DISP6D: Disentangled Implicit Shape and Pose Learning for Scalable 6D Pose Estimation

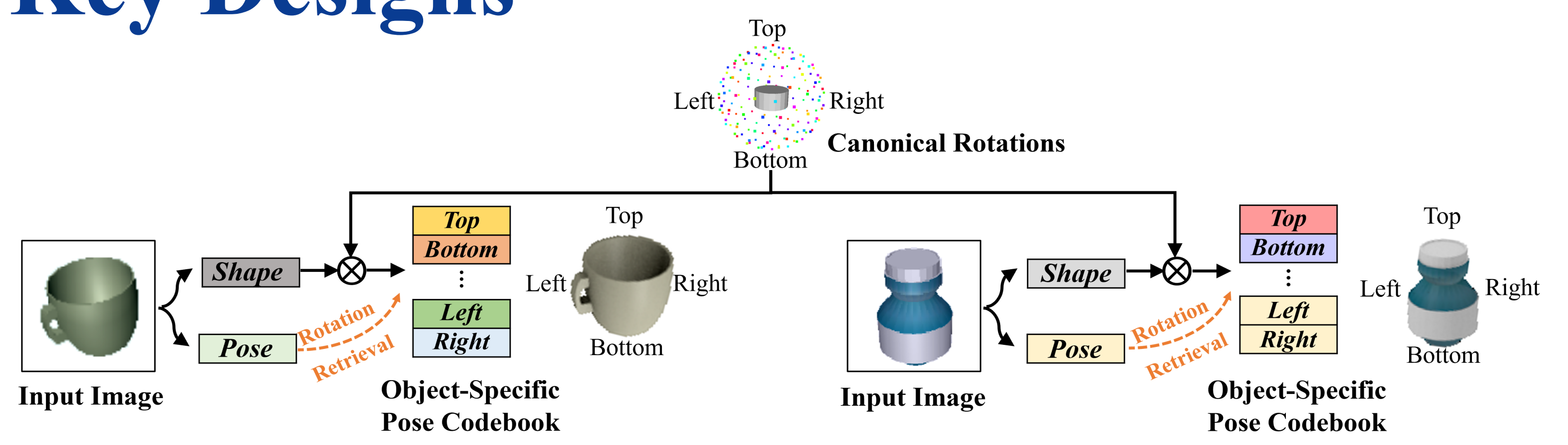
Yilin Wen[1,†], Xiangyu Li[2,†], Hao Pan[3], Lei Yang[1,4], Zheng Wang[5], Taku Komura[1], Wenping Wang[6]

[1]The University of Hong Kong [2]Brown University [3]Microsoft Research Asia [4]TransGP [5]SUSTech [6]Texas A&M University

Task

Scalable 6D pose estimation for rigid objects from RGB images: Aiming at ***handling multiple objects*** and ***generalizing to novel objects*** with a single framework.

Key Designs

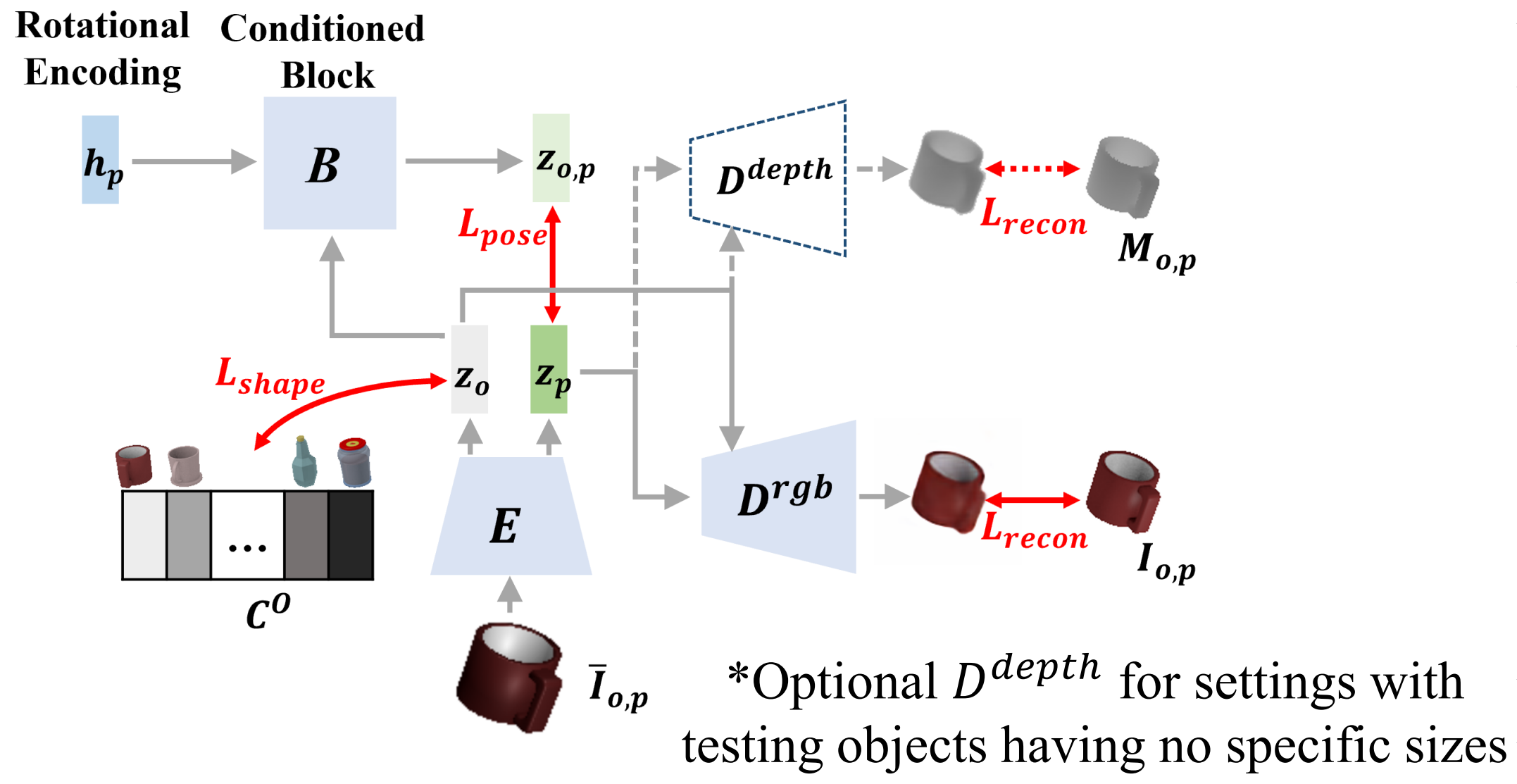


*Symmetries marked by code color

We extend the auto-encoding framework for RGB-based rotation estimation, by:

- ***Disentangling* the object shape and pose code to improve scalability.** A regular shape space is learned with contrastive learning, and the pose code is compared with canonical rotations for pose estimation.
- ***Re-entangling* the shape and canonical rotation to model the different pose spaces due to different object symmetries.** Object-conditioned pose codebooks are generated for rotation retrieval.

Framework



*Optional D^{depth} for settings with testing objects having no specific sizes

Disentangled shape and pose learning with the auto-encoding framework

- The encoder E maps the input image to its implicit shape and pose code z_o, z_p. Image $I_{o,p}$ is augmented is augmented into $\bar{I}_{o,p}$ for the input in training.
- The decoder D^{rgb}(or plus D^{depth}) tries to recover the canonical image $I_{o,p}$ (or plus the canonical depth map $M_{o,p}$) from z_o, z_p, by conditioning the per-view reconstruction on the shape code z_o with the AdaIN modulation.
- Training Objective: $L_{recon} = \sum_{o,p}\left(\left\|I_{o,p} - D^{rgb}\left(E(\bar{I}_{o,p})\right)\right\|^2 + \left\|M_{o,p} - D^{depth}\left(E(\bar{I}_{o,p})\right)\right\|^2\right)$

Contrastive Metric Learning for Object Shapes

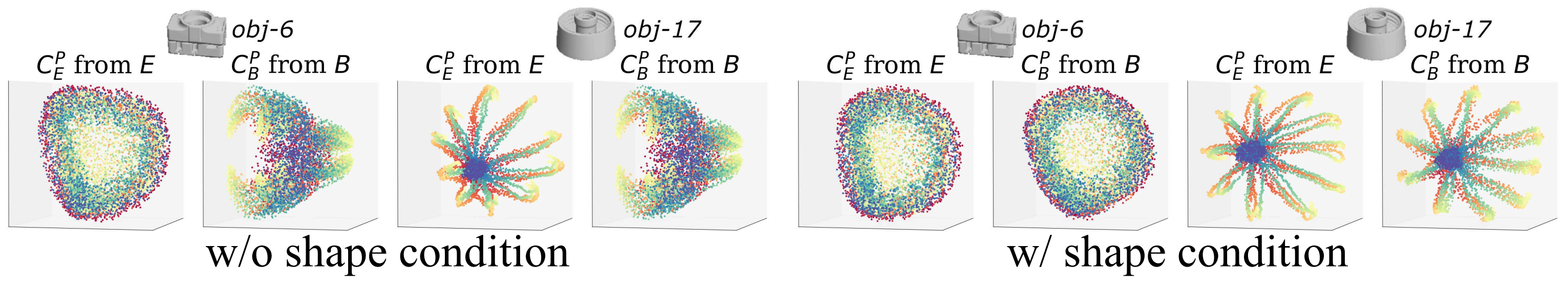
- A metric space for the shape codes is built with contrastive metric learning, where we establish a shape embedding C^o with each $c_i \in C^o$ representing a training object, and model the proximity between z_o and C^o.
- Training Objective: $L_{shape} = -\sum_{o,p}\sum_{i=1}^{N_o} w_i^o \log \Pr(c_i|z_o)$, with w^o as a one-hot vector for the target distribution.

Re-entanglement of Shape and Pose

- The conditioned block B entangles the rotational position encoding h_p and the shape code z_o with a tensor product structure, and outputs a pose code $z_{o,p}$ that is comparable with the z_p generated by E.
- Training Objective: $L_{pose} = -\sum_{o,p} \hat{z}_{o,p} \cdot \hat{z}_p$, with $\hat{z}$ denoting the normalized unit-length vector for z.

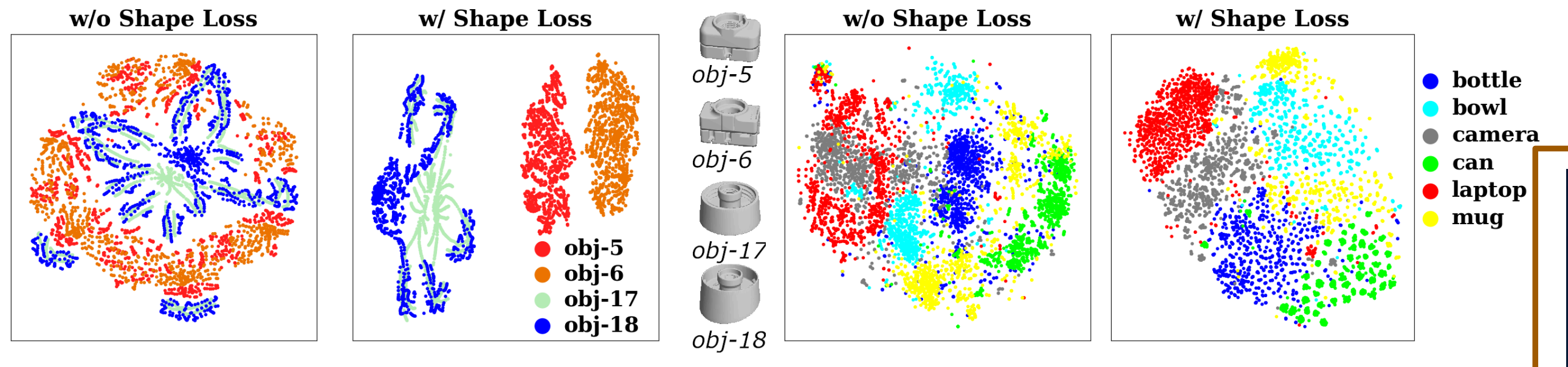
Ablation and Visualization

- ***Shape Conditioned Pose Code Generation***



Top three PCA projections of pose codes. Point colors (blue→green→red) encode rotations as viewpoints change from north pole to south pole.

- ***Contrastive Metric Learning of Latent Shape Space***

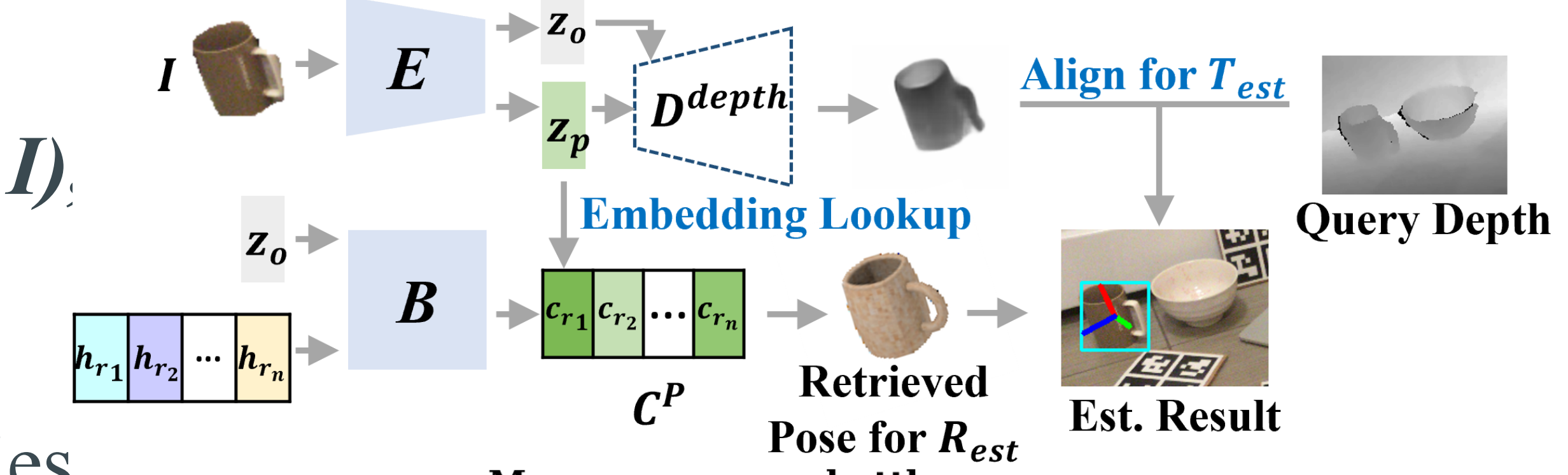


t-SNE embedding of z_o. Our network is unaware of category labels.

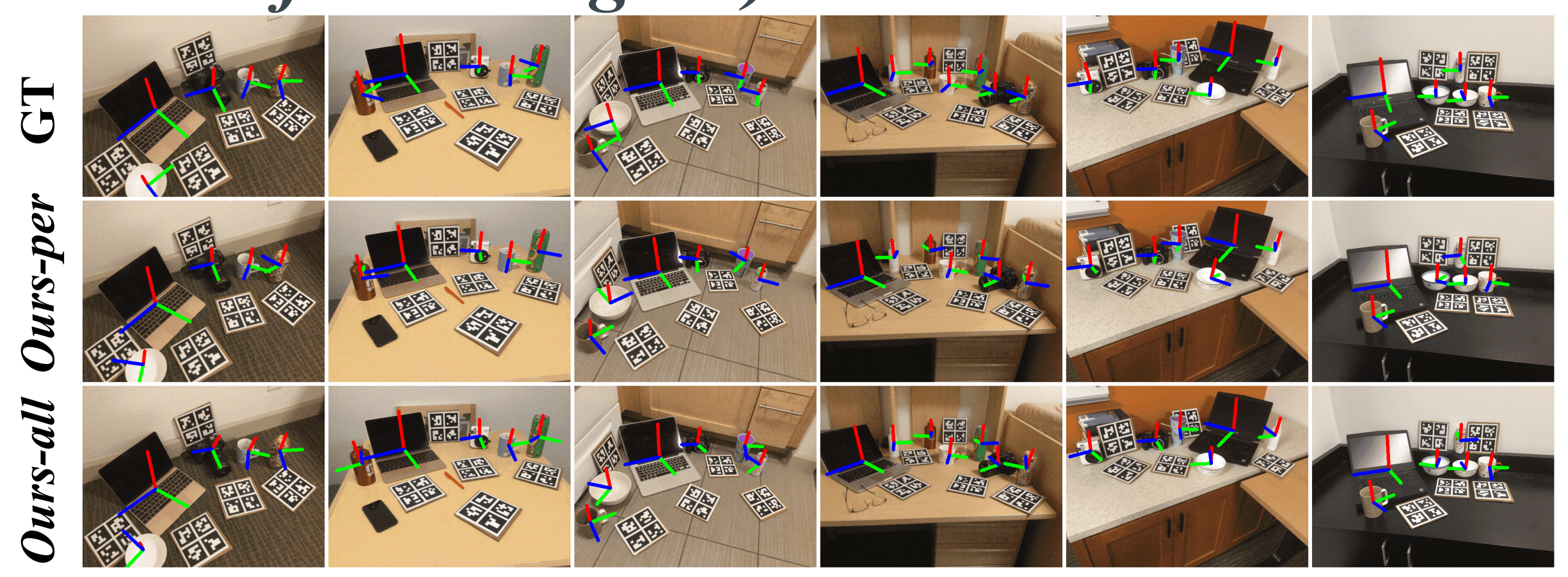
[†]Work partially done during internships with Microsoft Research Asia.

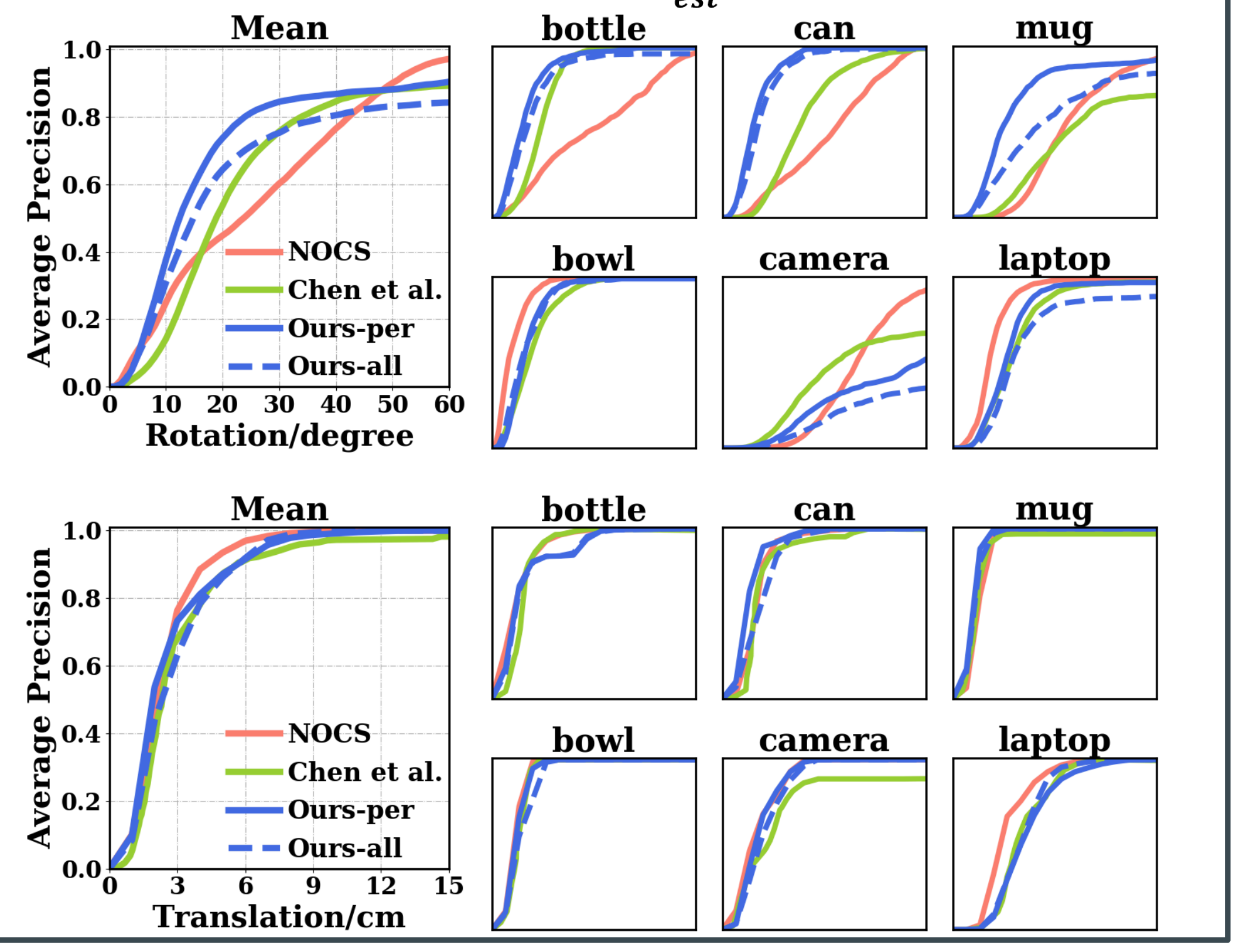
Inference Settings I&III

Novel objects in a given category (Setting I), or across categories (Setting III), without knowing 3D models. Setting III extends Setting I by combining objects of all categories into one set, without referring to predefined category labels in both training and testing.



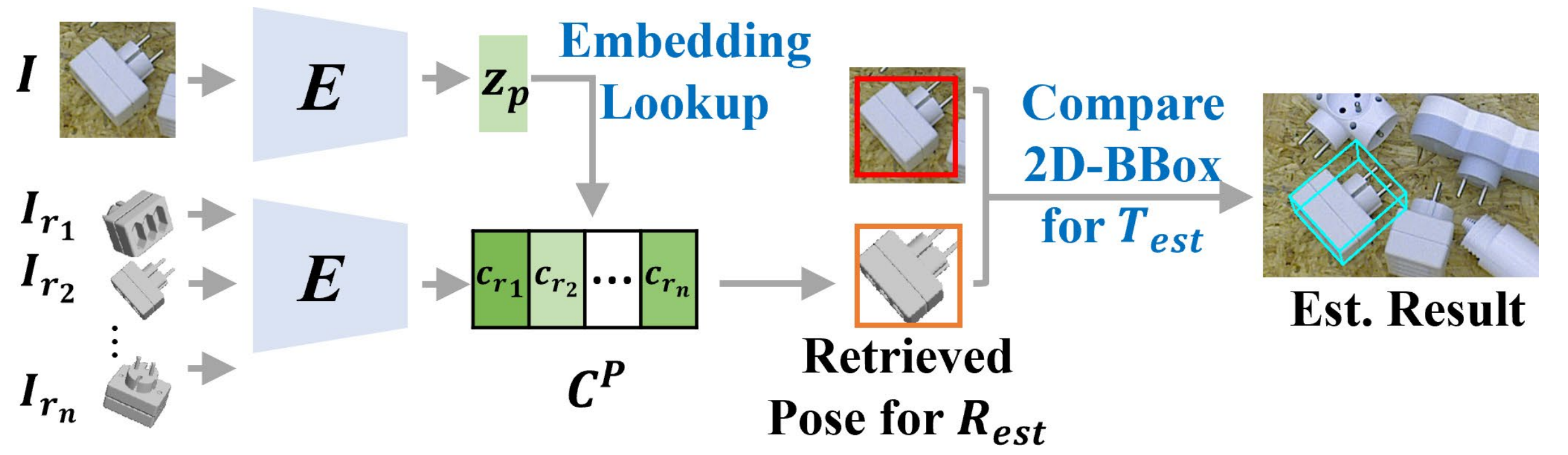
Results on REAL275 (Ours-per for Setting I; Ours-all for Setting III)





Inference Setting II

Novel objects with 3D models. Objects have drastic geometric differences and no specific category consistency.

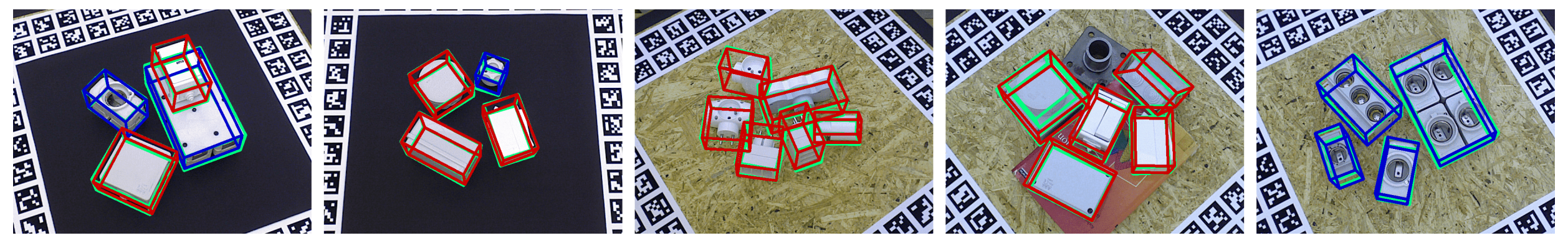

I
E
z_p
Embedding Lookup
Compare 2D-BBox for T_{est}
Est. Result
I_{r_1}
I_{r_2}
I_{r_n}
E
c_{r_1} c_{r_2} … c_{r_n}
C^P
Retrieved Pose for R_{est}

Results on T-LESS (Train on Obj. 1-18 only)

w/ 2D GT	*Obj. 1-18*	*Obj. 19-30*	*Obj. 1-30*
MP-AAE	60.75	59.89	60.41
Nguyen *et al.*	59.62	57.75	58.87
Ours	**66.14**	**64.42**	**65.45**

w/ MaskRCNN	*Obj. 1-30*
MP-AAE	23.51
Pitteri *et al.*	23.27
Ours	**35.36**

Average recall rates with $e_{VSD} < 0.3$



Ours on trained objects/unseen objects; GT

October 23-27, 2022, Tel Aviv