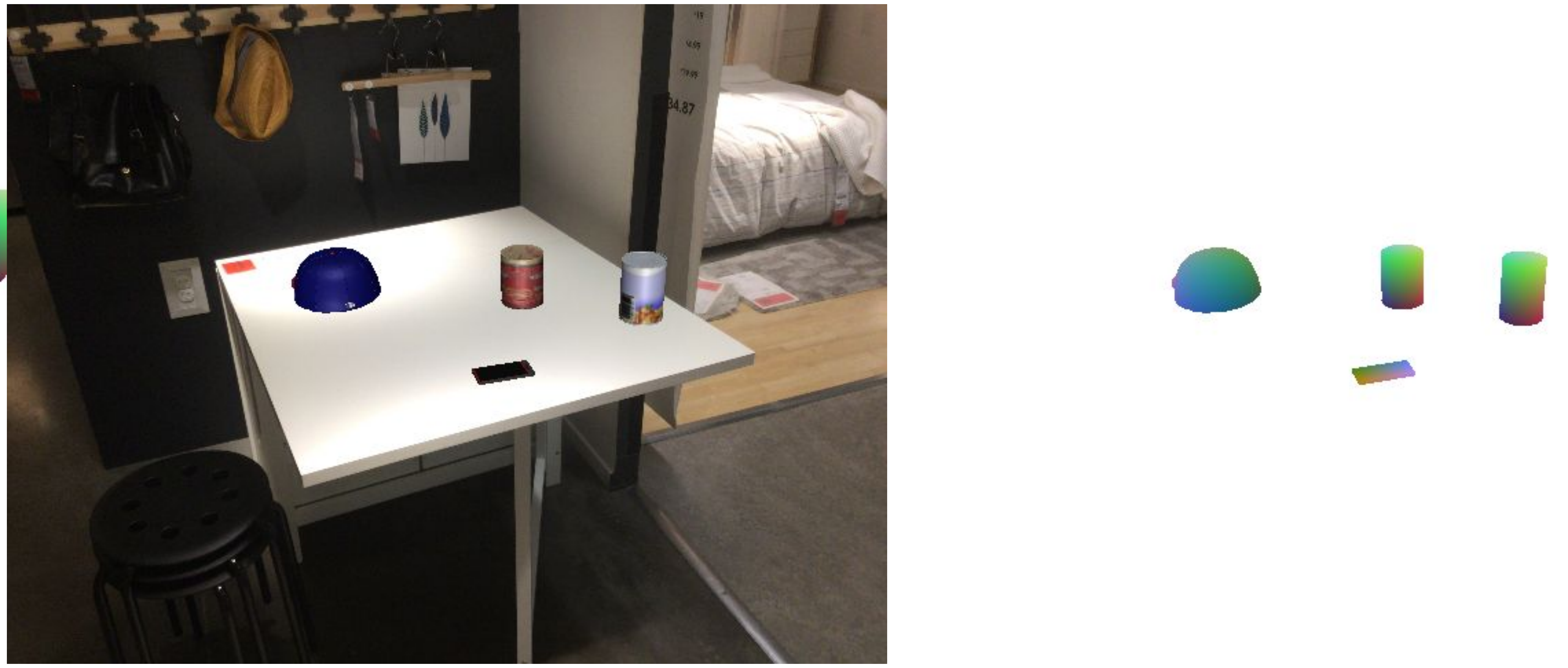
Show 50 entries Search:

	Date (UTC)	Method	Test image	AR_{Core}	AR_{LM-O}	AR_{T-LESS}	AR_{TUD-L}	AR_{IC-BIN}	AR_{ITODD}	AR_{HB}	AR_{YCB-V}	Time (s)
1	2022-10-15	GDRNPP-PBRReal-RGBD-MModel	RGB-D	0.837	0.775	0.874	0.966	0.722	0.679	0.926	0.921	6.263
2	2022-10-15	GDRNPP-PBR-RGBD-MModel	RGB-D	0.827	0.775	0.852	0.929	0.722	0.679	0.926	0.906	6.264
3	2022-10-14	GDRNPP-PBRReal-RGBD-MModel-Fast	RGB-D	0.805	0.792	0.872	0.936	0.702	0.588	0.909	0.834	0.228
4	2022-10-13	GDRNPP-PBRReal-RGBD-MModel-OfficialDet	RGB-D	0.798	0.758	0.824	0.966	0.708	0.543	0.890	0.896	6.406
5	2022-10-11	Extended FCOS+PFA-MixPBR-RGBD	RGB-D	0.787	0.797	0.850	0.960	0.676	0.469	0.869	0.888	2.317
6	2022-10-12	Extended FCOS+PFA-MixPBR-RGBD-Fast	RGB-D	0.771	0.792	0.779	0.958	0.671	0.460	0.860	0.880	0.639
7	2022-10-16	RCVPose 3D_SingleModel_VIVO_PBR	RGB-D	0.768	0.729	0.708	0.966	0.733	0.536	0.863	0.843	1.336
8	2022-10-15	ZebraPoseSAT-EffnetB4 + ICP (DefaultD...	RGB-D	0.765	0.752	0.727	0.948	0.652	0.527	0.883	0.866	0.500
9	2022-10-12	Extended FCOS+PFA-PBR-RGBD	RGB-D	0.762	0.797	0.802	0.893	0.676	0.469	0.869	0.826	2.631
10	2021-12-22	SurfEmb-PBR-RGBD	RGB-D	0.758	0.760	0.828	0.854	0.659	0.538	0.866	0.799	9.048

Context Aware Mixed Reality (CAMERA)

- Combines real background (tabletop) with synthetic objects for efficient data generation
- 6 object categories from ShapeNetCore: bottle, bowl, camera, can, laptop, and mug
  - 1085 object instances, 184 set aside for validation
- Distractor categories for robustness (phone , guitar, etc.)
- 300k images, 25k set aside for validation





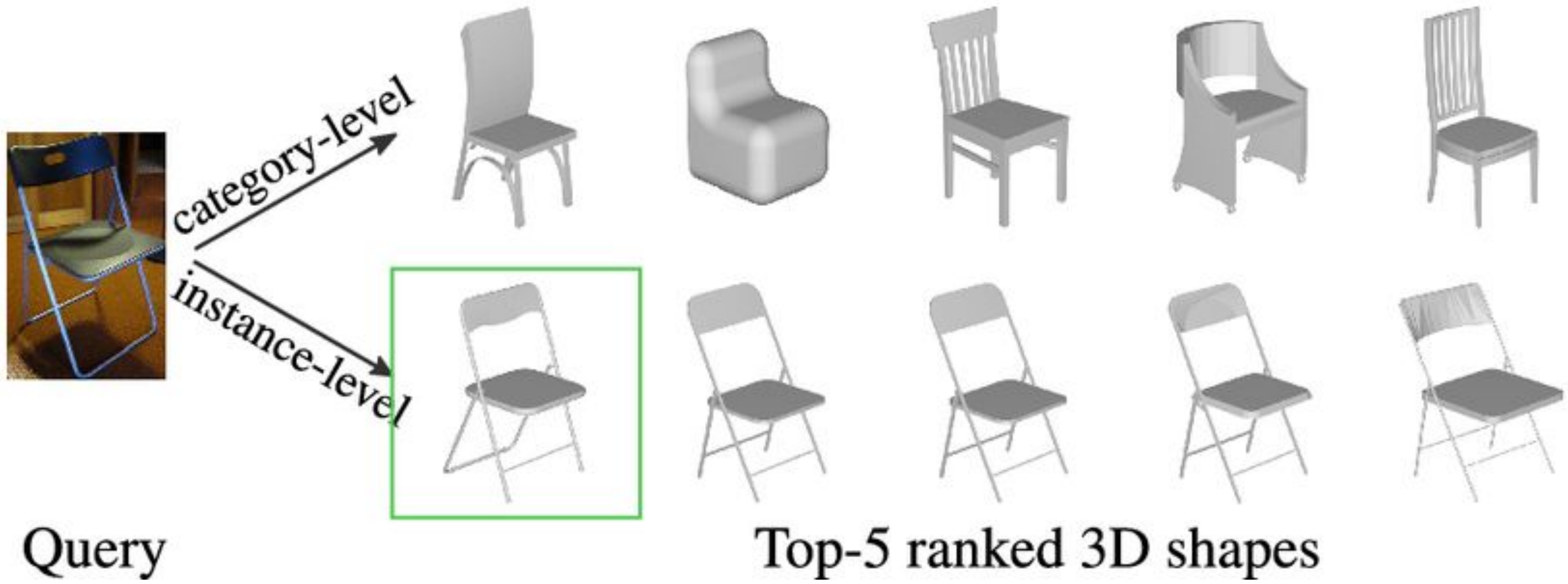
Chang, Angel X., et al. "Shapenet: An information-rich 3d model repository." *arXiv preprint arXiv:1512.03012* (2015).

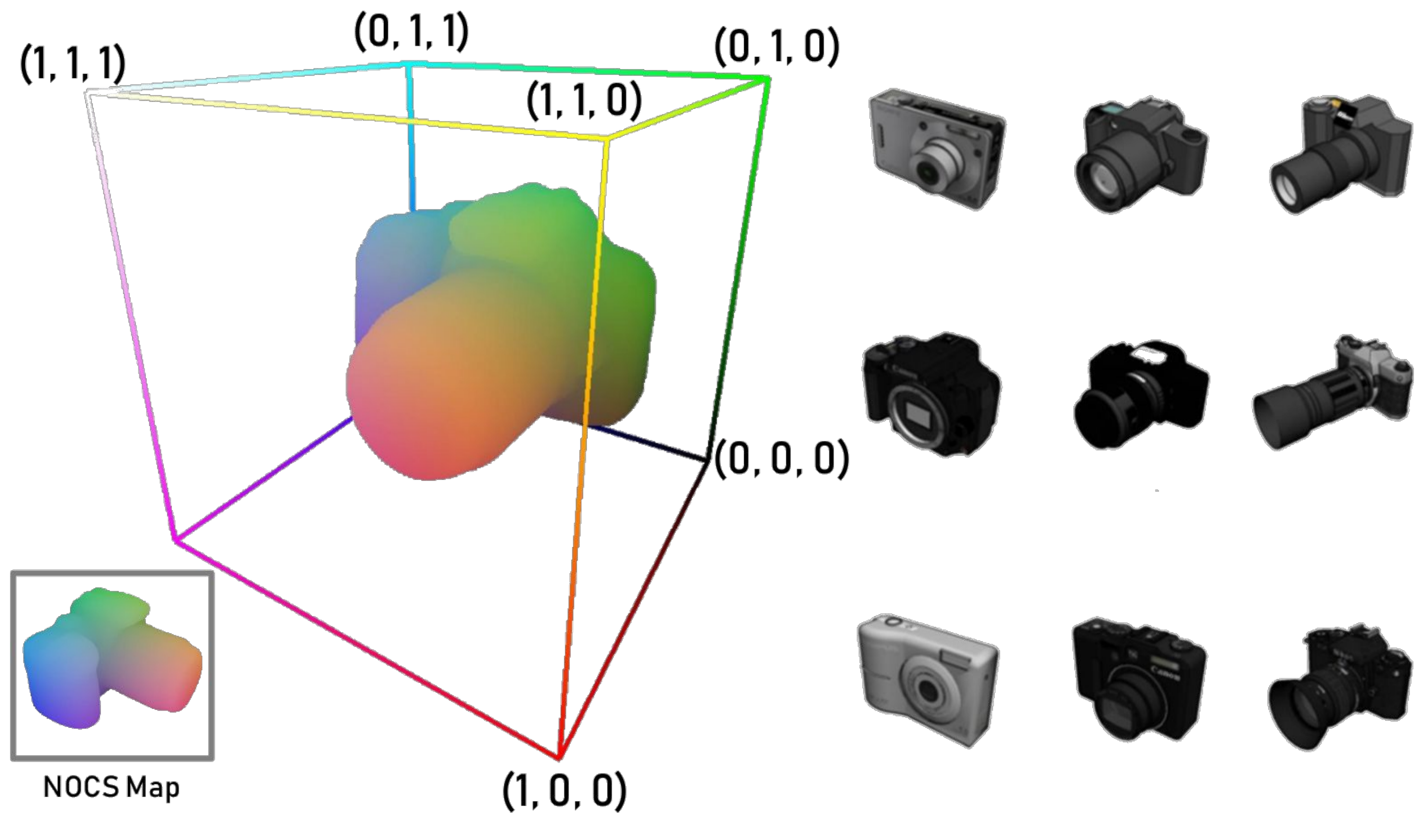
Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, Leonidas J. Guibas

Computer Vision and Pattern Recognition Conference 2019

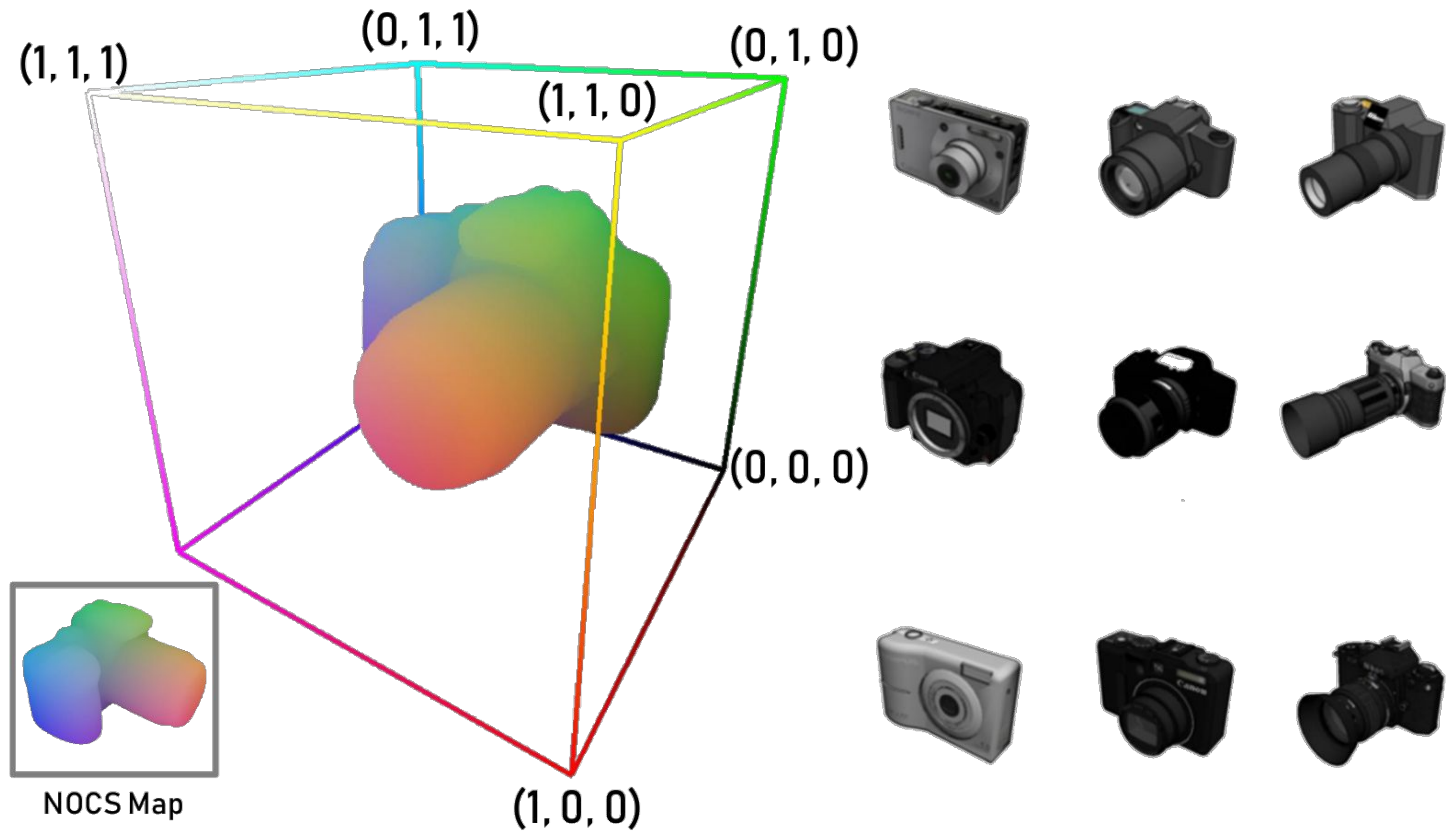
Category vs Instance level representation





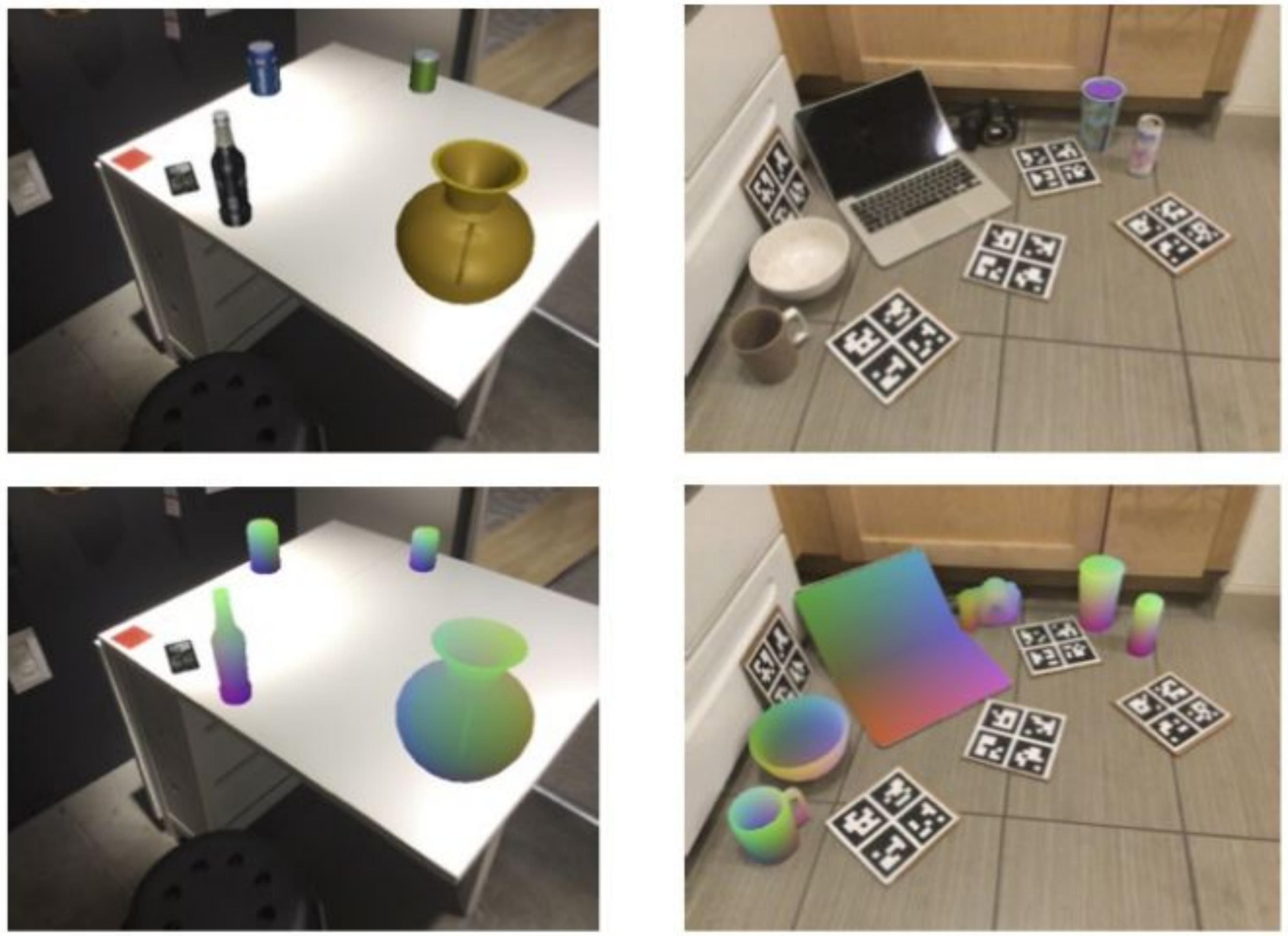
Zou, Qian-Fang, Ligang Liu, and Yang Liu. "Instance-level 3D shape retrieval from a single image by hybrid-representation-assisted joint embedding." *The Visual Computer* 37 (2021): 1743-1756.

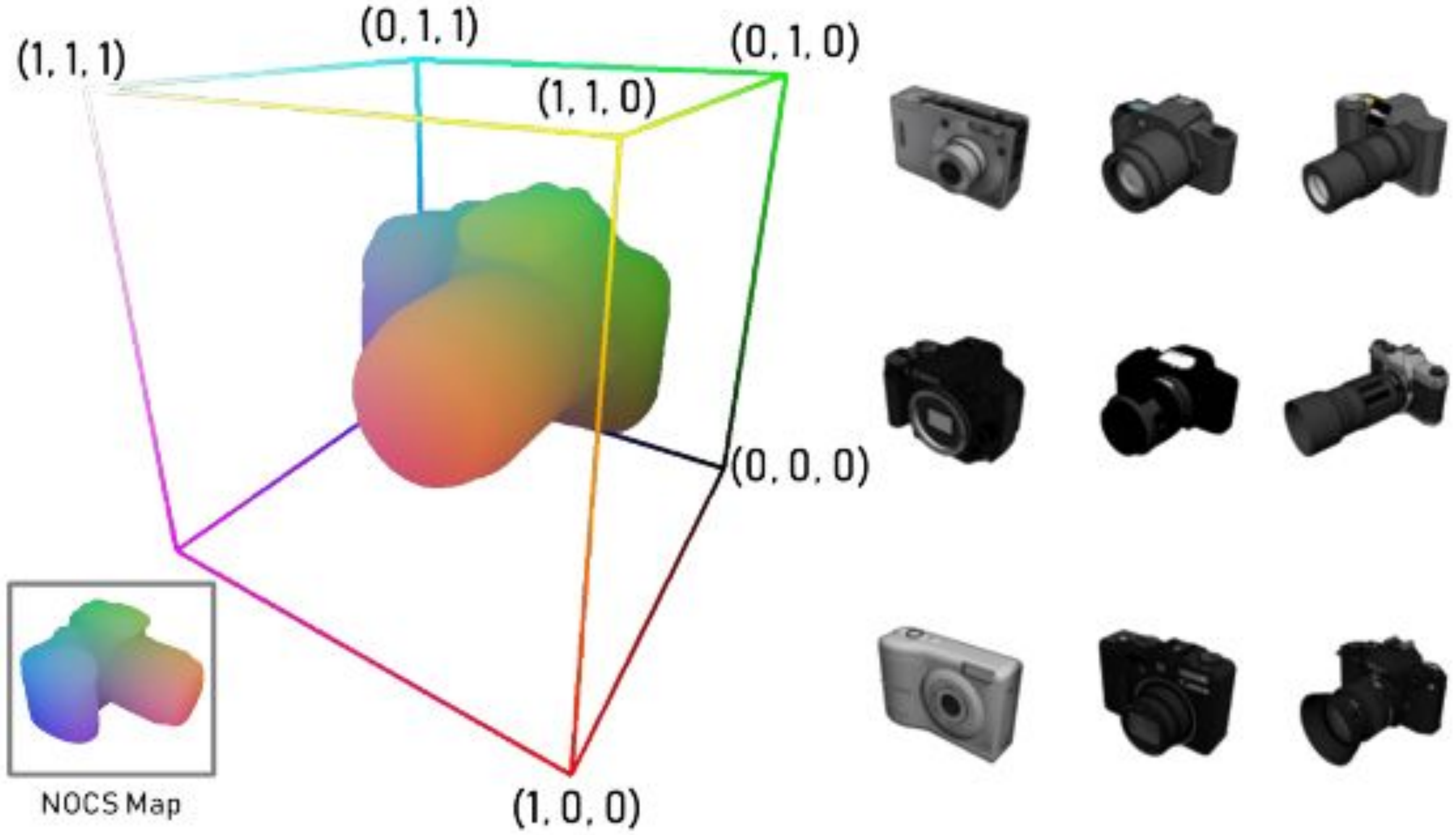
Category vs Instance level representation



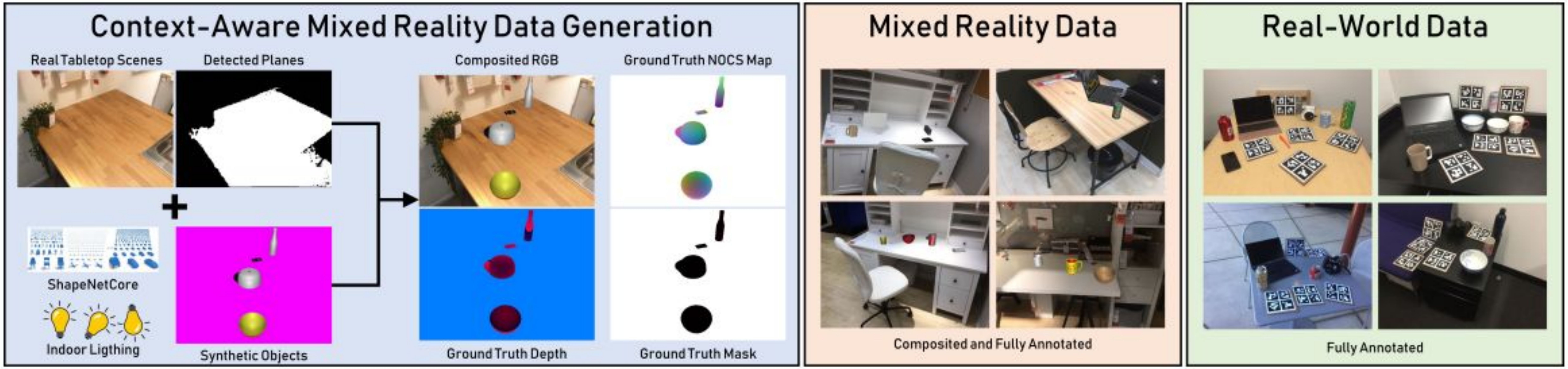
Zou, Qian-Fang, Ligang Liu, and Yang Liu. "Instance-level 3D shape retrieval from a single image by hybrid-representation-assisted joint embedding." *The Visual Computer* 37 (2021): 1743-1756.

Normalized Object Coordinate Space

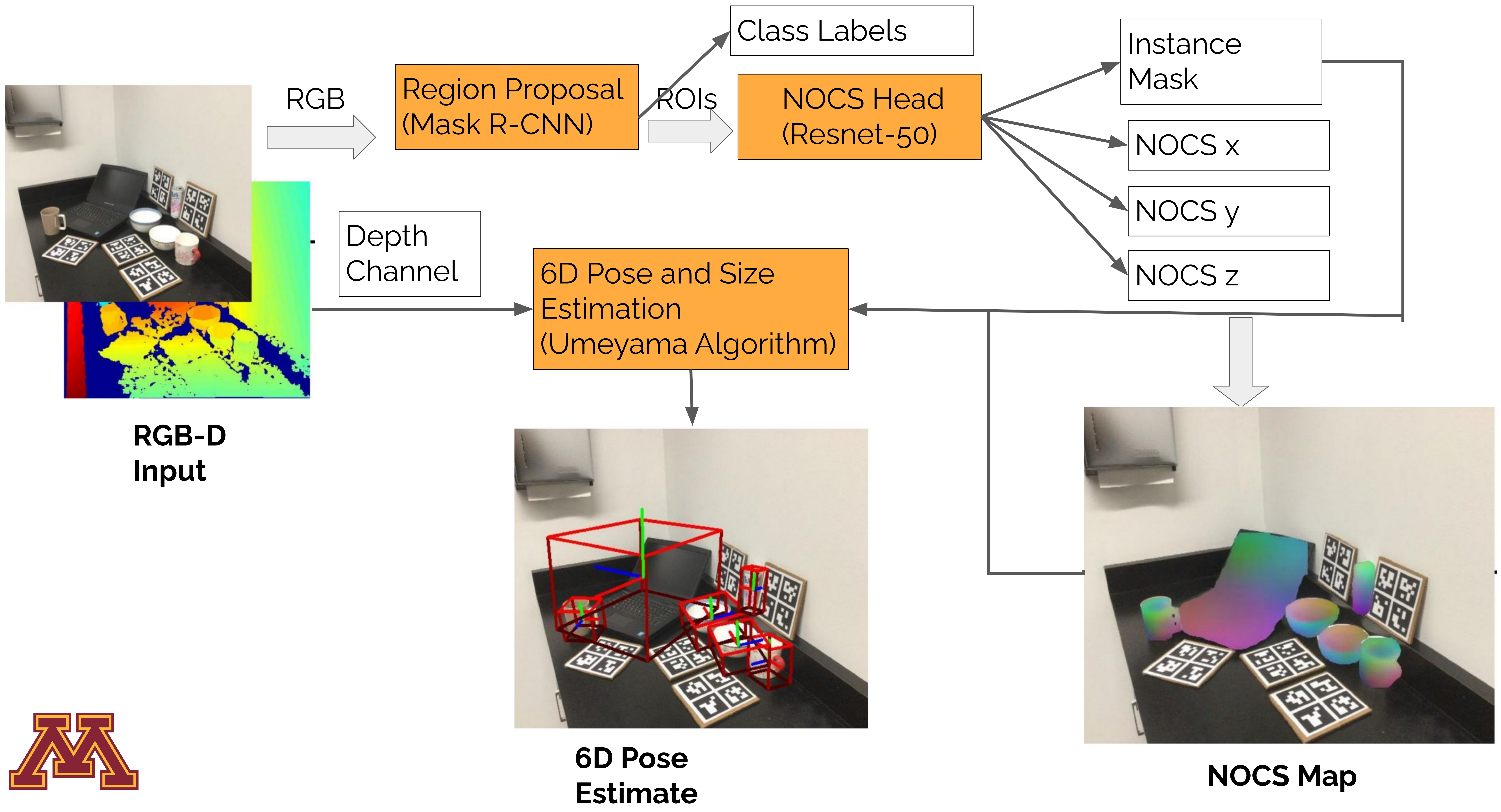




Dataset


Context-Aware Mixed Reality Data Generation
Real Tabletop Scenes
Detected Planes
Composited RGB
Ground Truth NOCS Map
ShapeNetCore
Indoor Ligthing
Synthetic Objects
Ground Truth Depth
Ground Truth Mask
Mixed Reality Data
Composited and Fully Annotated
Real-World Data
Fully Annotated

Model Architecture



Model Training

Input: RGB images, GT NOCS maps.

Initialize ResNet50 backbones, with weights trained on the COCO dataset for 2D instance segmentation.

Set batch size, learning rate and use SGD optimizer.

ResNet 50

Stage 1: Weights are frozen and only the layers in the heads, the RPN, and FPN trained for 10K iterations. LR decreased by 10x

Stage 2: Freeze below level 4, train for 3K iterations. LR decreased by 10x

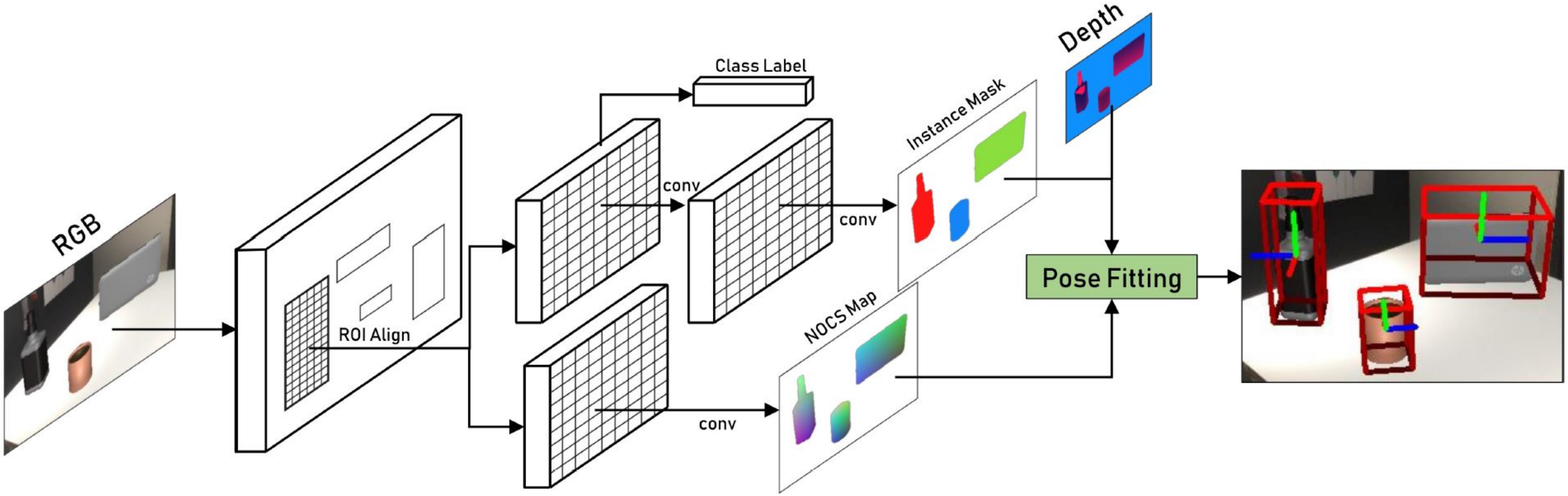
Stage 3: Freeze below level 3 and train for 70K iterations

Output: class, bbox, mask, and NOCS map

Two loss functions for NOCS map heads: softmax and soft L1

Symmetric loss is found between GT NOCS map and predicted NOCS map

Model Training



L1 Loss

$$L(\mathbf{y}, \mathbf{y}^*) = \frac{1}{n} \begin{cases} 5\,(\mathbf{y} - \mathbf{y}^*)^2, & |\mathbf{y} - \mathbf{y}^*| \leq 0.1 \\ |\mathbf{y} - \mathbf{y}^*| - 0.05, & |\mathbf{y} - \mathbf{y}^*| > 0.1 \end{cases},$$

$$\forall \mathbf{y} \in N, \mathbf{y}^* \in N_p,$$

y - ground truth NOCS map pixel value
y* - predicted NOCS map pixel value,
n - number of mask pixels in ROI.

Softmax Loss

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log\left(\frac{e^{f_j(x_i)}}{\sum_{k=1}^{C} e^{f_k(x_i)}}\right)$$

N - number of samples
C - number of classes
xi - i-th input sample
fj (xi) - score of the j-th class for the i-th sample
yij - 1 if true label of i =j , 0 otherwise

Symmetric loss function

$$L_s = \min_{i=1,\ldots,|\theta|} L(\tilde{y_i}, y^*)$$

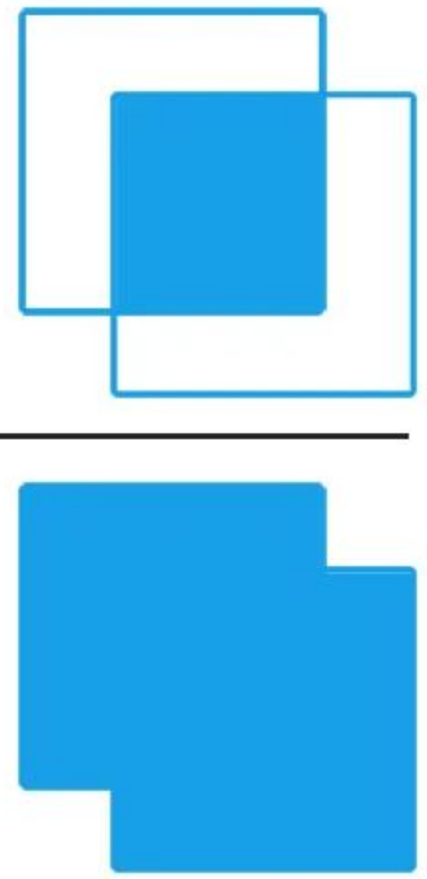
y - ground truth NOCS map pixel value
y* - predicted NOCS map pixel value
|θ| - angle to rotate the NOCS maps along the symmetry axis

Evaluation Metrics

- 3D detection and object dimension estimation
  - mAP at IoU at 25% and 50% threshold

- 6D Pose estimation
  - Average precision of object instances for which the error is less than m=5,10 cm for translation and n = 5°,10° for rotation

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

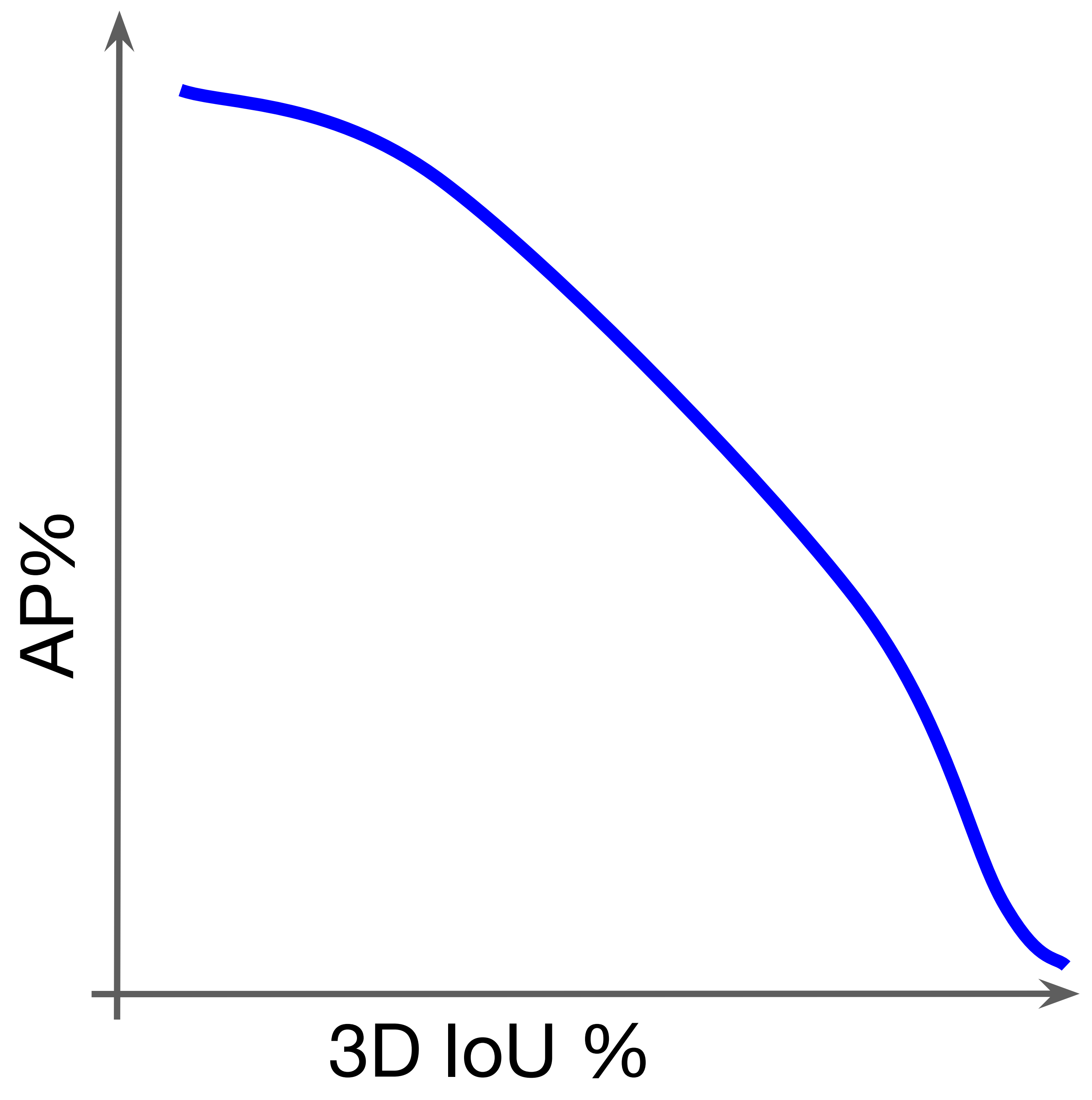


$$Precision = \frac{TP}{TP + FP}$$

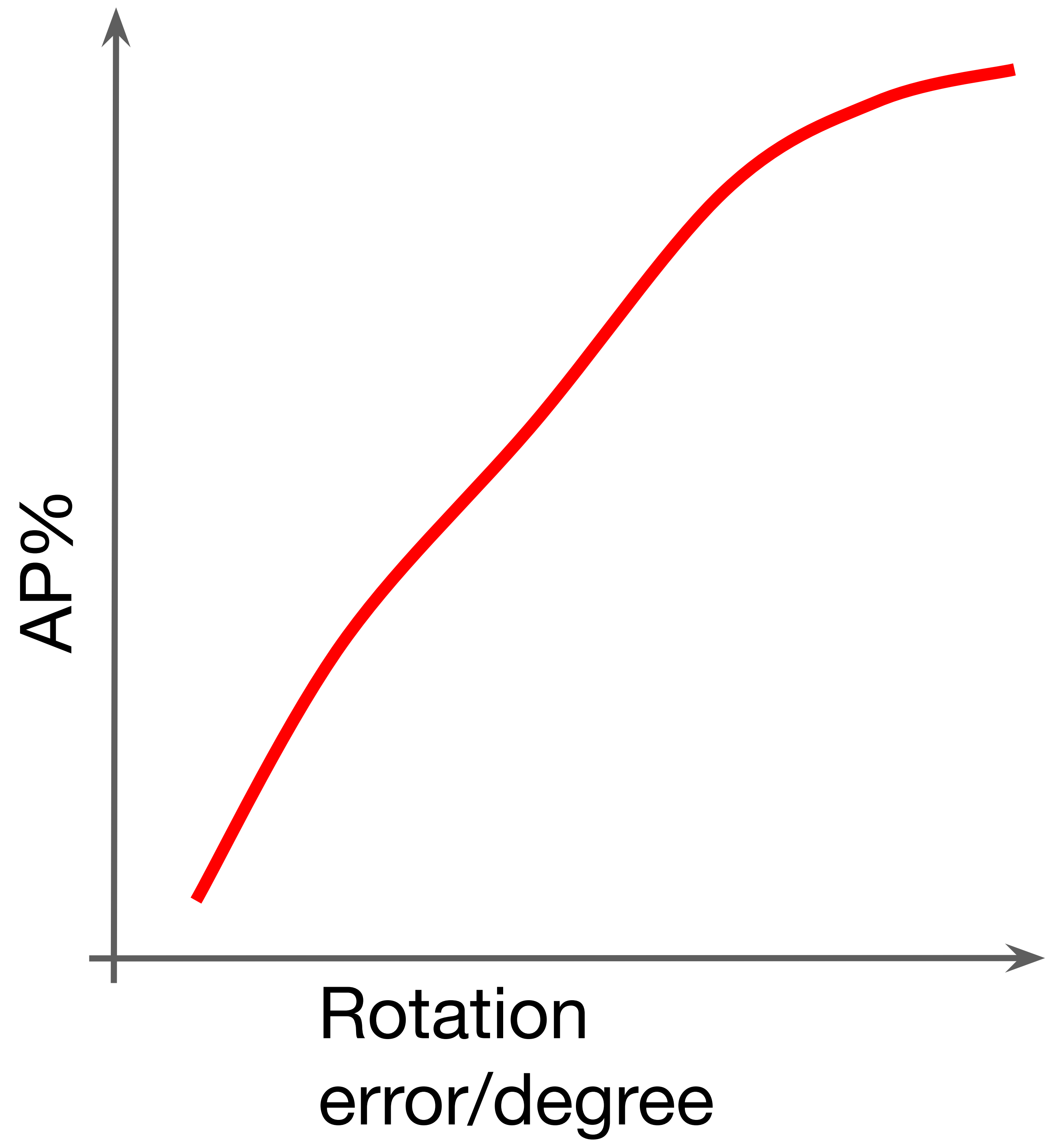
$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP(q)}}{Q}$$

Results: Hypothesis

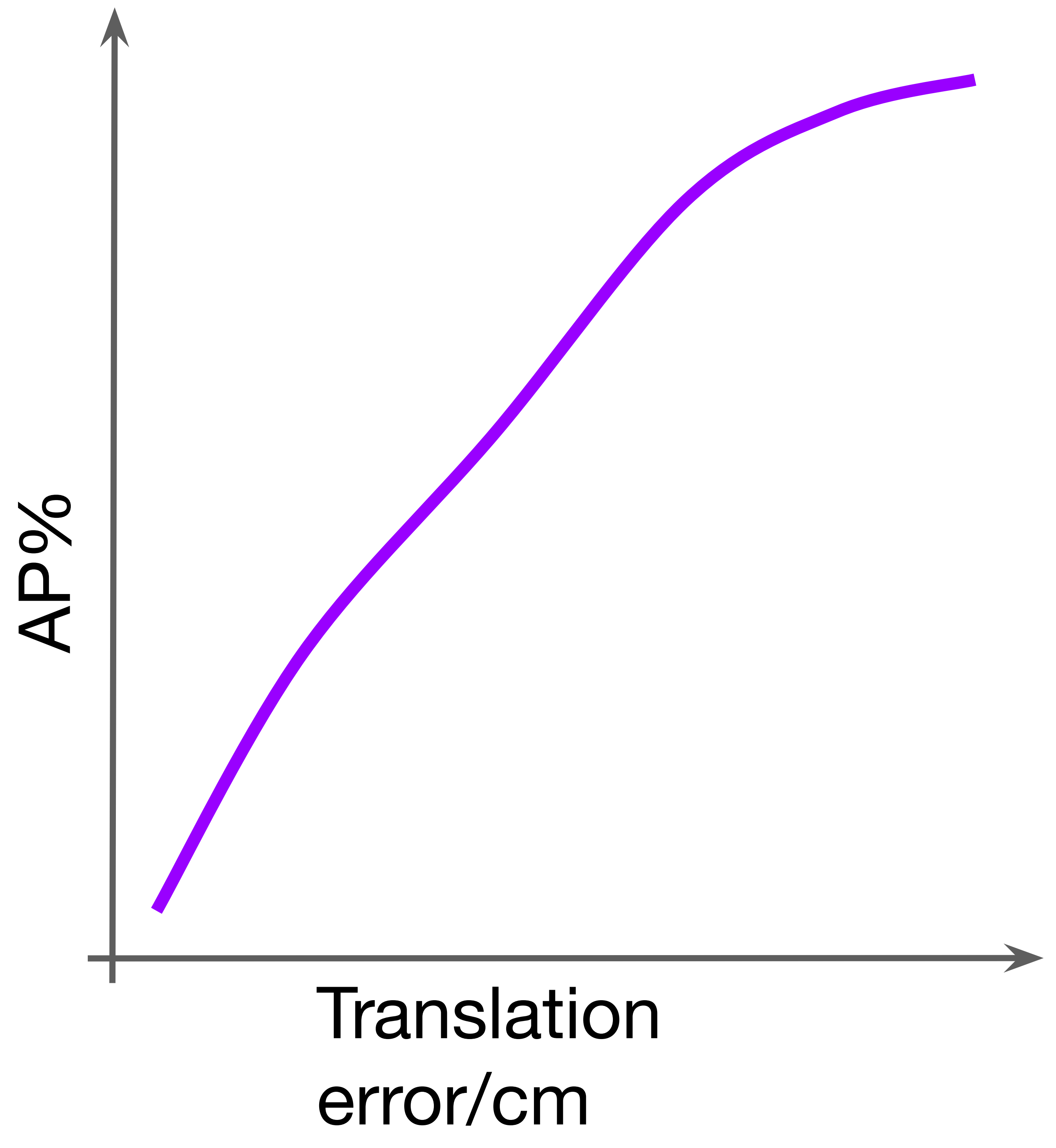
3D IoU AP



Rotation AP

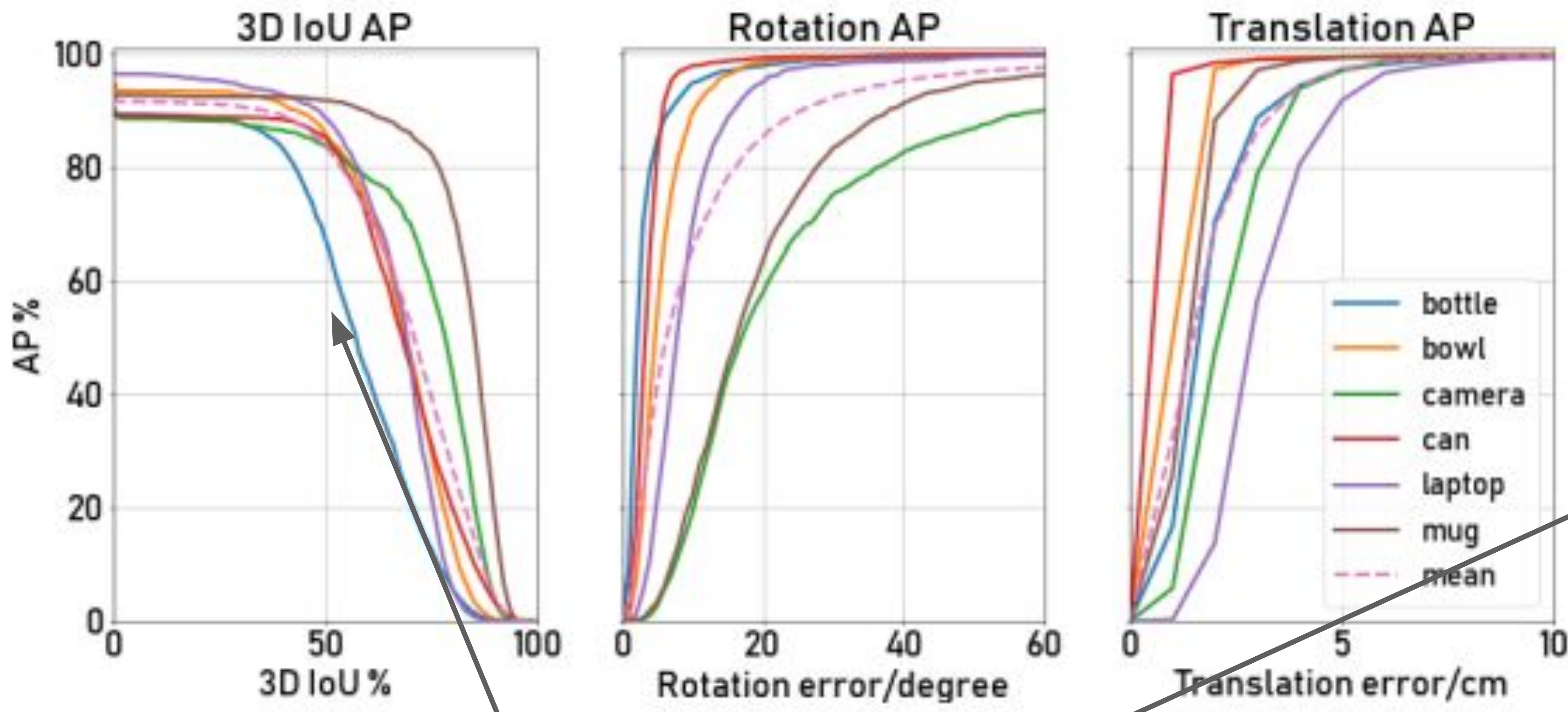


Translation AP

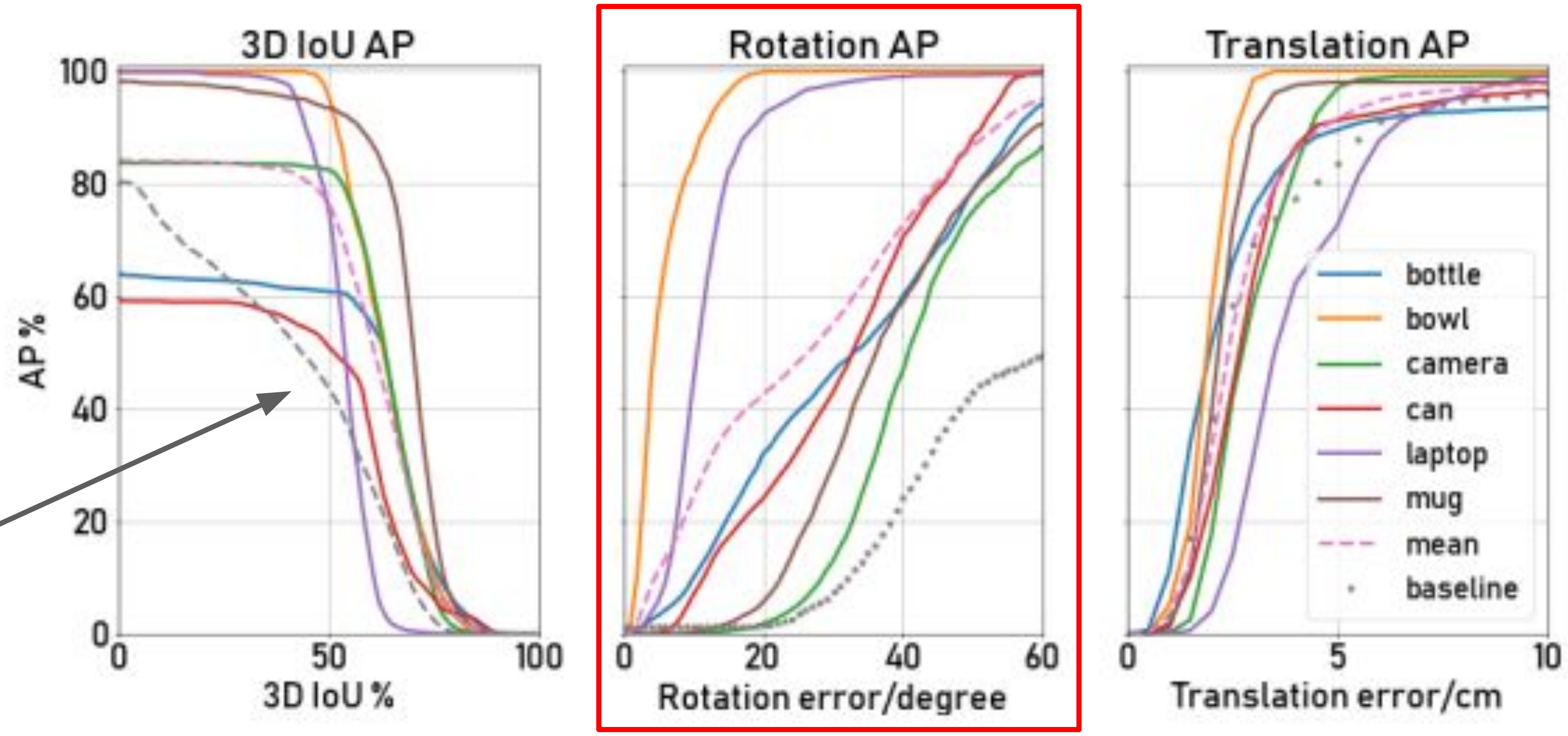


DR

Results: Actual



CAMERA25 Test Results



REAL275 Test Results

Sharp drop-off
after 50% IoU!

Ablation Studies

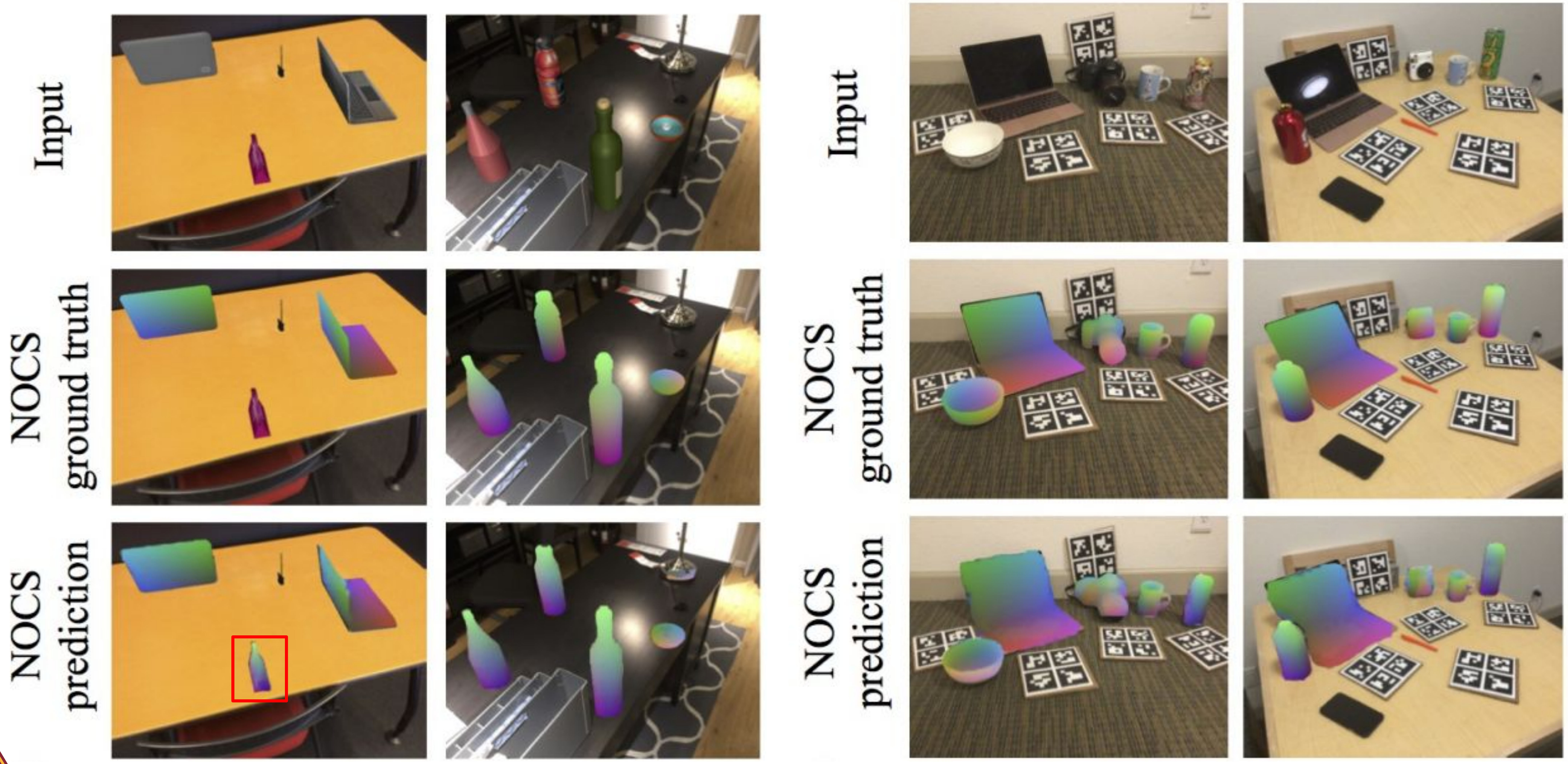
Data			mAP				
CAMERA*	COCO	REAL*	$3D_{25}$	$3D_{50}$	5° 5 cm	10° 5 cm	10° 10cm
C			51.7	36.7	3.4	20.4	21.7
C	✓		57.6	41.0	3.3	17.0	17.1
		✓	61.9	47.5	6.5	18.5	18.6
	✓	✓	71.0	53.0	7.6	16.3	16.6
C		✓	79.2	69.7	6.9	20.0	21.2
C	✓	✓	**79.6**	**72.4**	**8.1**	**23.4**	**23.7**
B			42.6	36.5	0.7	14.1	14.2
B	✓	✓	79.1	71.7	7.9	19.3	19.4

Testing on Real275

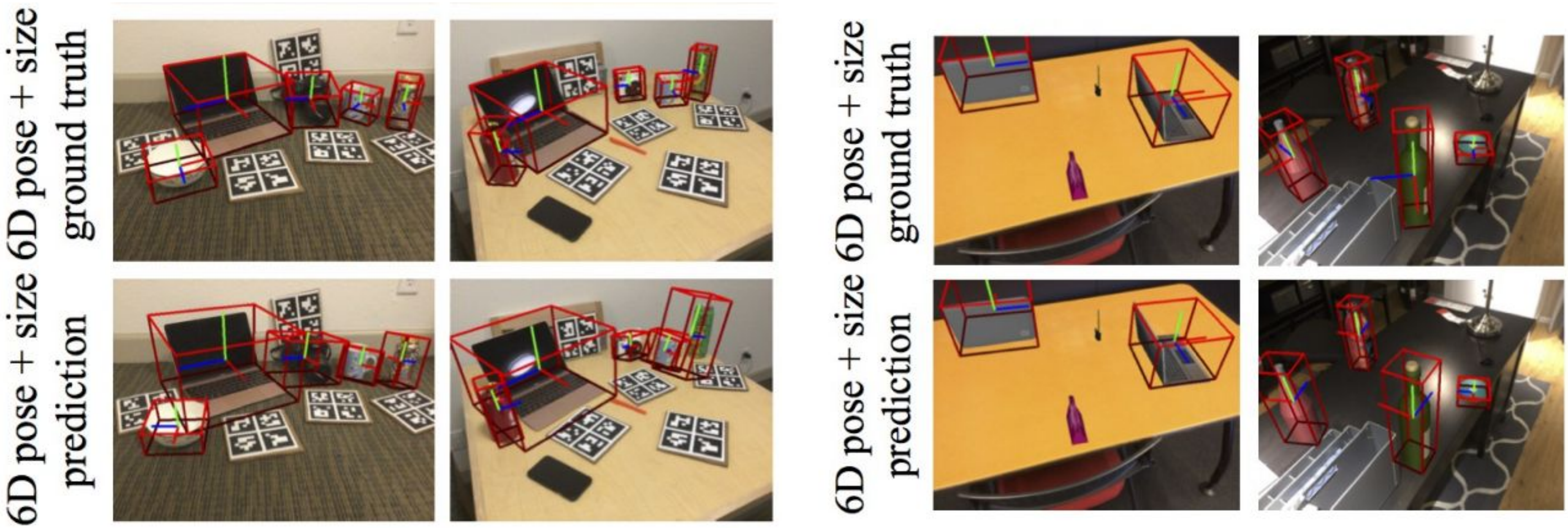
Data	Network	mAP				
		$3D_{25}$	$3D_{50}$	5° 5 cm	10° 5 cm	10° 10cm
CAMERA25	Reg.	89.3	80.9	29.2	53.7	54.5
	Reg. w/o Sym.	86.6	79.9	14.7	38.5	40.0
	32 bins	91.1	83.9	**40.9**	**64.6**	**65.1**
	128 bins	**91.4**	**85.3**	38.8	61.7	62.2
REAL275	Reg.	79.6	72.4	8.1	23.4	23.1
	Reg. w/o Sym.	82.7	73.8	1.3	9.1	9.3
	32 bins	84.8	78.0	**10.0**	25.2	25.8
	128 bins	**84.9**	**80.5**	9.5	**26.7**	**26.7**

Different losses

Qualitative Results



Qualitative Results



Conclusions

Primary Contributions:

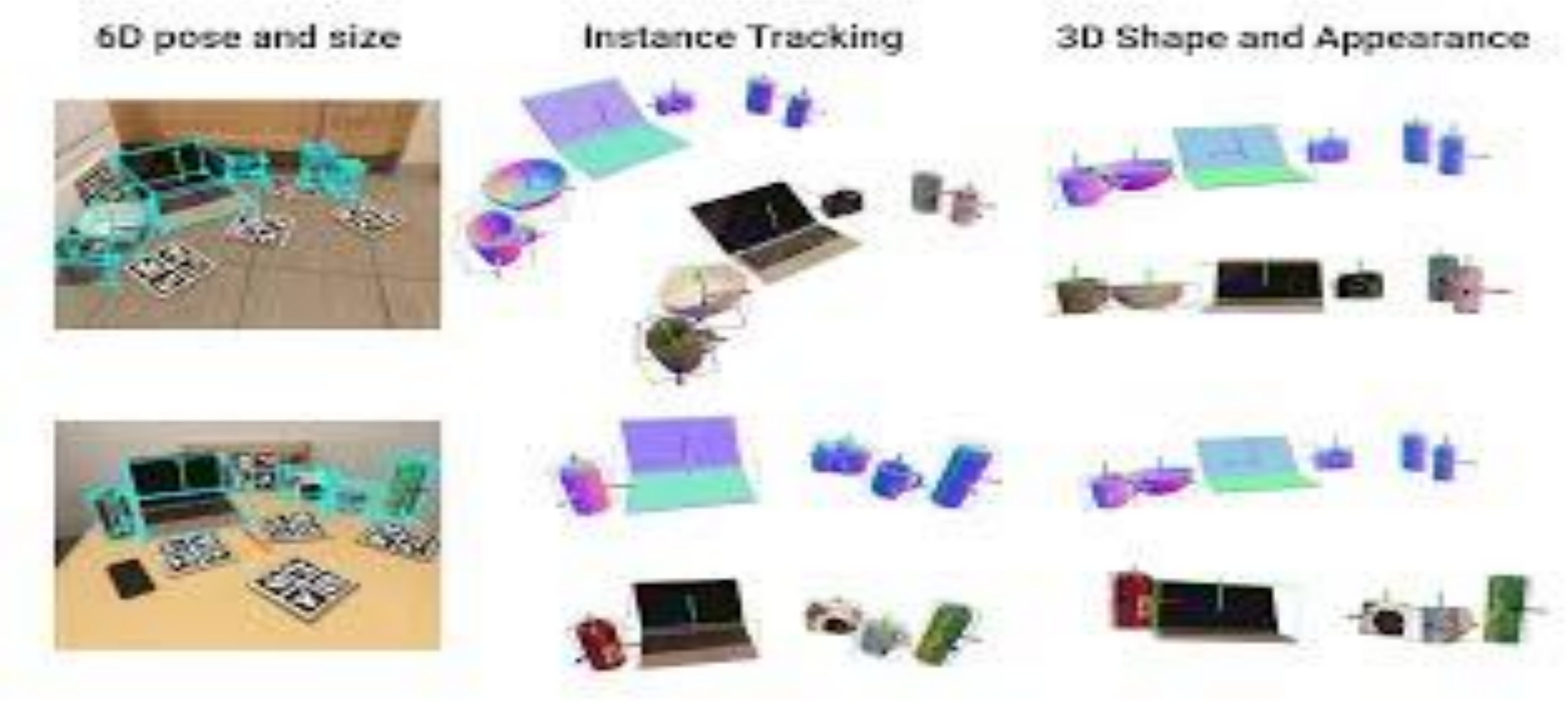
1. NOCS, a method which allows for different but related (same category) objects to have the same representation, allowing for 6D pose and size estimate
2. CNN which allows joint prediction of class label, instance mask, and NOCS map of multiple unseen objects in an image
3. Synthetic data generation technique in addition to the resultant CAMERA and Real data

Further work:

1. Incorrect region proposal or category prediction could result in failures
2. Relies on depth image to fully utilize the NOCS map
3. Does not talk about articulate objects

Extensions of NOCS

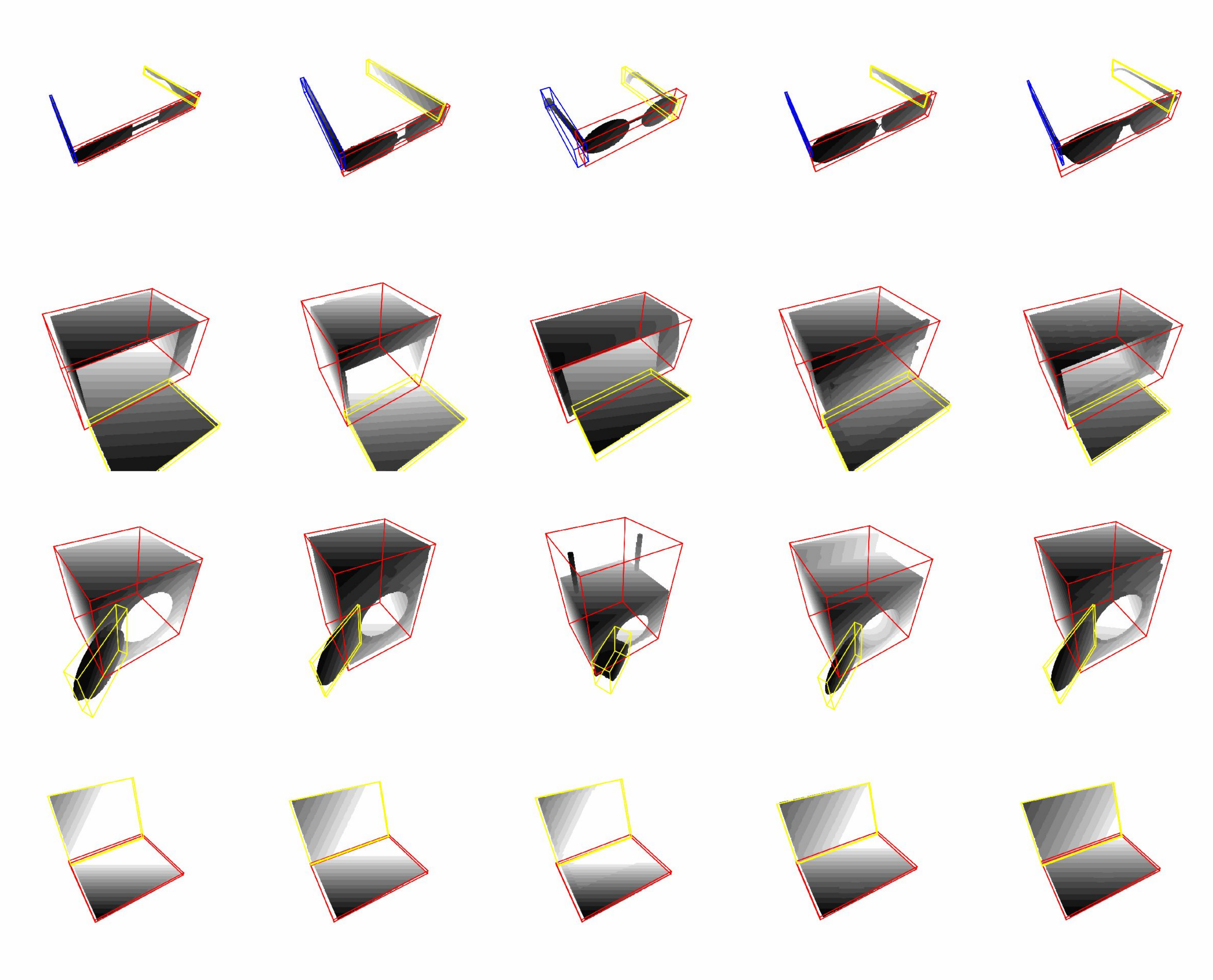
ShAPO: Implicit Representations for Multi-Object Shape, Appearance, and Pose Optimization

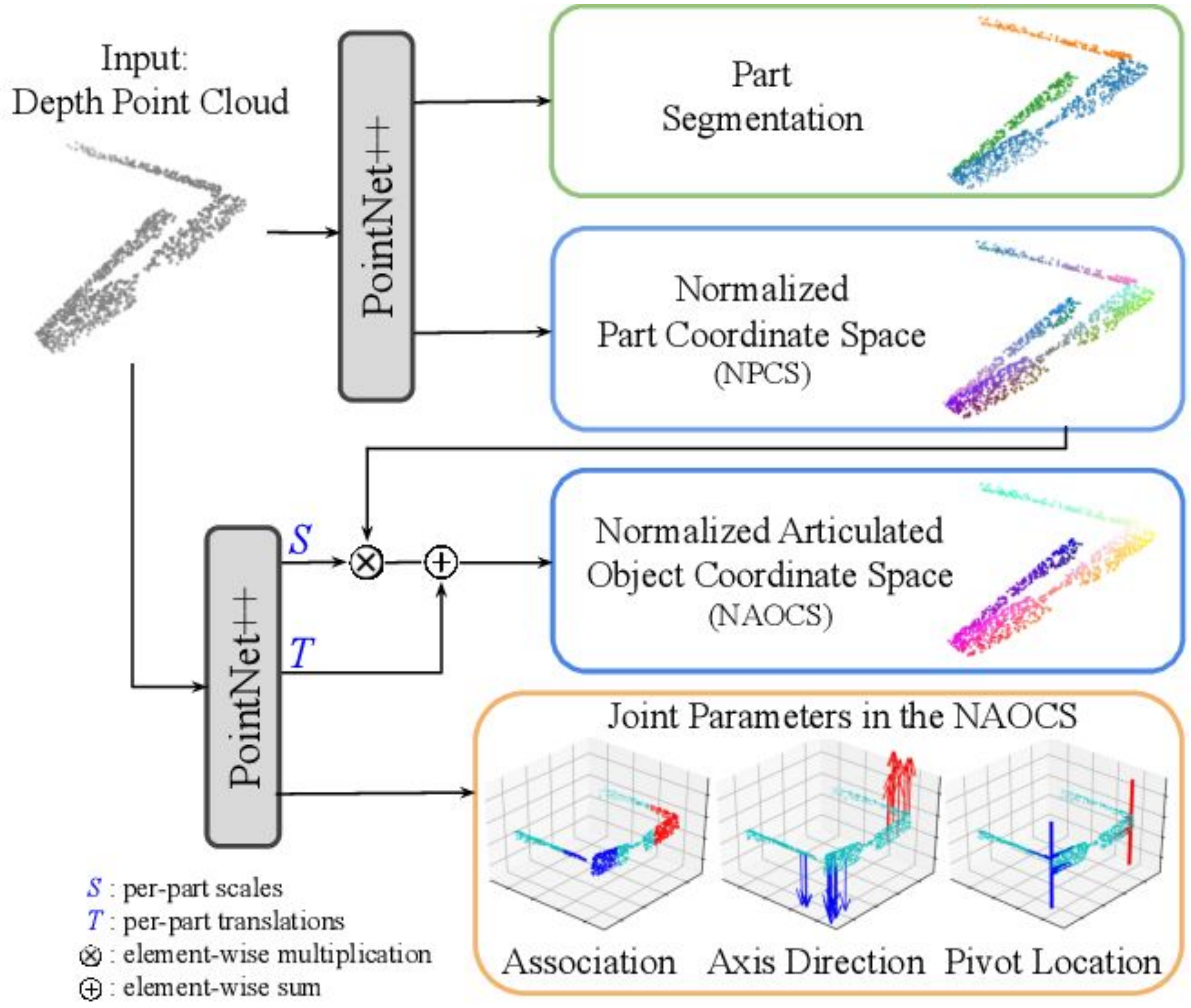


Irshad, M. Z., Zakharov, S., Ambrus, R., Kollar, T., Kira, Z., & Gaidon, A. (2021). ShAPO: Implicit Representations for Multi-Object Shape, Appearance, and Pose Optimization. arXiv preprint arXiv:2104.08901.

Other Relevant Works

Category-Level Articulated Object Pose Estimation





Li, X., Wang, H., Yi, L., Guibas, L., Abbott, A.L., & Song, S. (2018). Category-Level Articulated Object Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1572-1580).

Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation

What is the best object representation for robot manipulation?

Common Representation Across a category

Picking up an object at same point in different orientations

Florence, Peter R., Lucas Manuelli, and Russ Tedrake. "Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation." *Conference on Robot Learning*. PMLR, 2018.

Questions?

DeepRob

[Student] Lecture 15

by Bharath Sivaram, Sahith Reddy, Prakadeeswaran Manivanan

Rigid Object Perception, Dense Descriptors, Category-level Object Pose Estimation

University of Michigan and University of Minnesota