Pose Estimation of Novel Rigid Objects

Van Nguyen Nguyen

Jury:

Markus Vincze (Prof, TU Wien) - reviewer

Benjamin Busam (Researcher, TU Munich) - reviewer

Josef Sivic (Prof, CTU Prague) - examiner

Dima Damen (Prof, University of Bristol & Google Deepmind) - examiner

Slobodan Ilic (Researcher, TU Munich & Siemens) - examiner

Vincent Lepetit (Prof, ENPC) - supervisor

PhD defense

19th December 2024

ÉCOLE NATIONALE DES PONTS ET CHAUSSÉES

IP PARIS

Object pose estimation

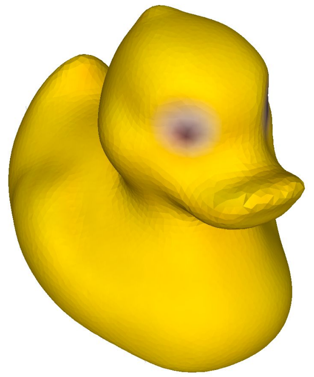


Image courtesy of BOP challenges

Object pose estimation

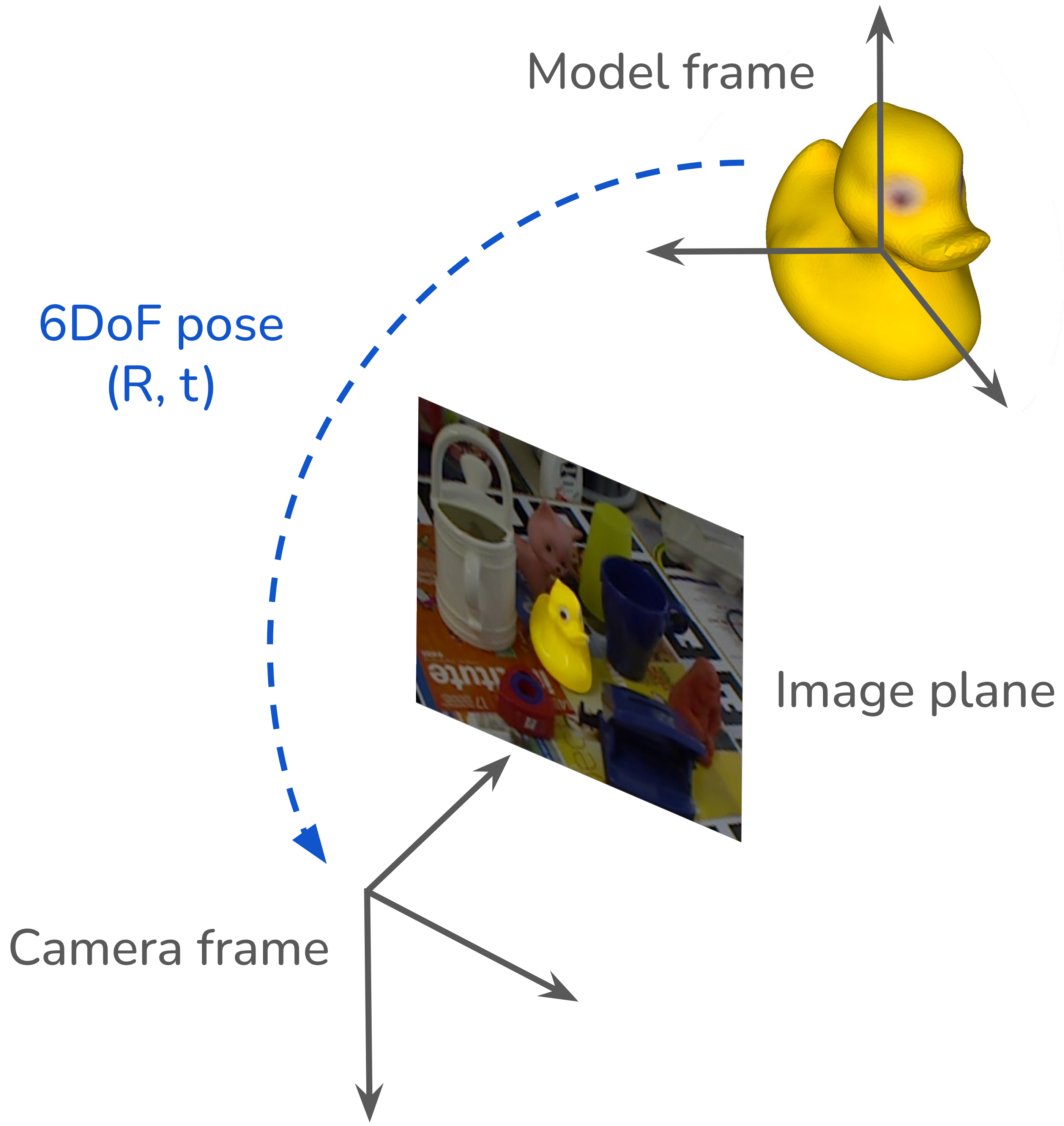
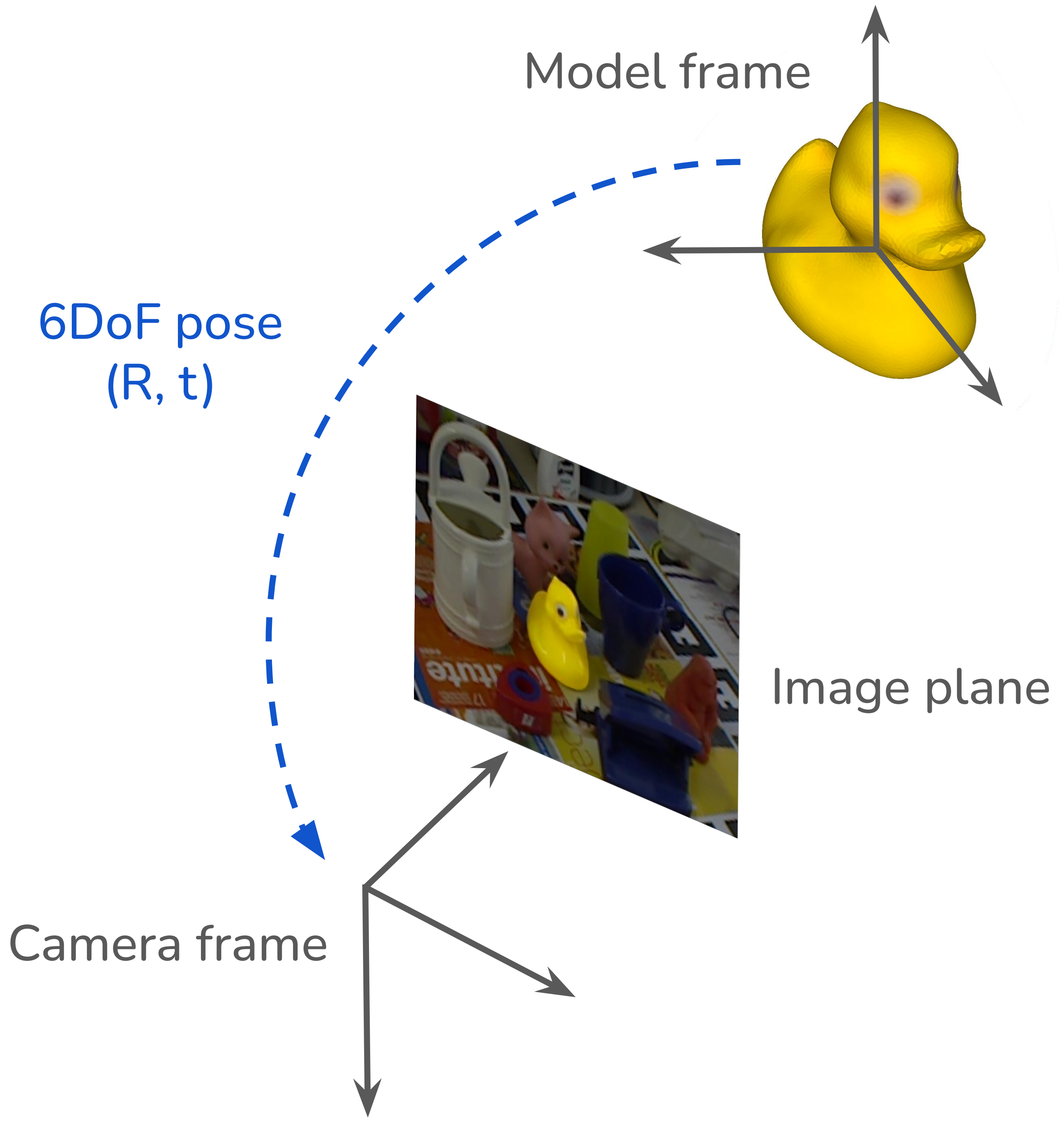


Image courtesy of BOP challenges

Object pose estimation



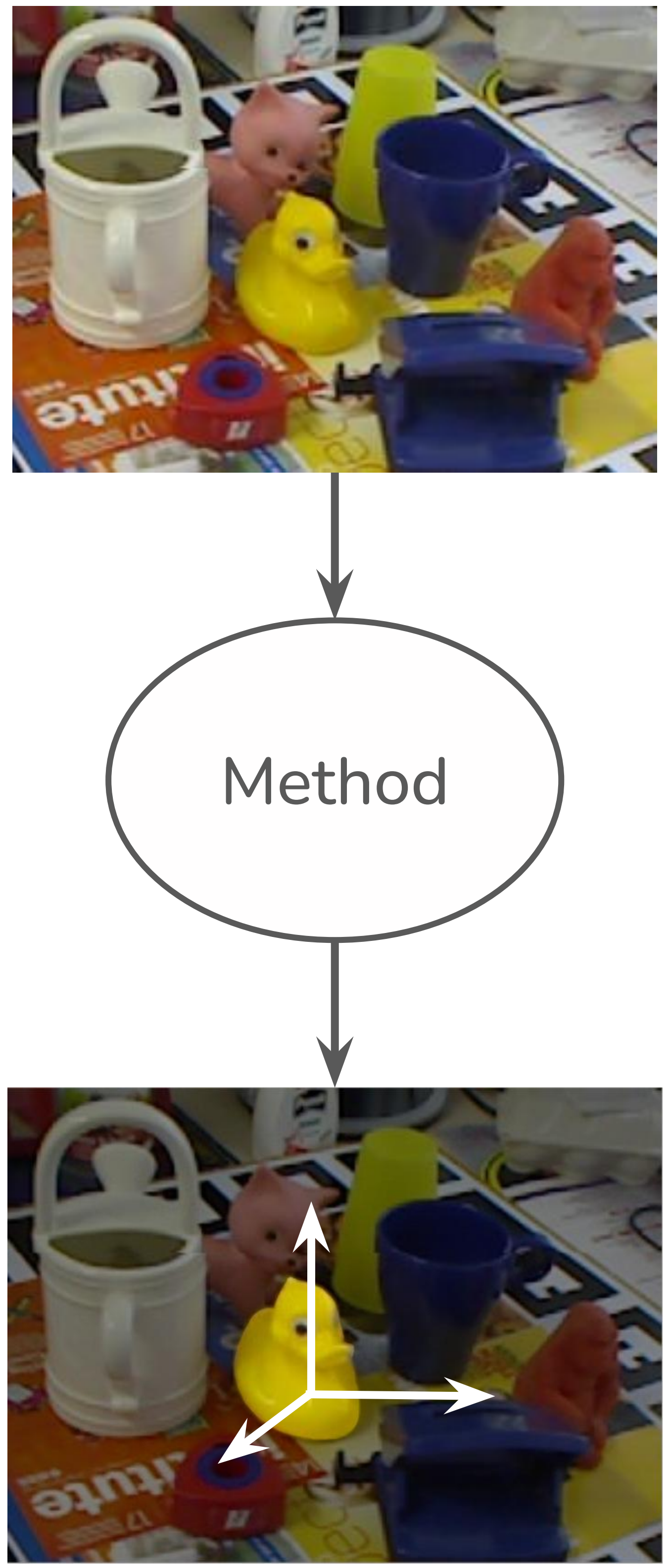


Image courtesy of BOP challenges

Applications of object pose estimation

Robotics: bin picking, grasping, robot navigation
(Demo from Pickit)

Augmented reality
(Vacchetti *et al.*, ISMAR 2004)

Combining Edge and Texture Information for Real-Time Accurate 3D Camera Tracking, Vacchetti *et al.*, ISMAR 2004

Seen object pose estimation

Image courtesy of BOP challenges

Seen object pose estimation

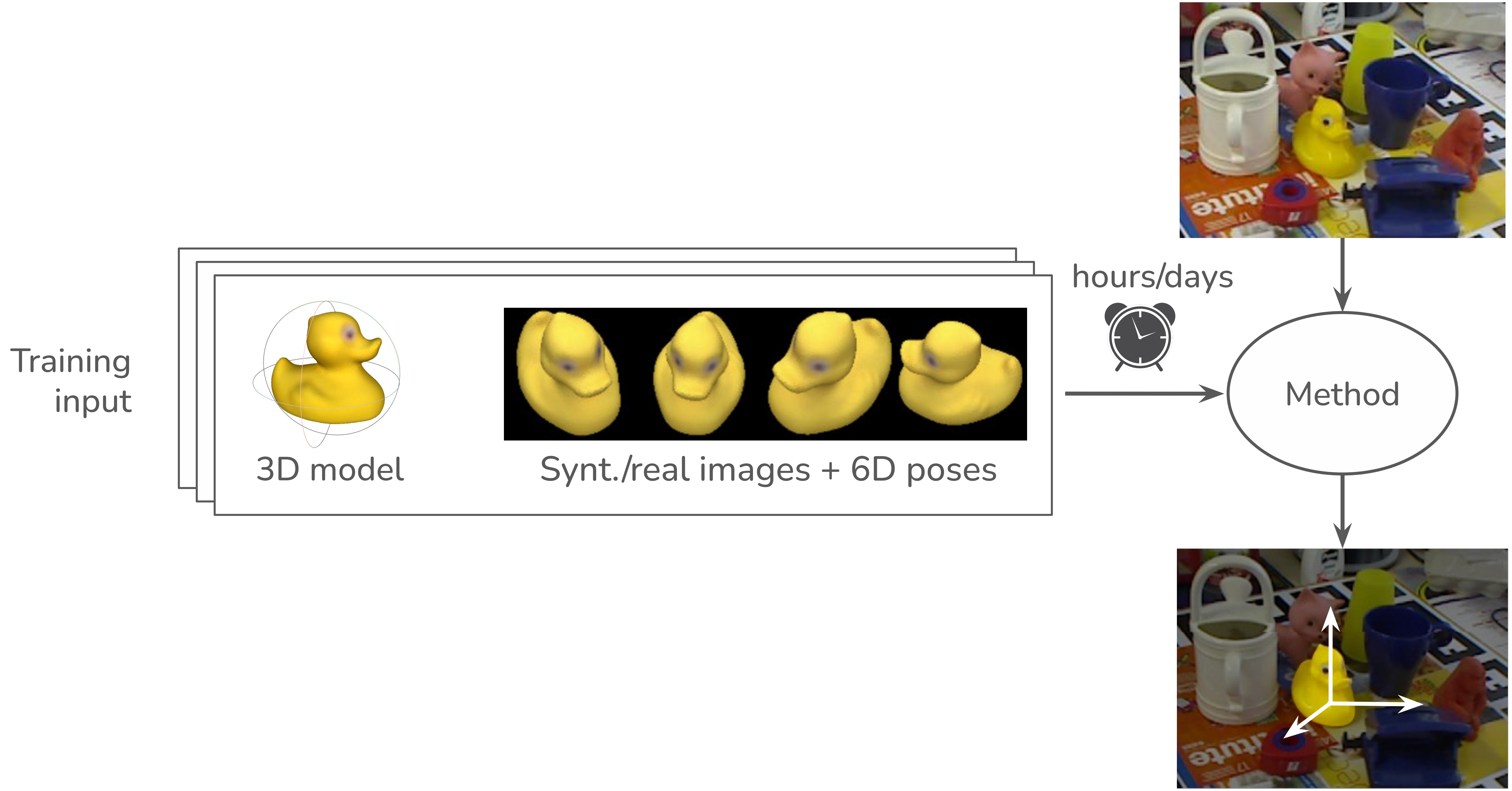


Image courtesy of BOP challenges

Novel object pose estimation

Novel = not seen during the training

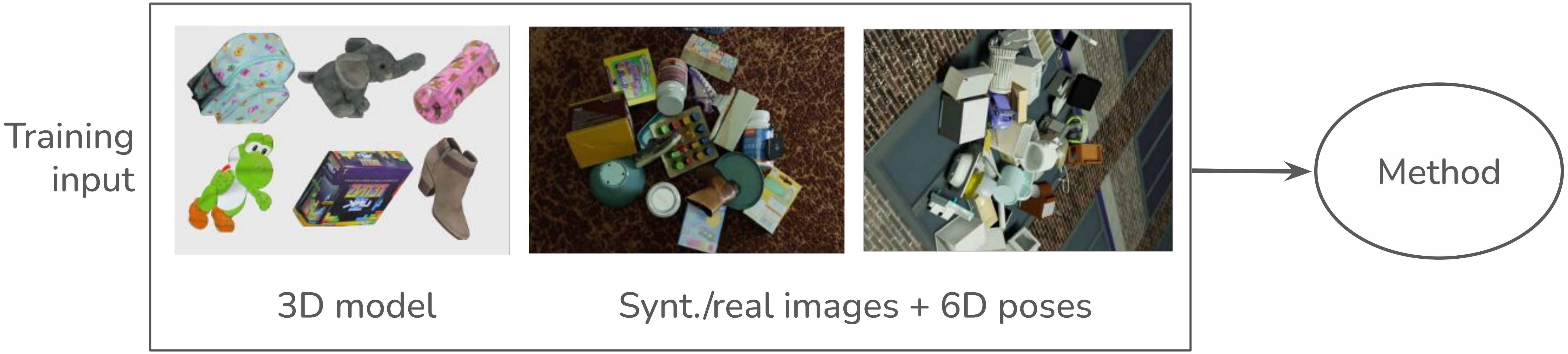


Image courtesy of BOP challenges

Novel object pose estimation

Novel = not seen during the training

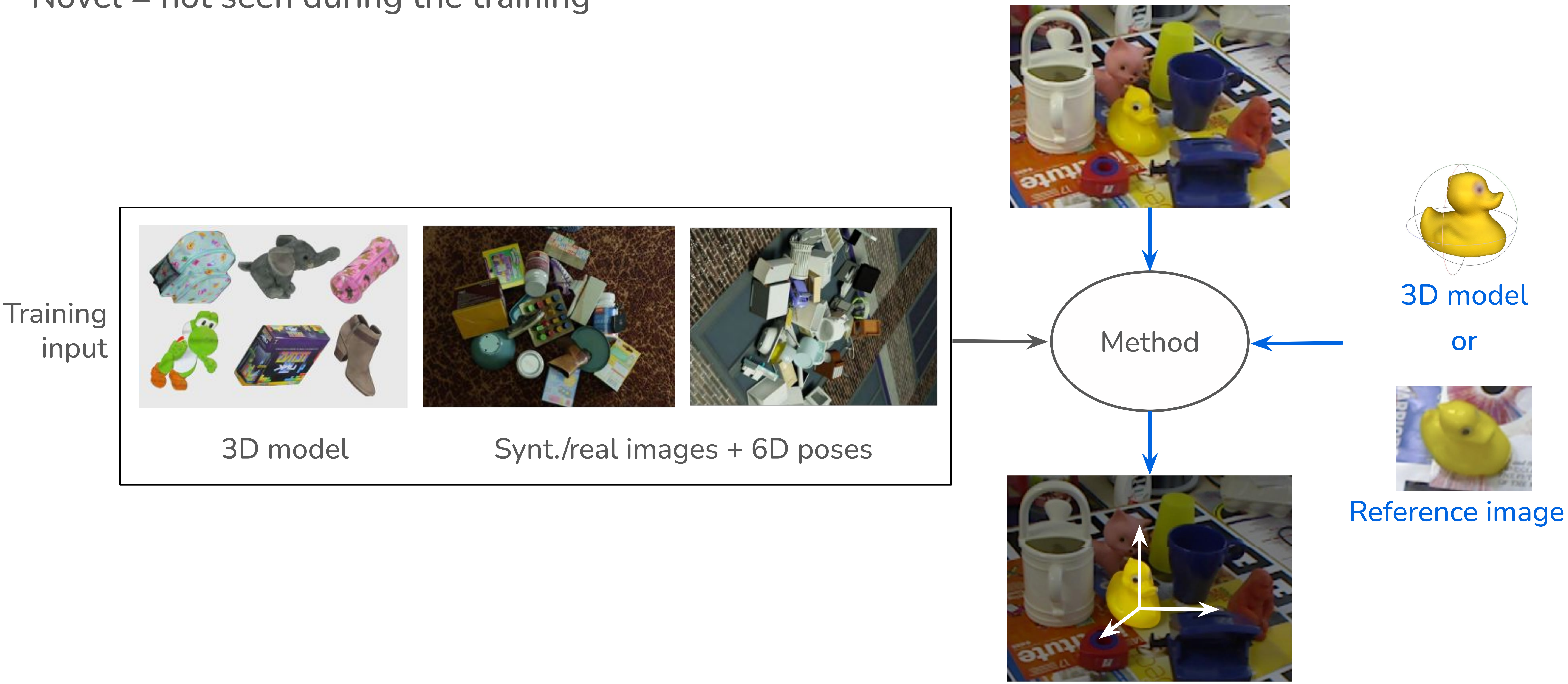
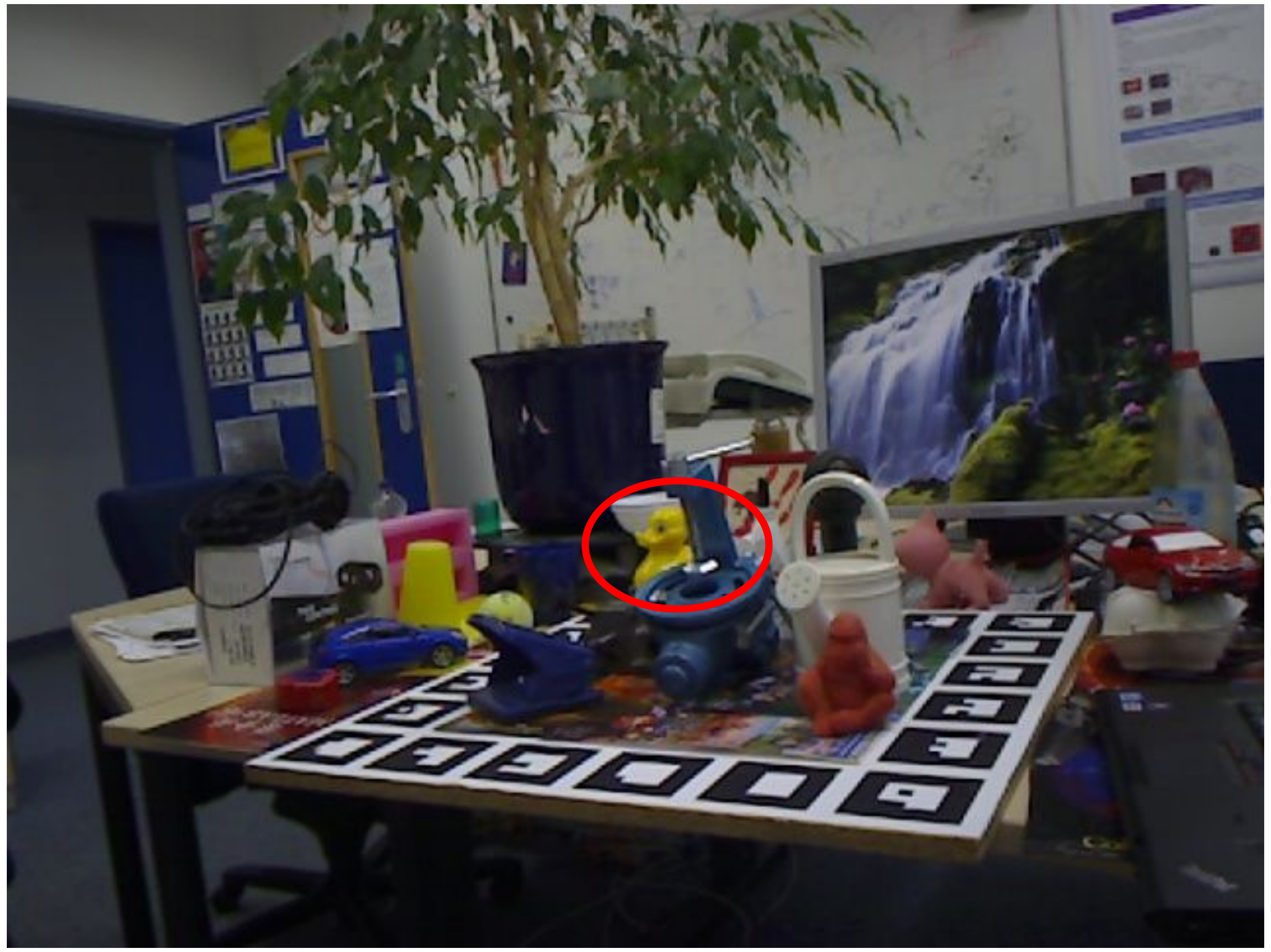


Image courtesy of BOP challenges

Challenges

- Cluttered background and occlusions



Credit: LM, T-LESS datasets

Challenges

- Viewpoints and illumination

Credit: TUD-L dataset

Challenges

- Textureless objects

Credit: YCB-V, T-LESS datasets

Challenges

- Generalization to novel objects, domain gap

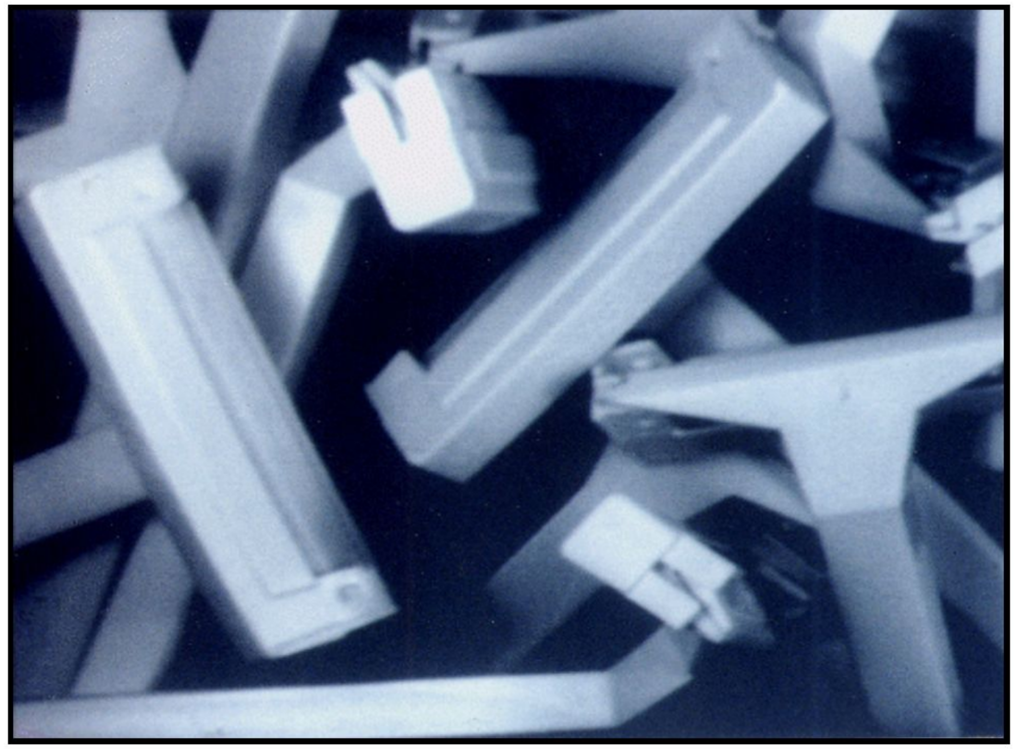
Synthetic training images generated by BlenderProc

Real-world images

Credit: MegaPose, BOP challenge datasets

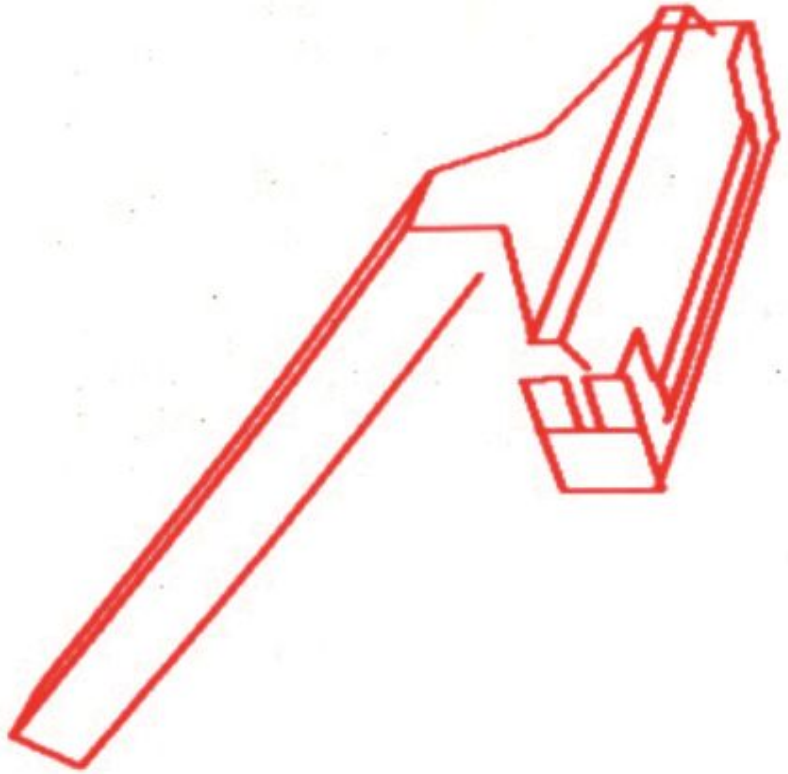
Early work

3D Object Recognition, David G. Lowe, 1987



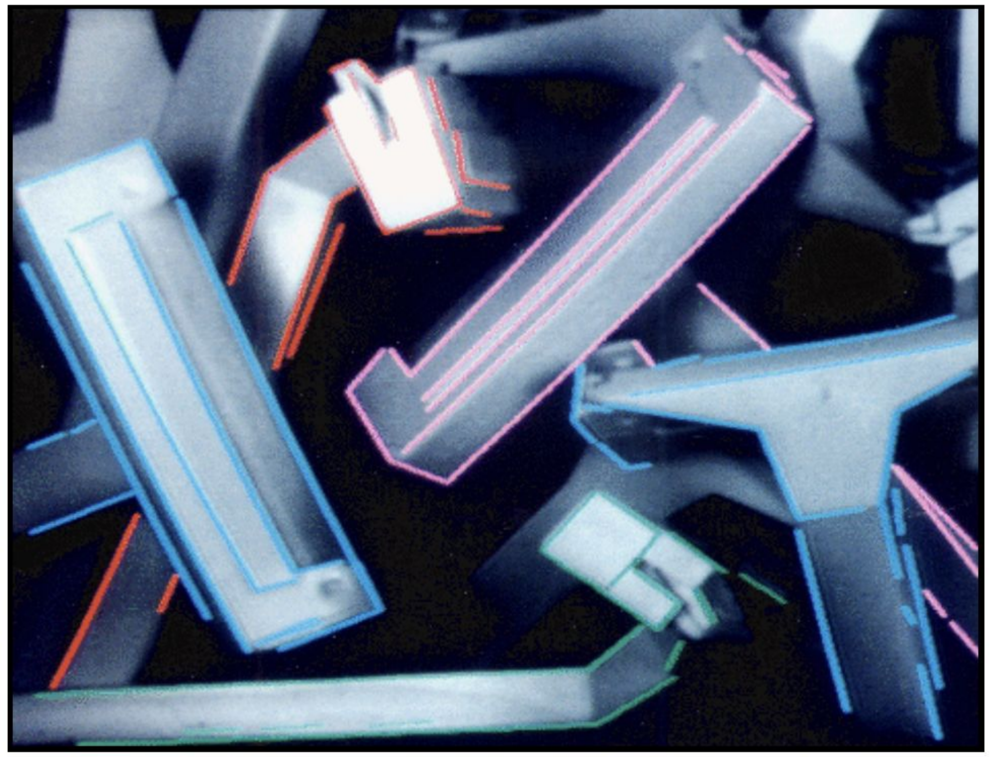
Input RGB

+



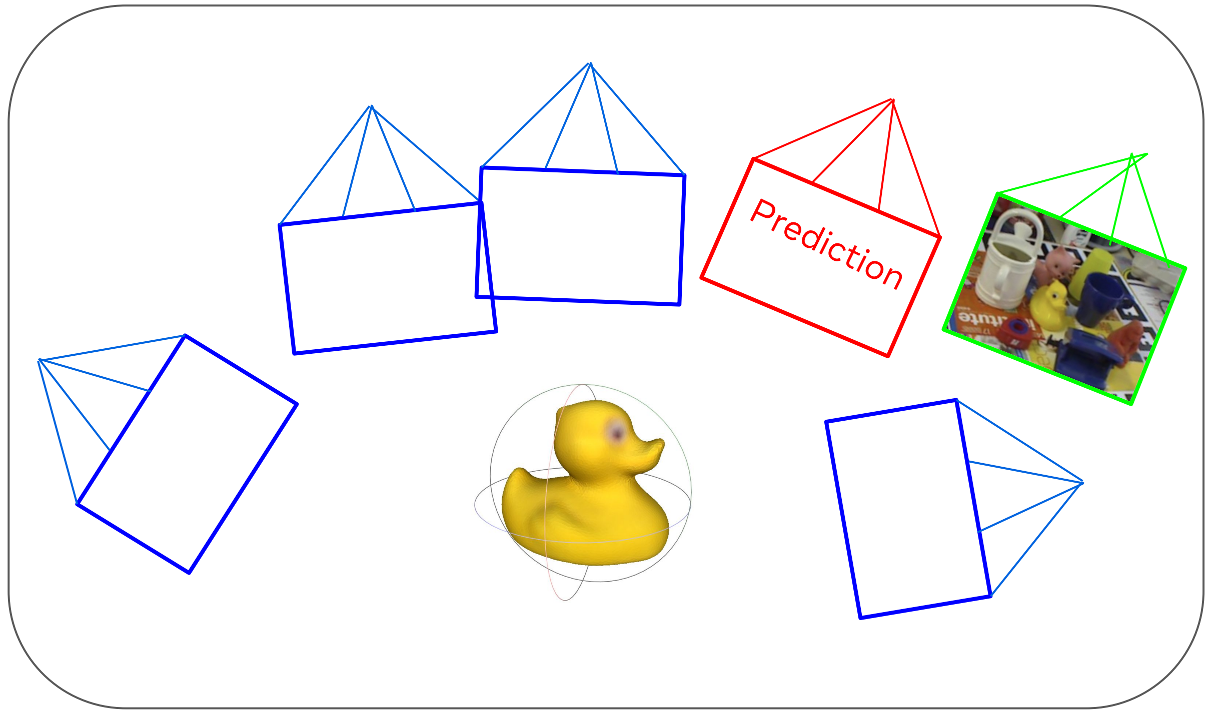
Wire-frame model

→



Prediction

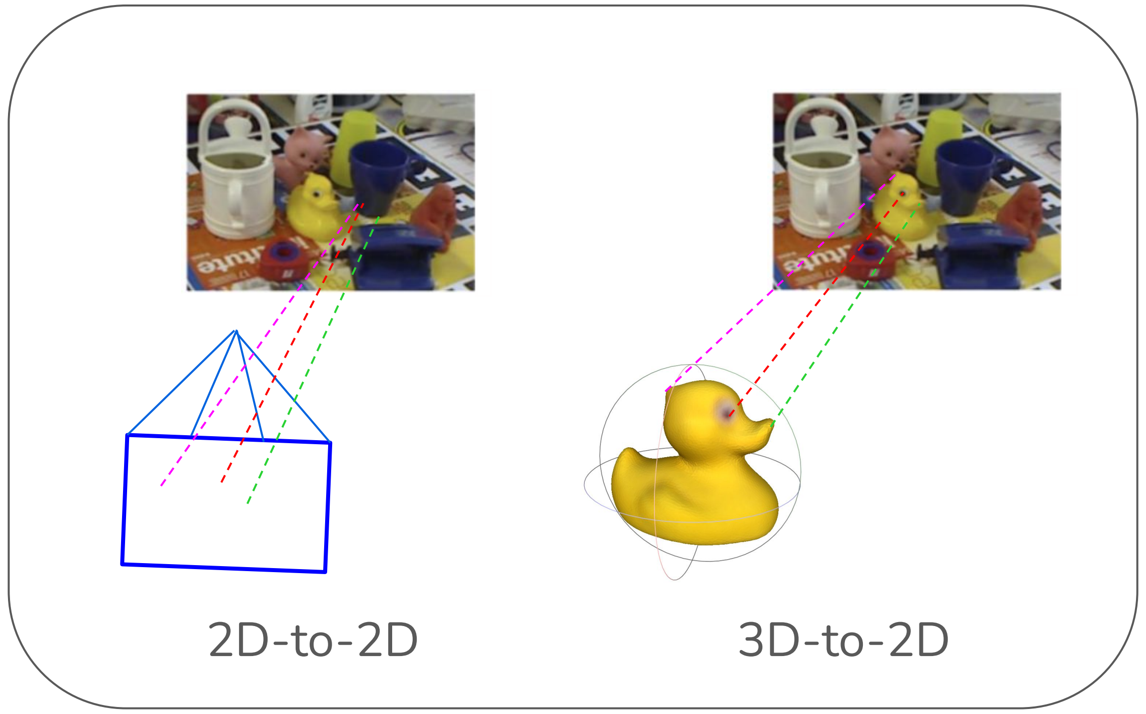
Related work: seen object pose estimation (-2021)



Template matching

- Learning descriptors [Wohlhart, CVPR 2015]
- Pose guided learning [Balntas, ICCV 2017]
- Implicit 3D learning [Sundermeyer, ECCV 2018]
- MultiPath learning [Sundermeyer, CVPR 2020]

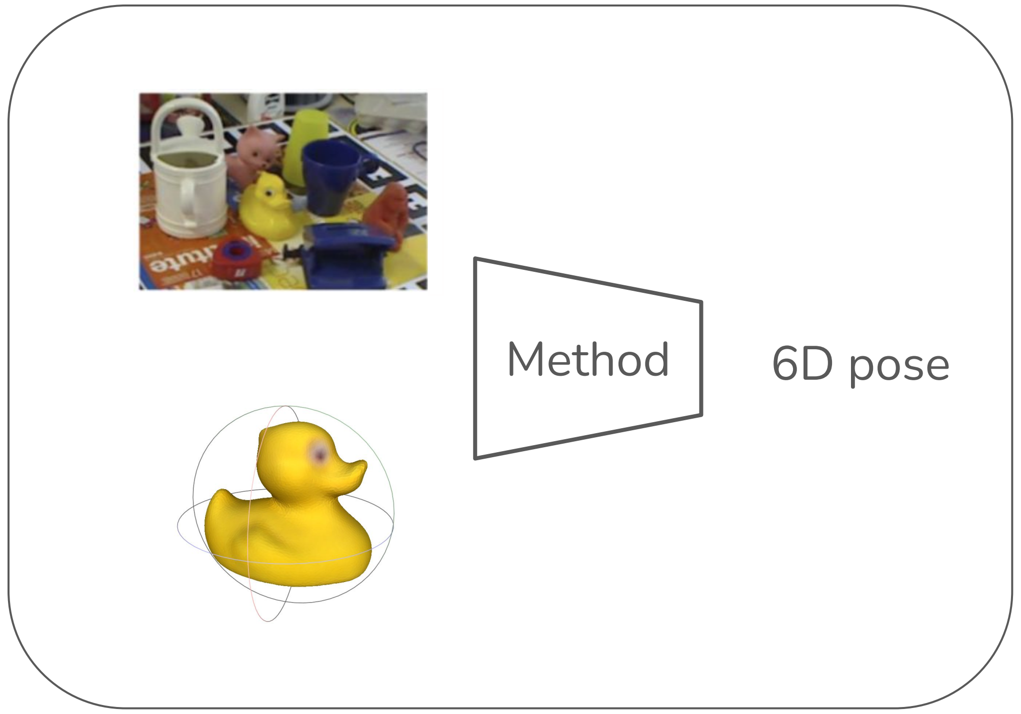
...



Correspondence-based

- BB8 [Rad, ICCV 2017]
- Heatmaps [Oberweger, ECCV 2018]
- PVNet [Peng, CVPR 2019]
- DPOD [Zakharov, ICCV 2019]
- Pix2Pose [Park, ICCV 2019]
- EPOS [Hodan, CVPR 2020]
- Single-stage [Hu, CVPR 2020]

...



Direct pose estimation

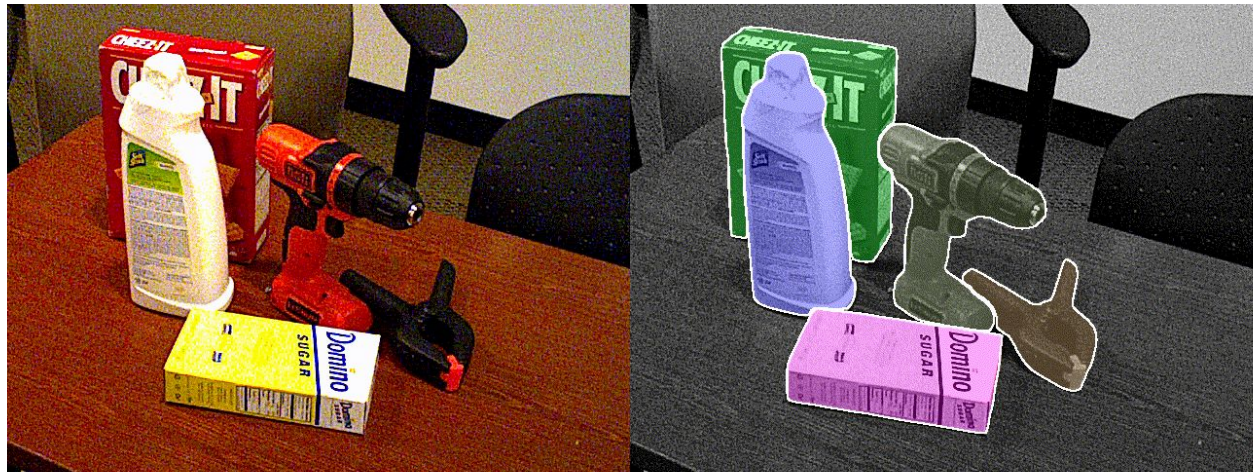
- SS6D [Kehl, ICCV 2017]
- Real-Time Seamless [Tekin, CVPR 2018]
- DeepIM [Li, ECCV 2018]
- PosefromShape [Xiao, BMVC 2019]
- I Like to Move It [Busam, arXiv, 2020]
- CosyPose [Labbé, ECCV 2020]
- GDR-Net [Wang, CVPR 2021]

...

Main contributions of this thesis

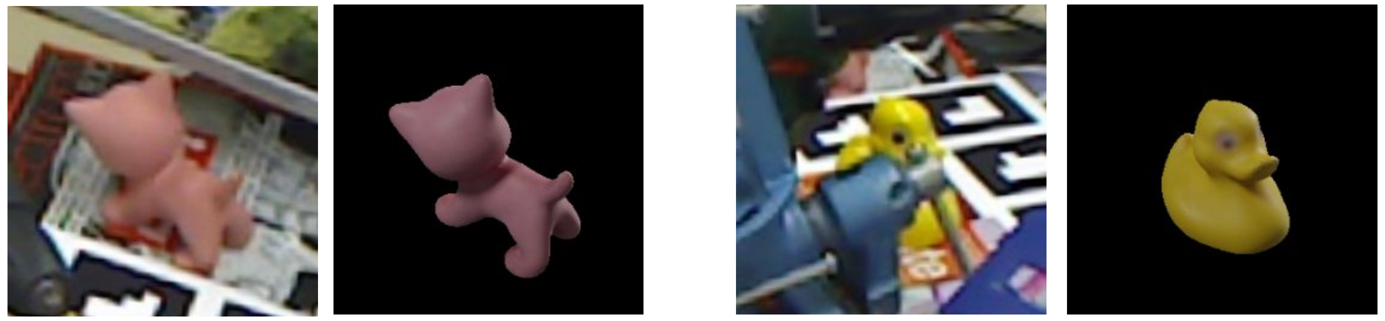
BOP 2023
4 awards

[ICCVW'23] CNOS: A Strong Baseline for CAD based Novel Object Segmentation



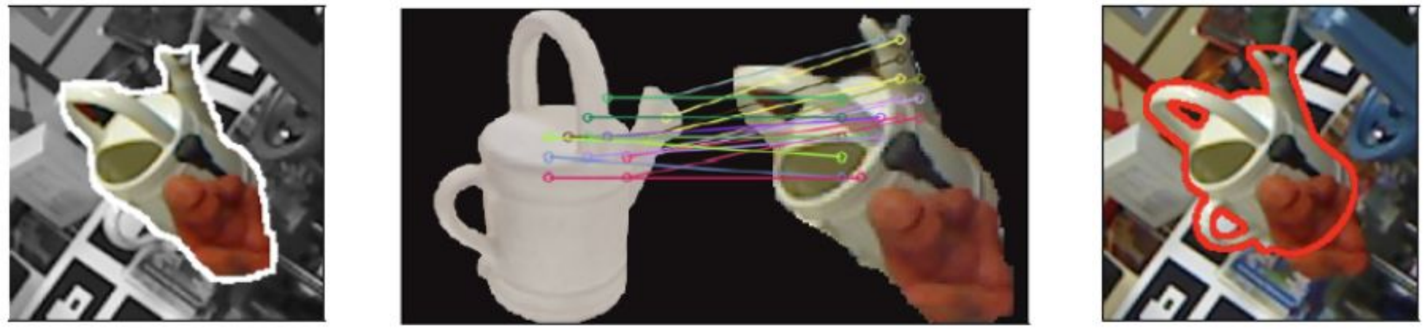
Input RGB | Prediction

[CVPR'22] Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects and Robustness to Occlusions



Query | Prediction | Query | Prediction

[CVPR'24] GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence



Query | Predicted 2D-2D correspondences | Prediction

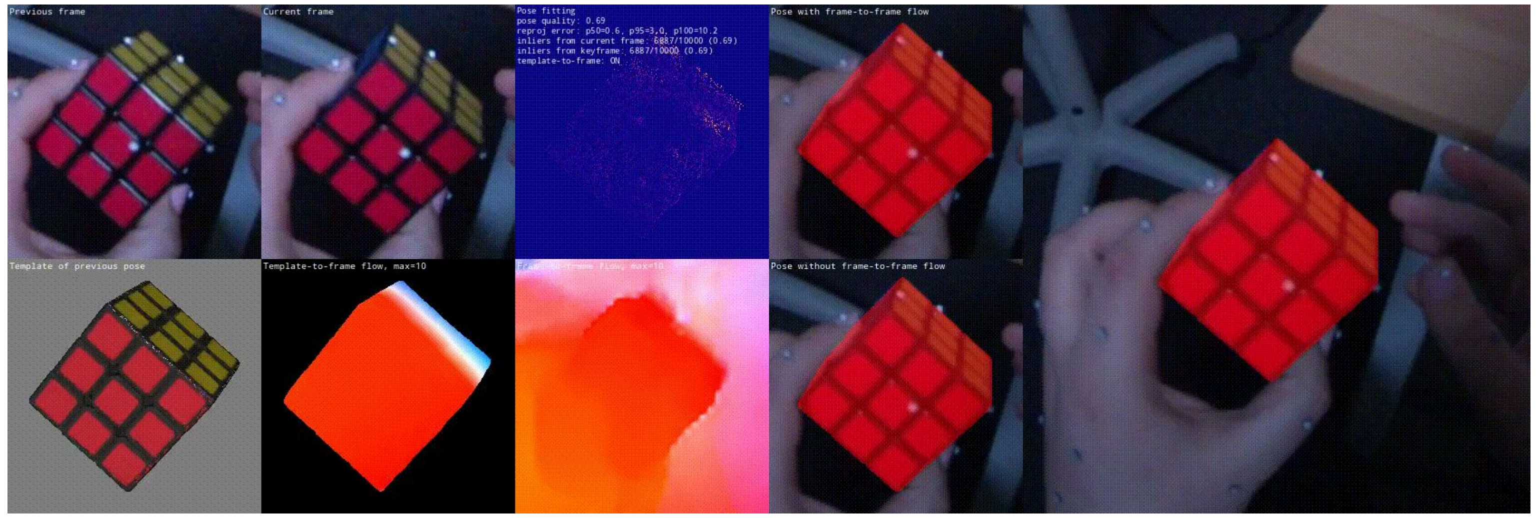
[CVPR'24] NOPE: Novel Object Pose Estimation from a Single Image

Reference | Query | Prediction | Predicted pose distribution

Other contributions

First-author contributions

PIZZA [3DV'2022 Oral] , GoTrack [In submission]: 6DoF pose tracking methods for novel objects



BOP challenge 2024 [ECCVW 2024]: Model-based, model-free detection, pose estimation of novel objects

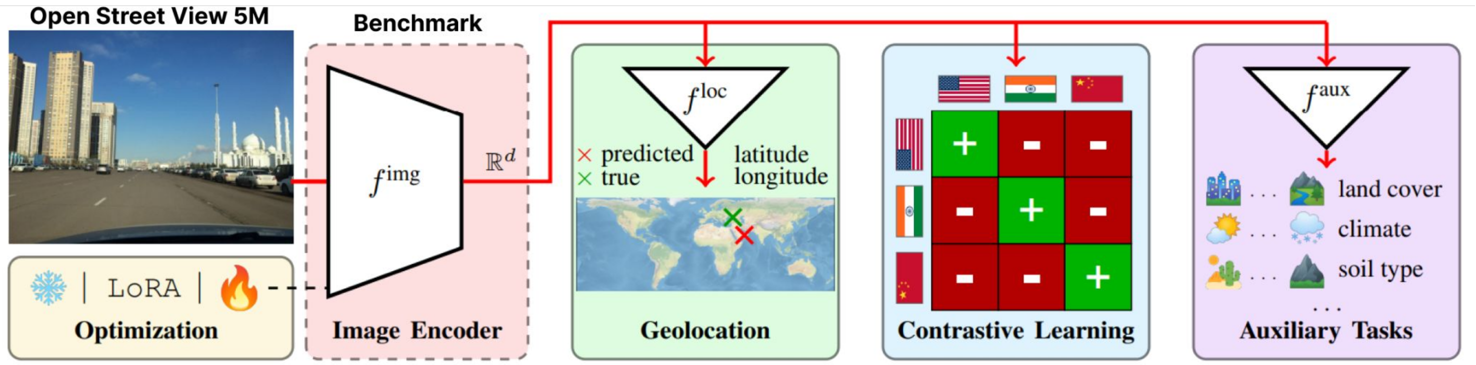
Co-author contributions

MaGIC-GS [In submission]: Monocular Articulated GenerIC object reconstruction with Gaussian Splatting

RGB-Video

Our reconstruction

OpenStreetView-5M [CVPR'24]: The Many Roads to Global Visual Geolocation

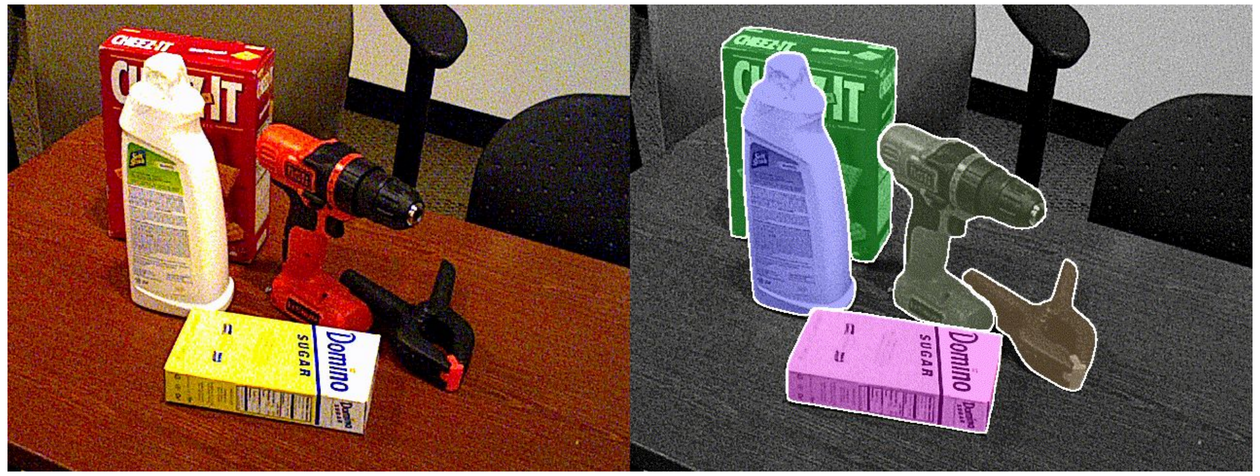


Outline

BOP 2023
4 awards

[ICCVW'23] CNOS: A Strong Baseline for CAD based Novel Object Segmentation

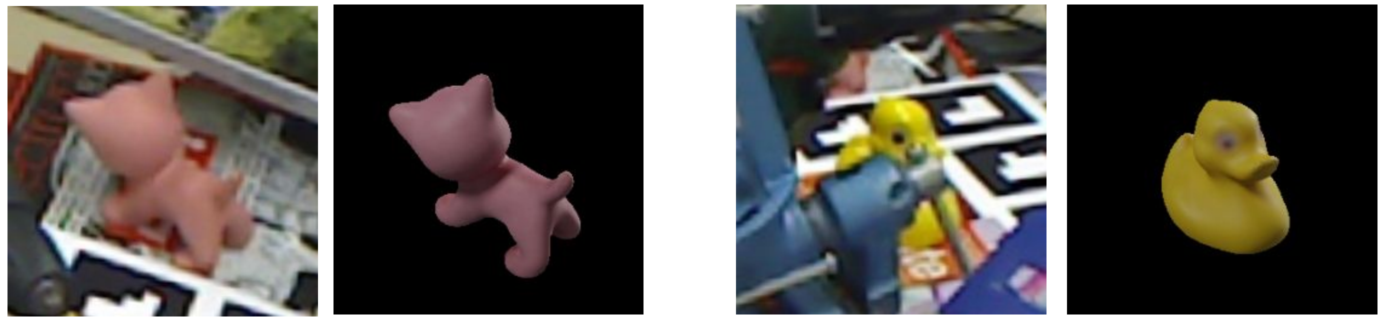
1



Input RGB   Prediction

[CVPR'22] Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects and Robustness to Occlusions

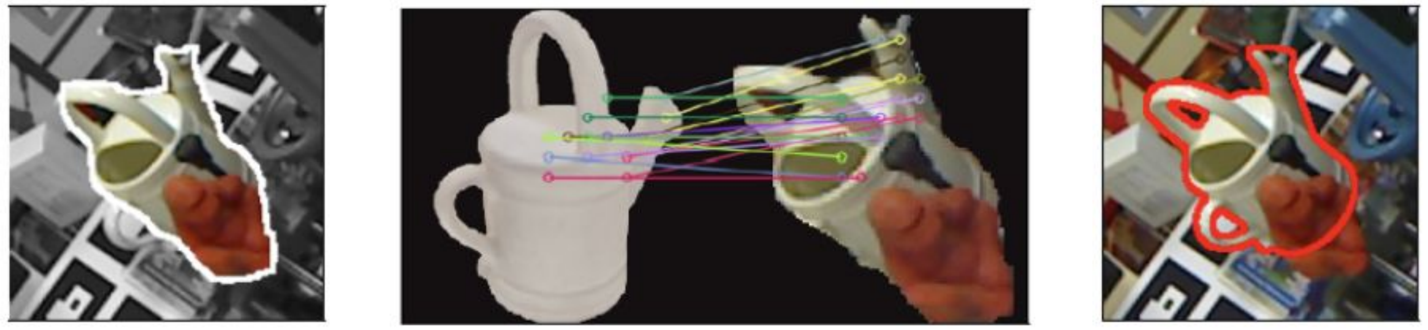
2



Query   Prediction   Query   Prediction

[CVPR'24] GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence

3



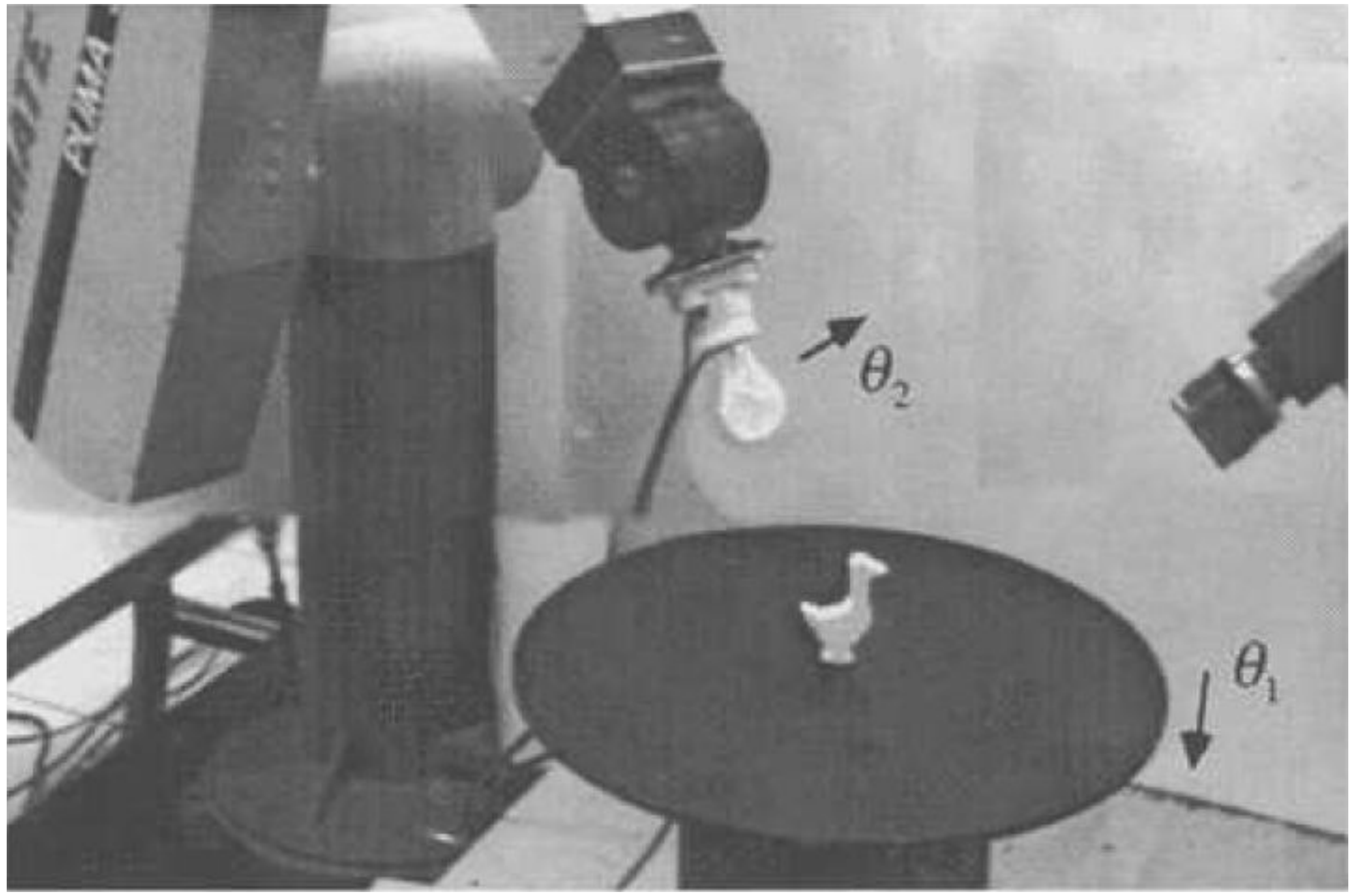
Query   Predicted 2D-2D correspondences   Prediction

[CVPR'24] NOPE: Novel Object Pose Estimation from a Single Image

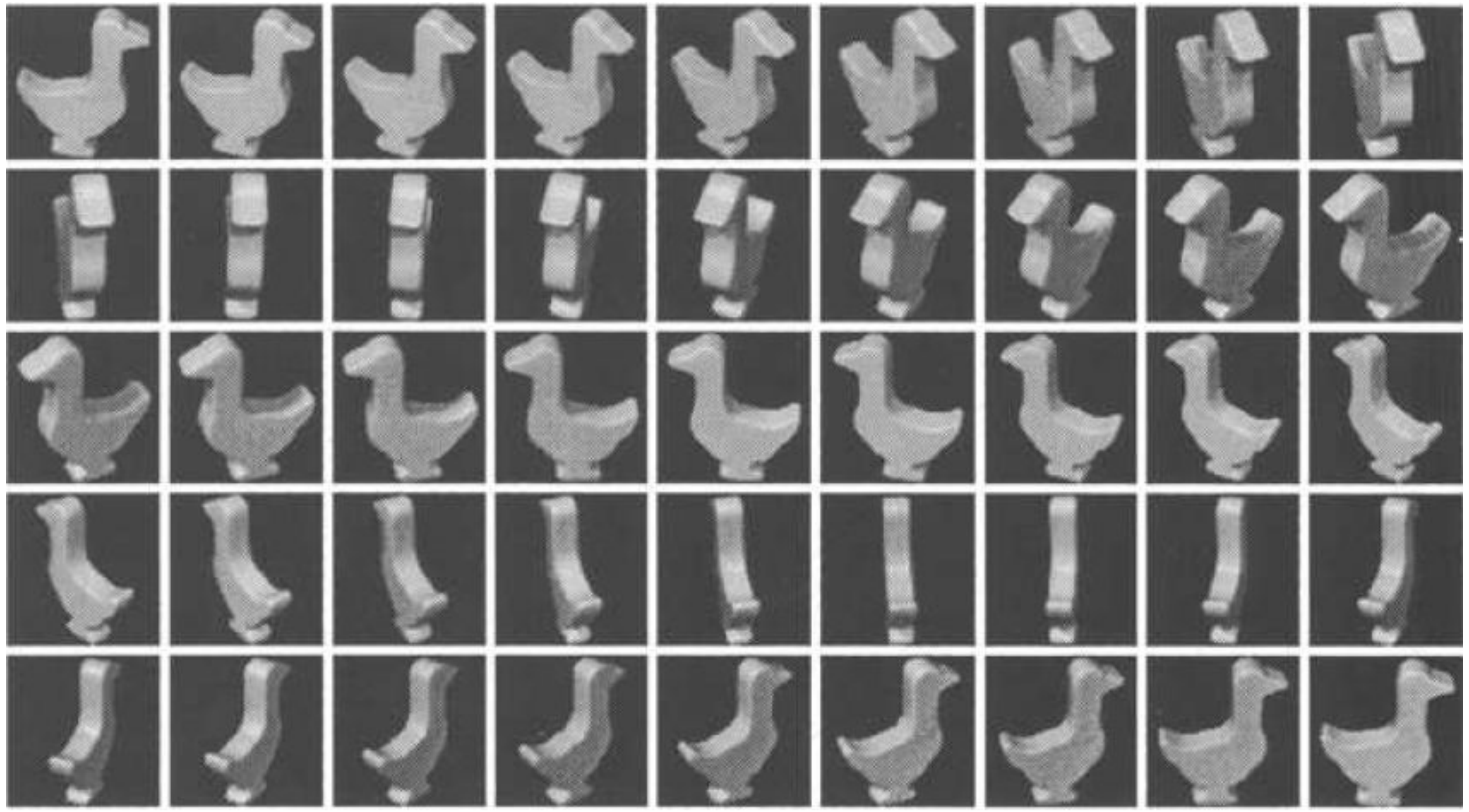
Reference   Query   Prediction   Predicted pose distribution

Templates for object pose estimation

Visual learning and recognition of 3D objects, Murase and Nayar, IJCV 1995



Robot and turntable to capture templates



Templates captured under different viewpoints

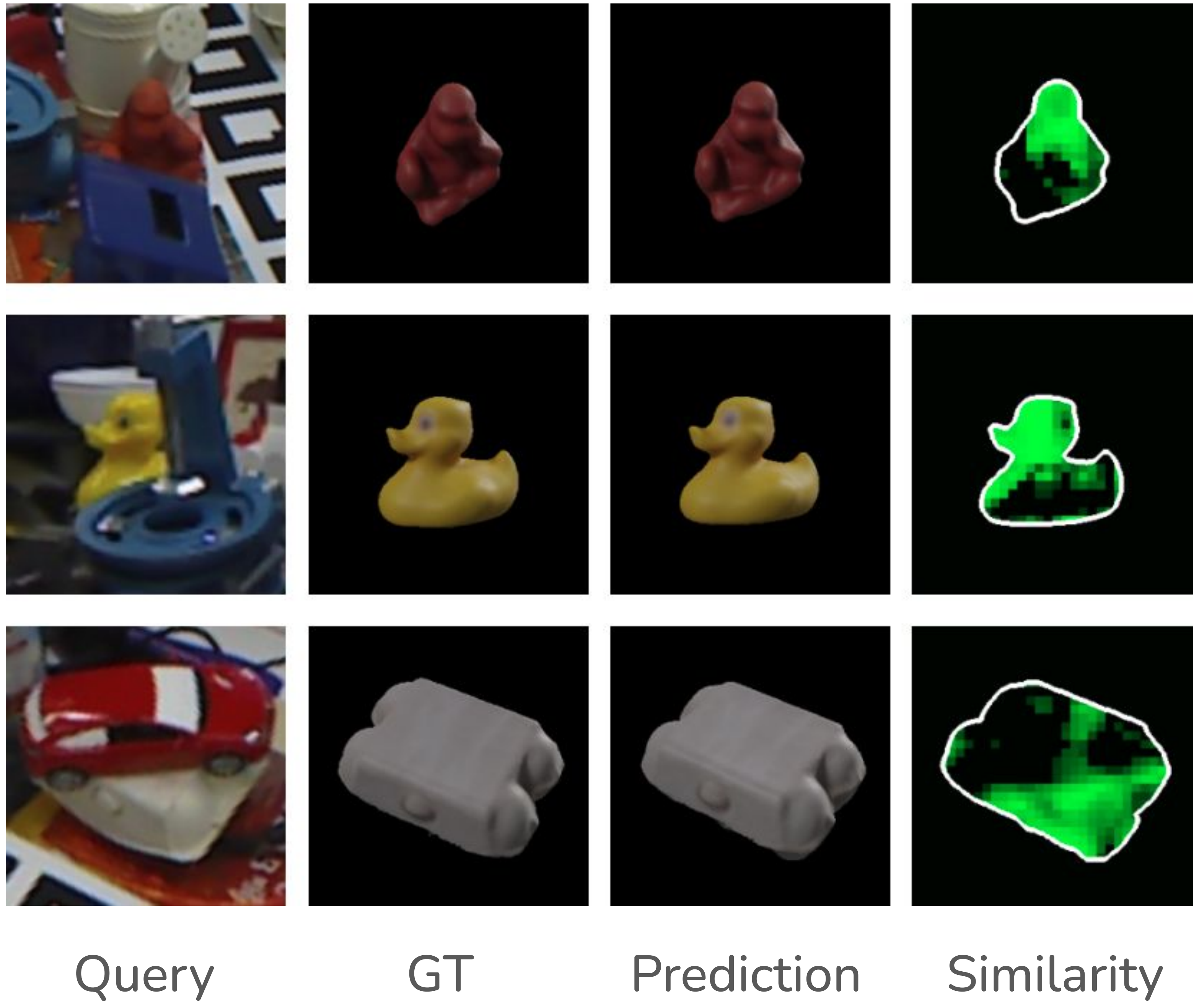
Templates for object pose estimation

- **Generalization to novel objects:** Templates generalize to novel objects, by extending the set of templates;

- **Robustness to domain gap and illumination**: by discriminative learning;

- **Robustness to clutter and partial occlusions:** image intensity correlation is very sensitive to partial occlusions, solved by representing with local feature vectors;

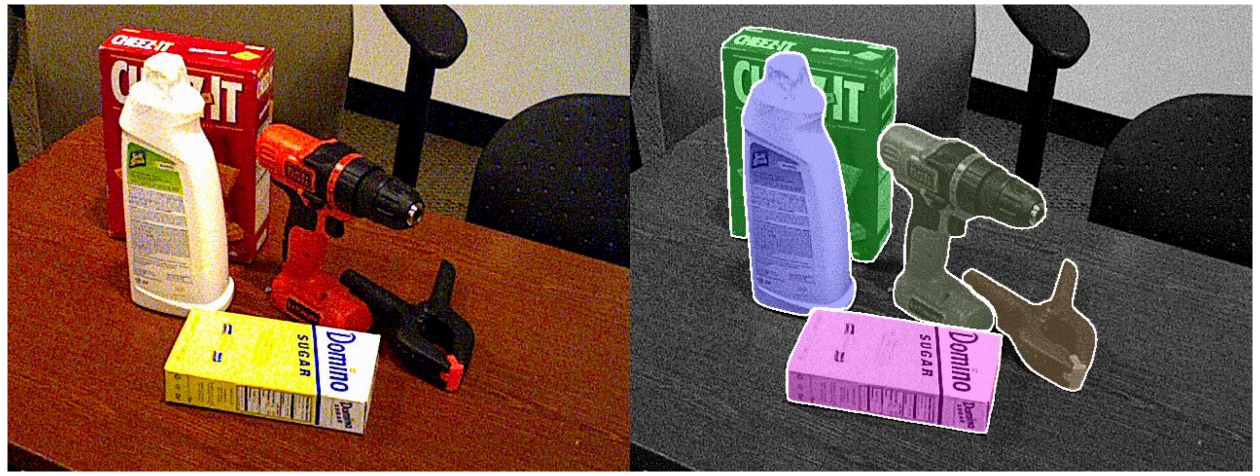
- **Robustness to the lack of texture:** in contrast with approaches based on local descriptors such as interest points, templates capture the appearance of an object as a whole.



Outline

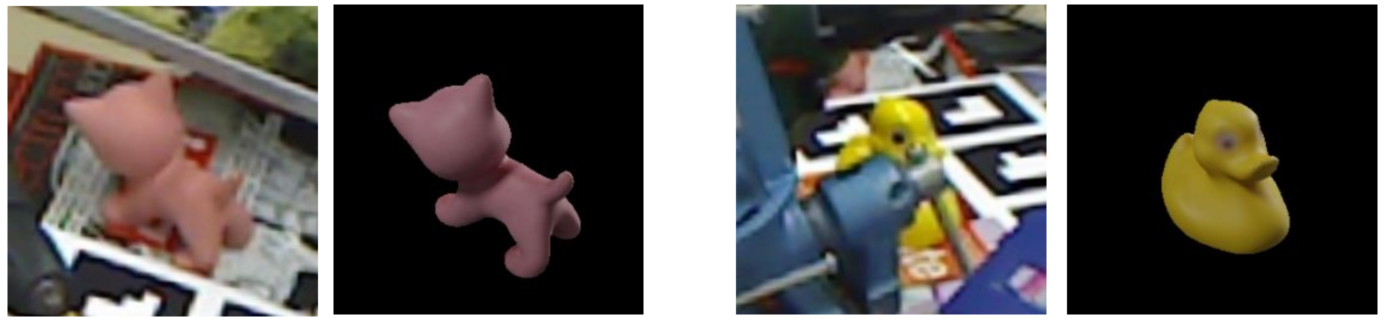
BOP 2023
4 awards

[ICCVW’23] CNOS: A Strong Baseline for CAD based Novel Object Segmentation



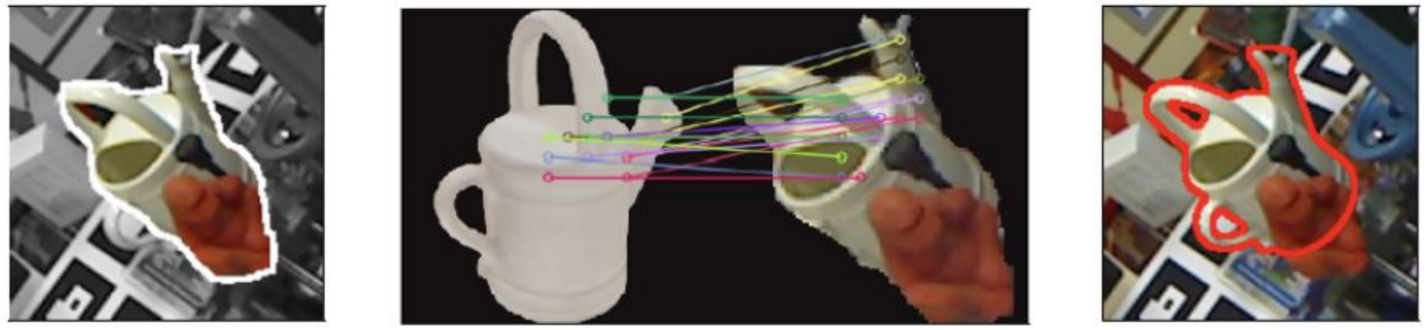
Input RGB | Prediction

[CVPR’22] Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects and Robustness to Occlusions



Query | Prediction | Query | Prediction

[CVPR’24] GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence



Query | Predicted 2D-2D correspondences | Prediction

[CVPR’24] NOPE: Novel Object Pose Estimation from a Single Image

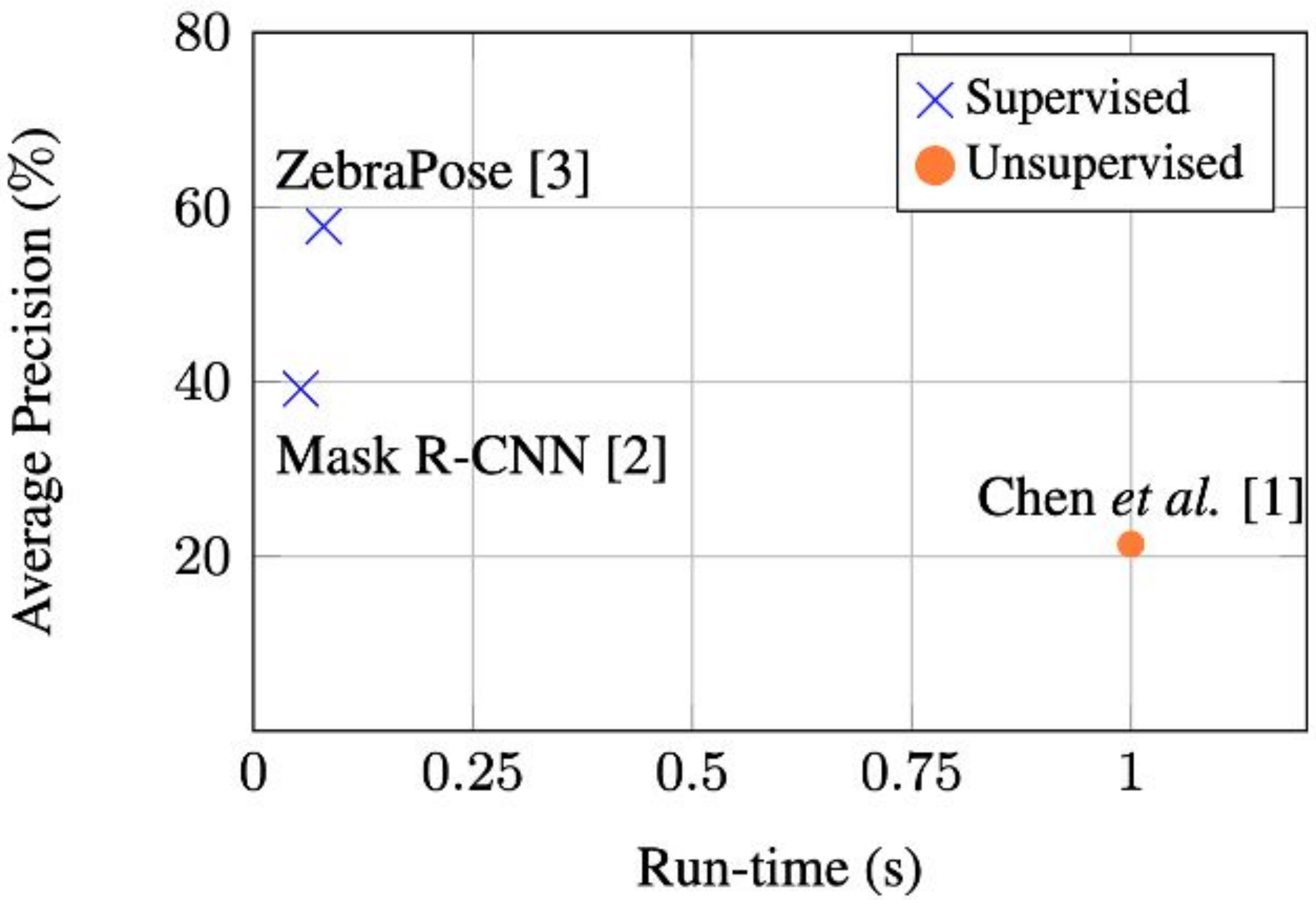
Reference | Query | Prediction | Predicted pose distribution

Motivation

RGB

2D detection/segmentation

6D pose

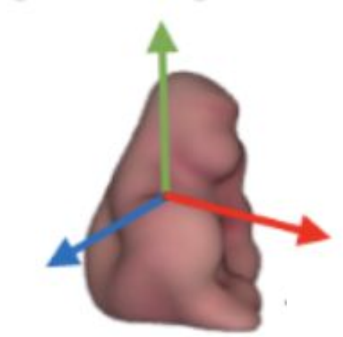


*supervised = test objects are seen during the training

*unsupervised = test objects are not seen during the training

Onboarding stage

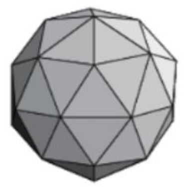
- Rendering templates from CAD models



3D model

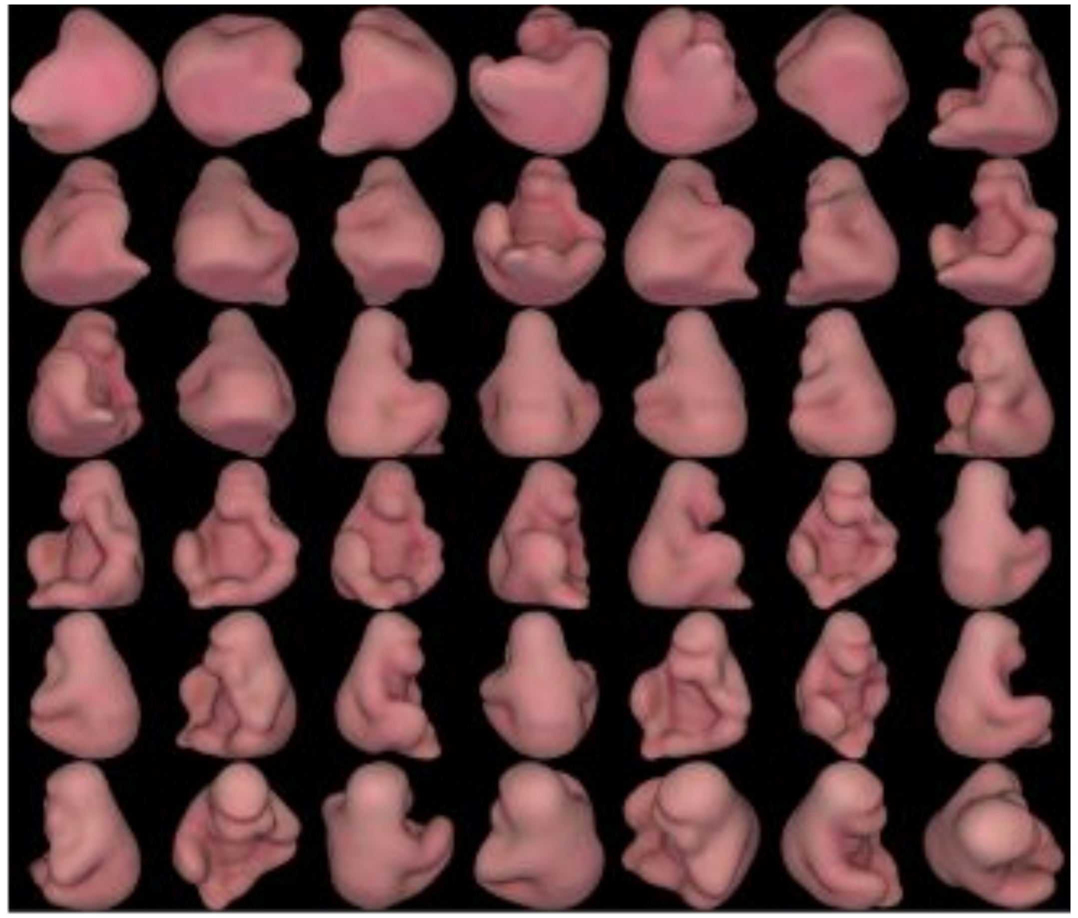
→

Offline rendering

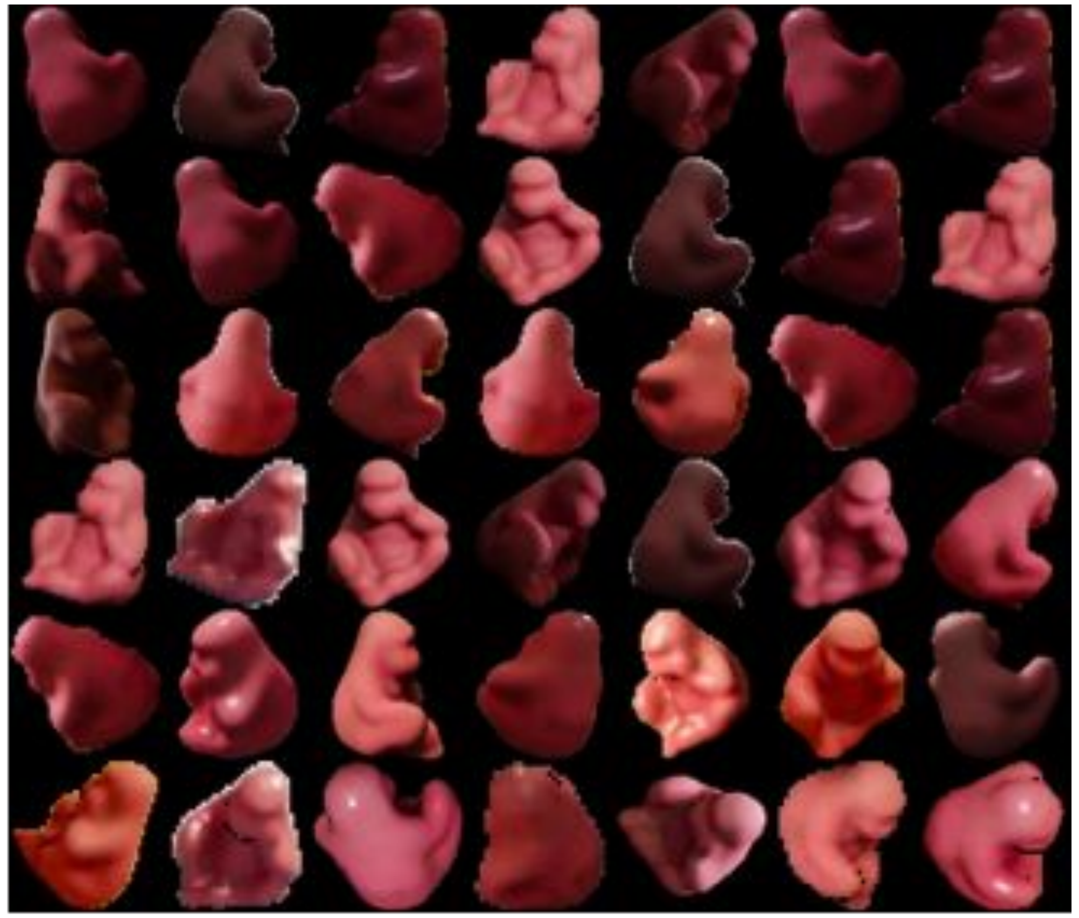


42 viewpoints

→



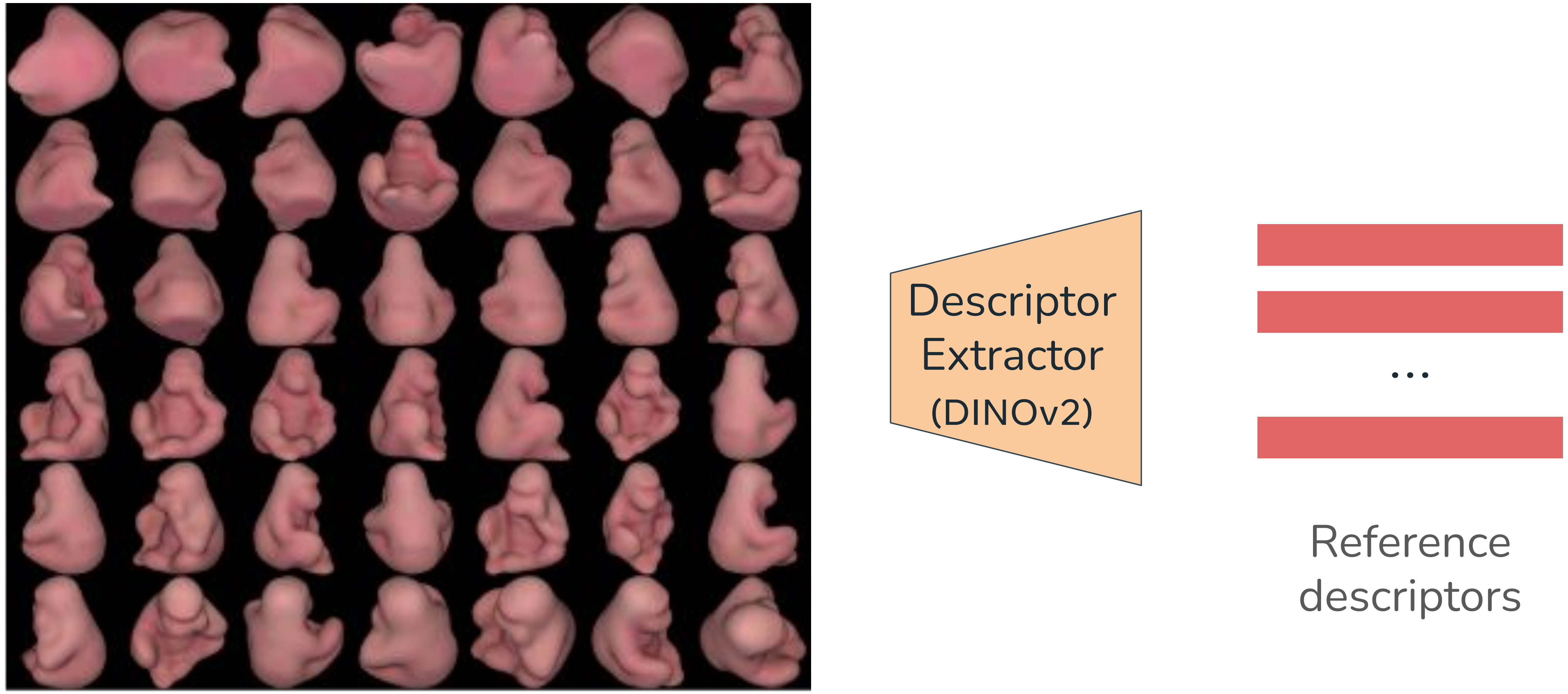
Templates rendered with Pyrender



Nearest templates generated with BlenderProc provided by BOP challenges

Onboarding stage

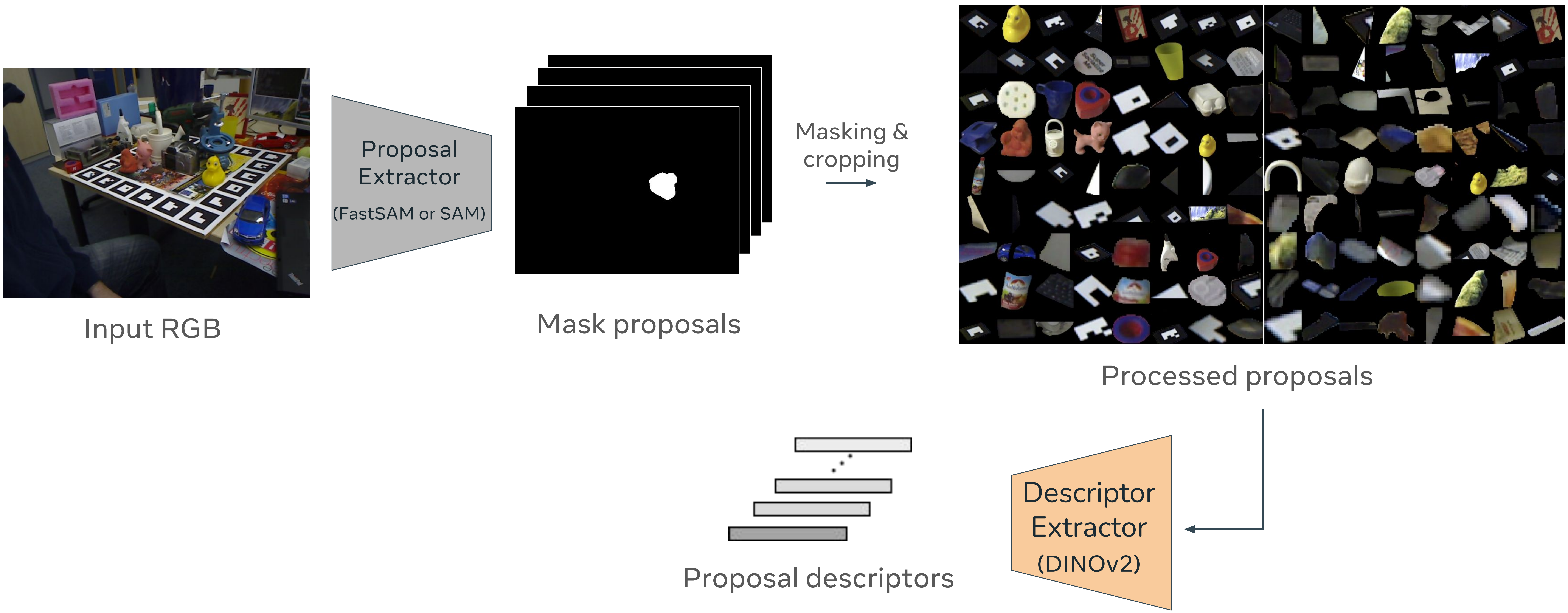
- Extracting visual descriptors (“cls” token of DINOv2) of templates



Templates

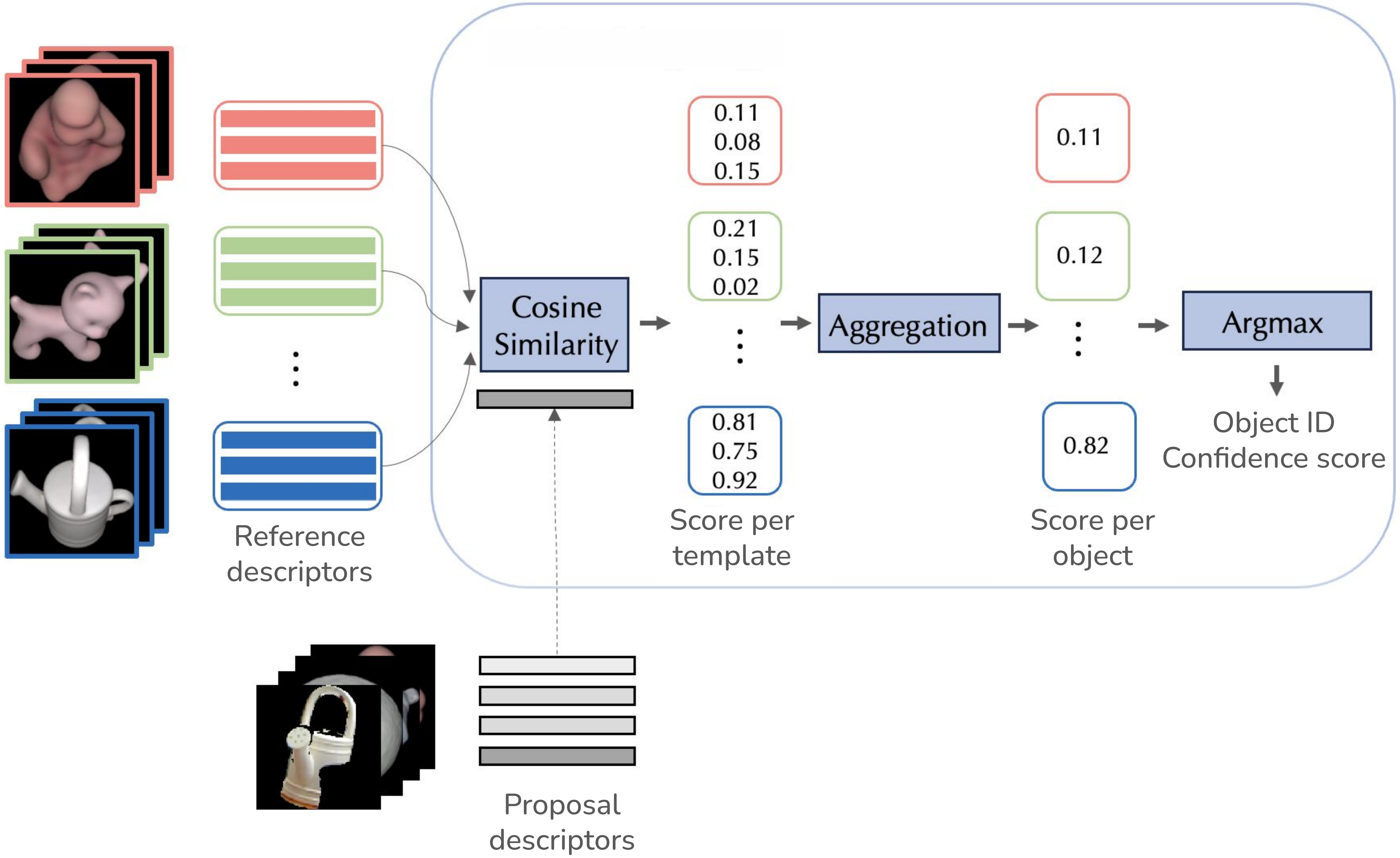
Proposal stage

- Extract all 2D masks from (Fast)SAM



Matching stage

- Finding object ID, confidence score for each proposal



Results

(Frame by frame prediction)

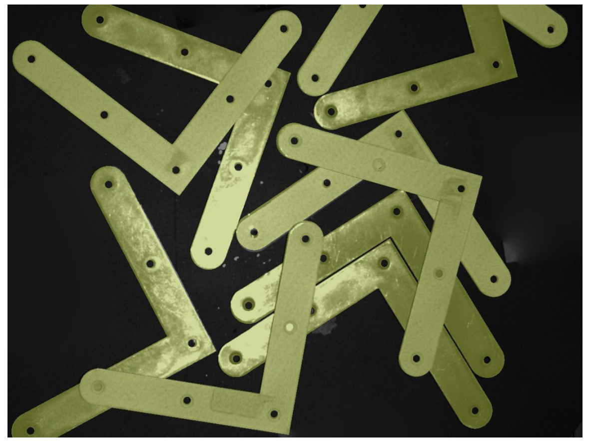
Evaluation protocol

- Training datasets: not required (CNOS is training-free)
- Testing datasets: 7 datasets of BOP-Classic-Core
- Evaluation metrics: COCO evaluation protocol (AP with IoU thresholds in [0.5, …, 0.95])

LM-O

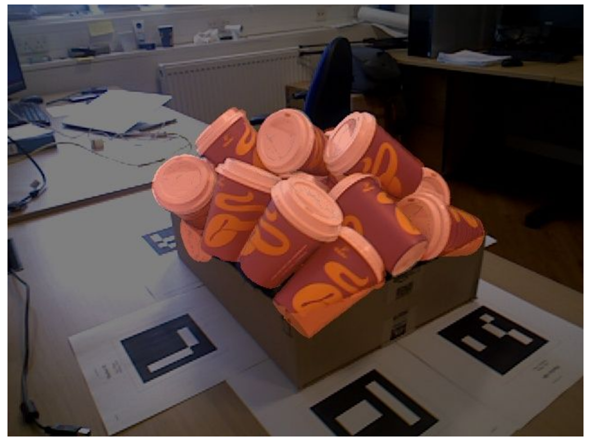
T-LESS

HB



ITODD

YCB-V

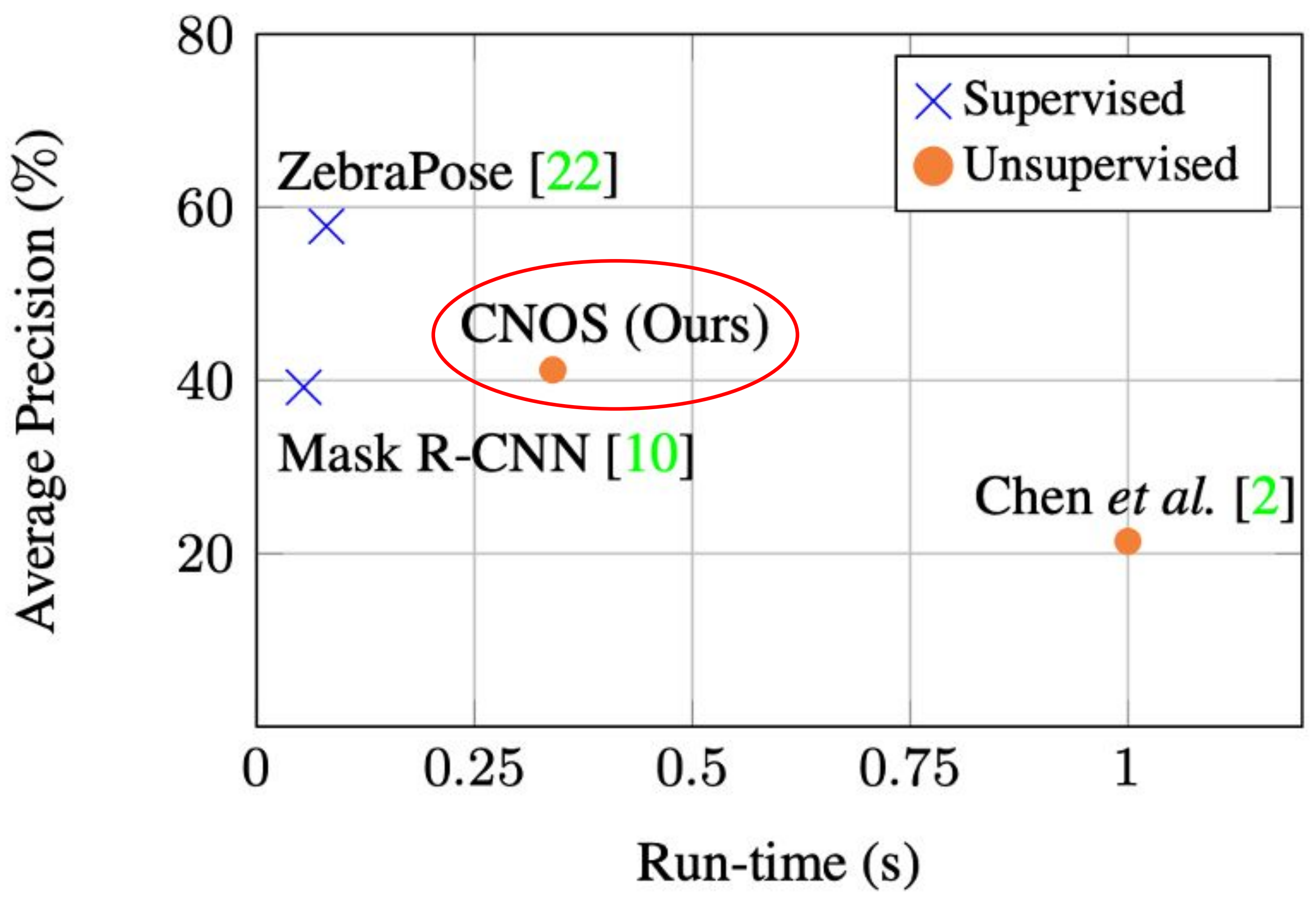


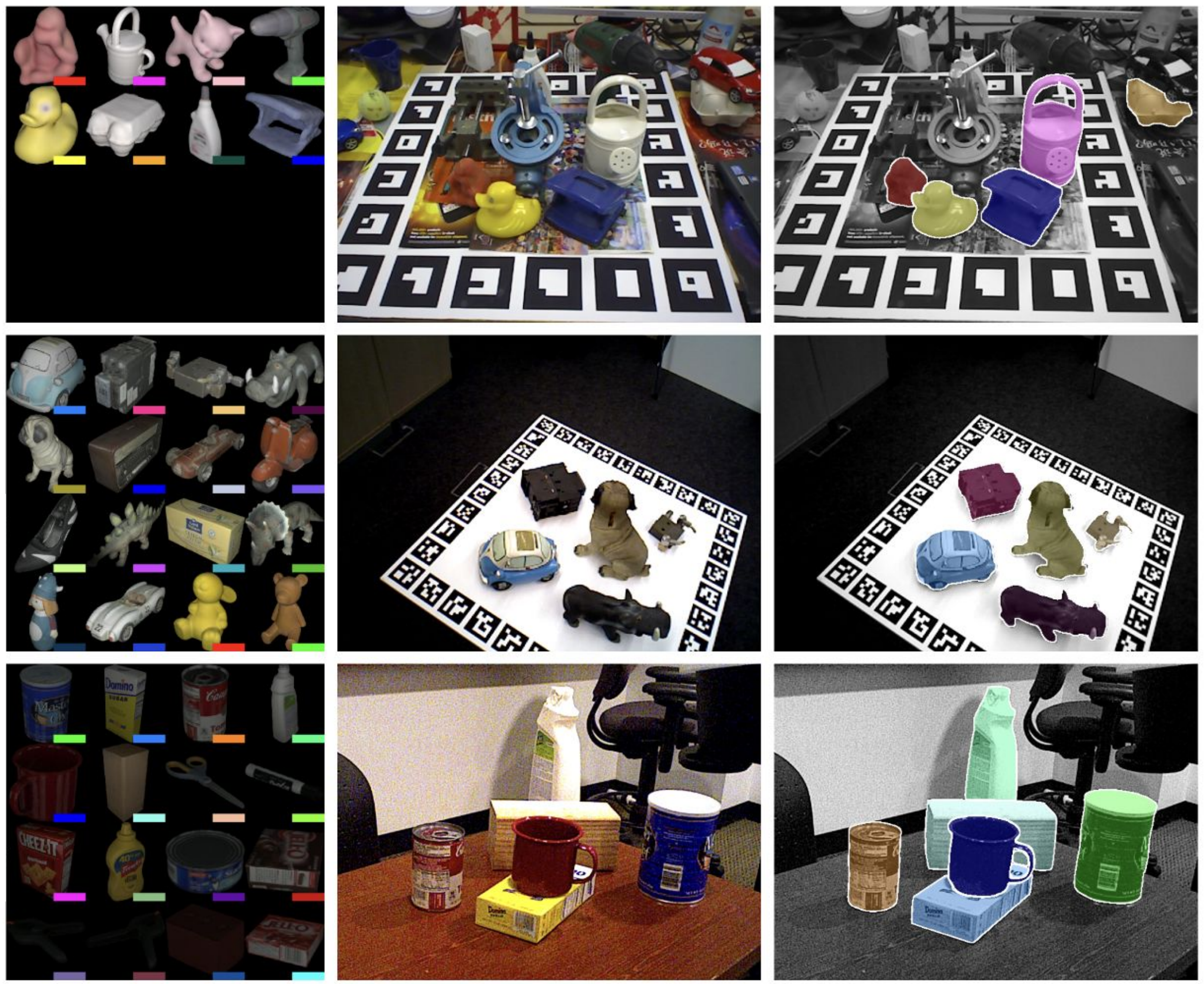
IC-BIN

TUD-L

Results

- **Accuracy**: CNOS improves SOTA by absolute 19.8% AP, outperforms the supervised Mask R-CNN;
- **Run-time:** CNOS (FastSAM) outperforms CNOS (SAM) by absolute 0.8% AP while being 7x faster;
- **Domain gap:** BlenderProc reduces the domain gap, with 4.3% AP improvement compared to Pyrender.



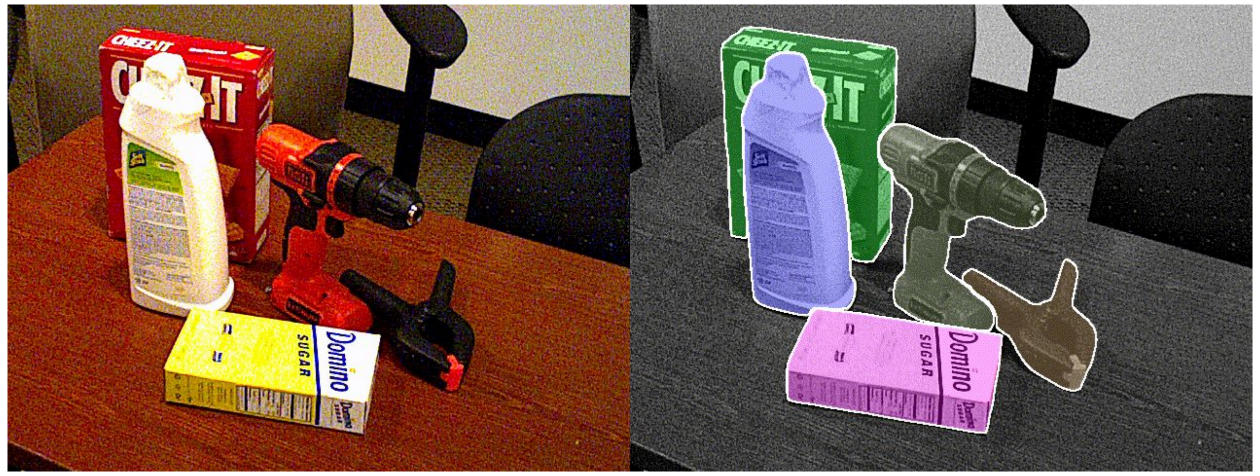


Input 3D models | Input RGB | Prediction (confidence > 0.5)

Outline

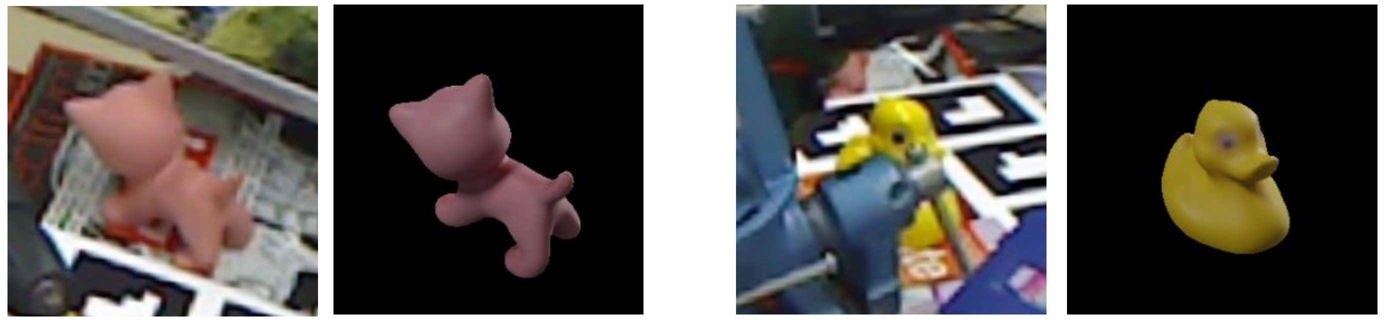
BOP 2023
4 awards

[ICCVW'23] CNOS: A Strong Baseline for CAD based Novel Object Segmentation



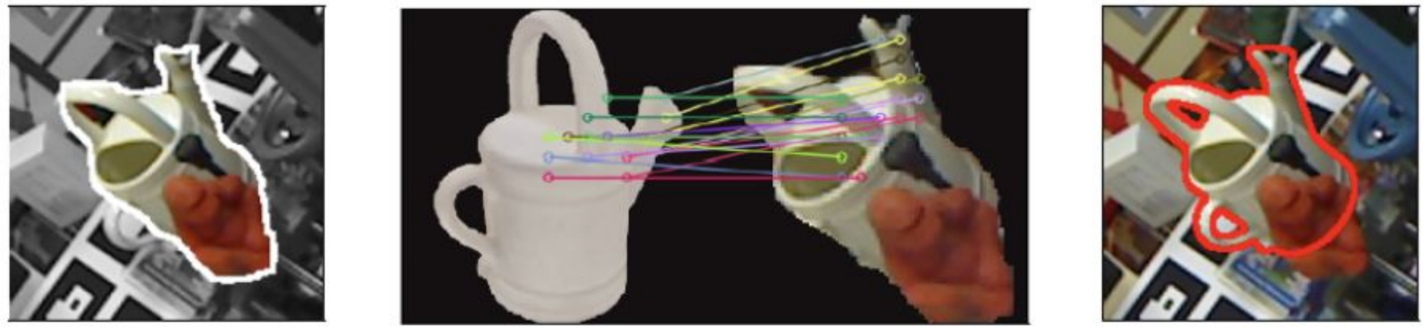
Input RGB Prediction

[CVPR'22] Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects and Robustness to Occlusions



Query Prediction Query Prediction

[CVPR'24] GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence



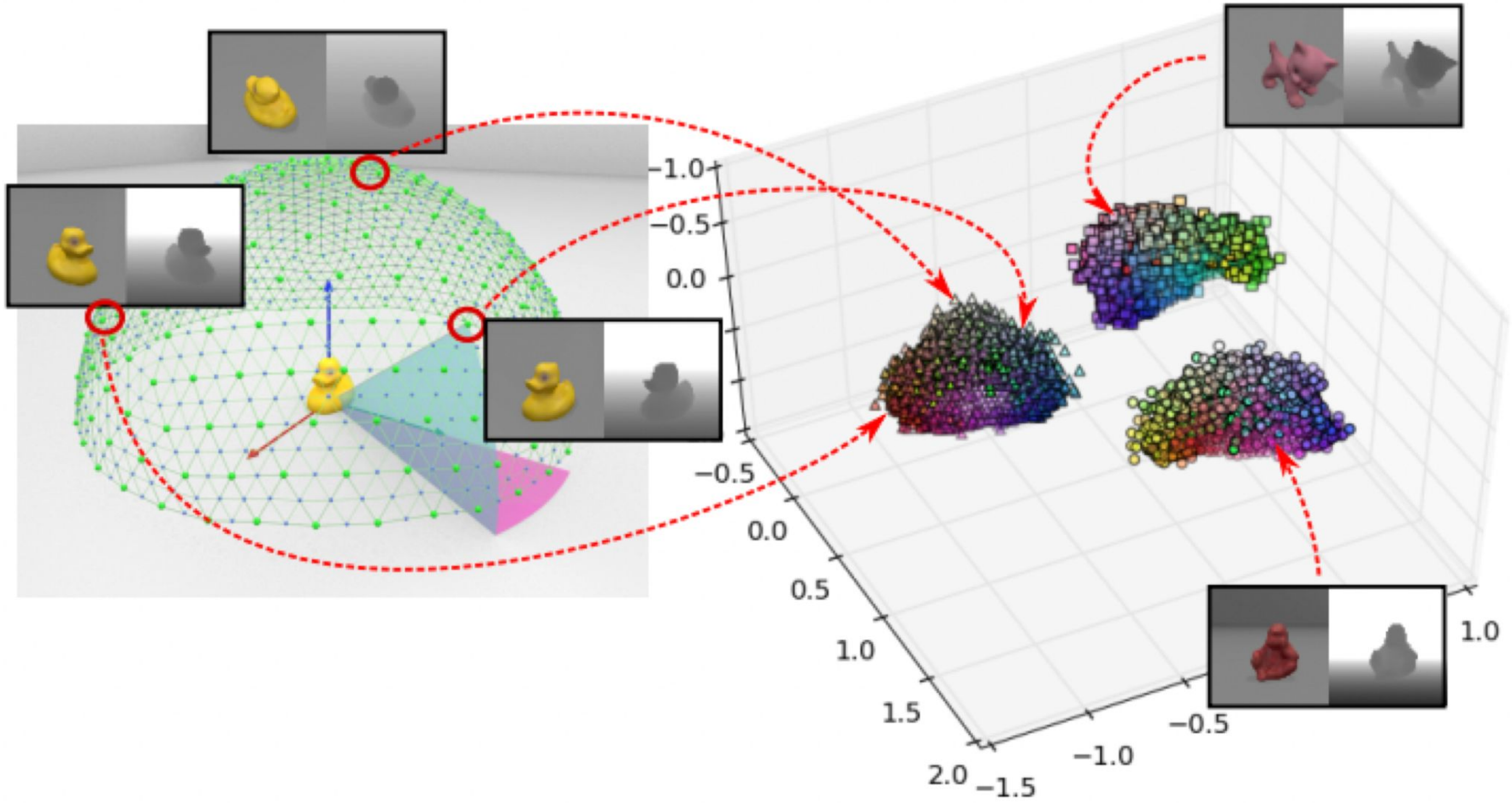
Query Predicted 2D-2D correspondences Prediction

[CVPR'24] NOPE: Novel Object Pose Estimation from a Single Image

Reference Query Prediction Predicted pose distribution

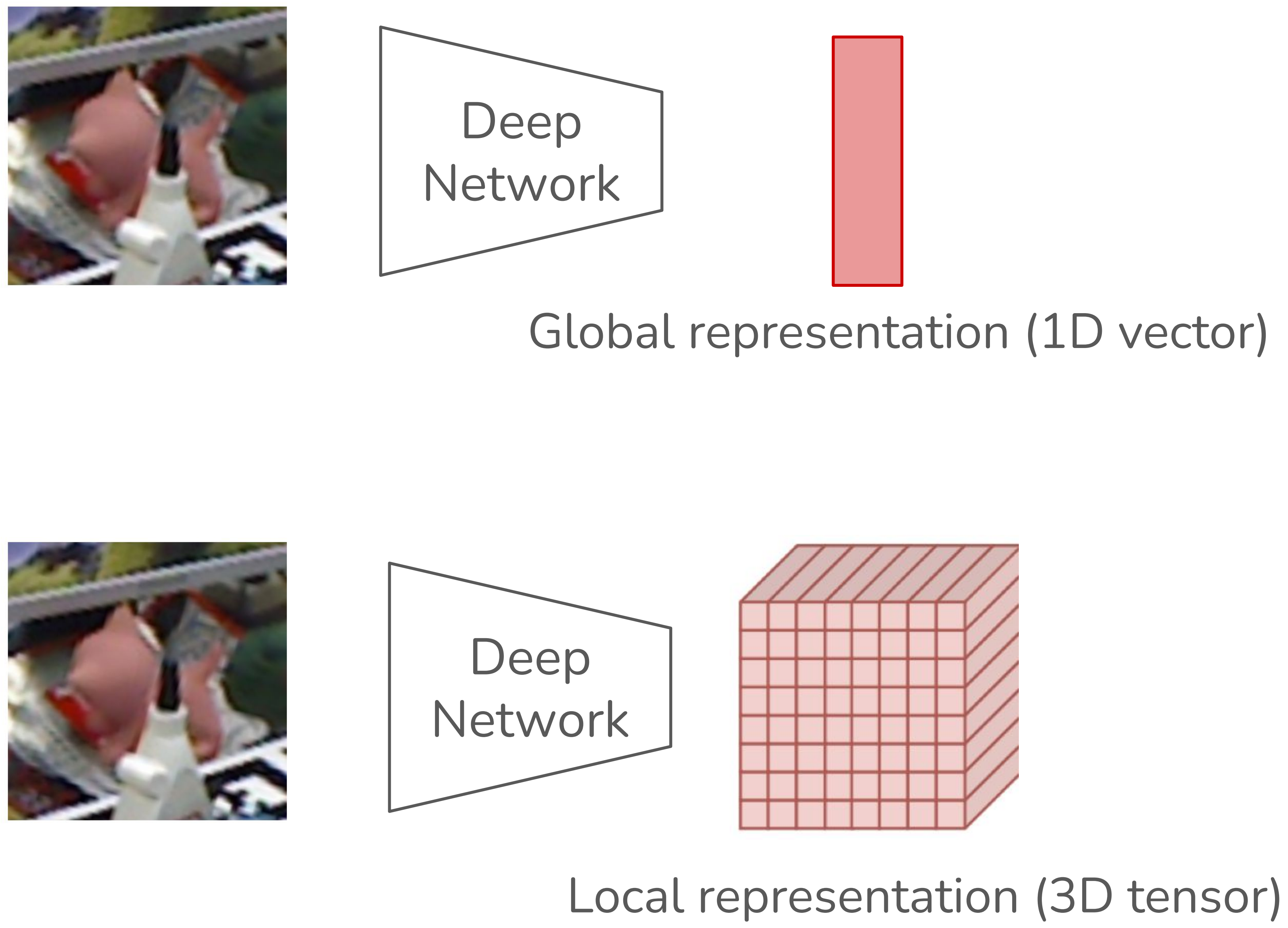
Template matching for 3D pose estimation

- “Good” image representation for pose estimation: Object discrimination and pose discrimination



Learning Descriptors for Object Recognition and 3D Pose Estimation, Wolhart and Lepetit, CVPR 2015

Global representation vs local representation



Failure cases of global representation

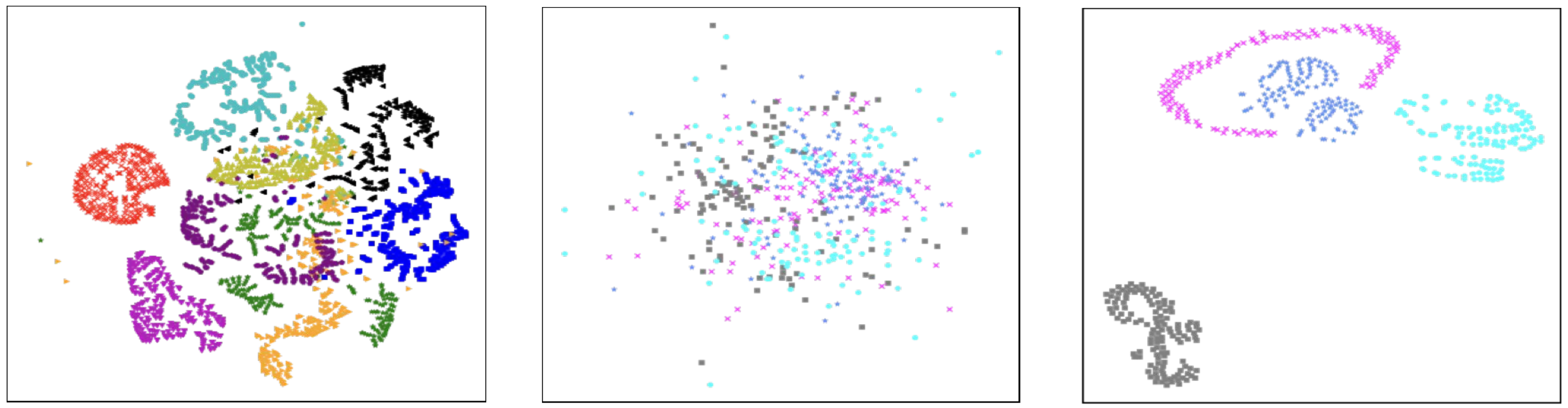
Training objects

ape	can	iron
bench.	glue	lamp
cam	hole.	phone

Testing (unseen) objects

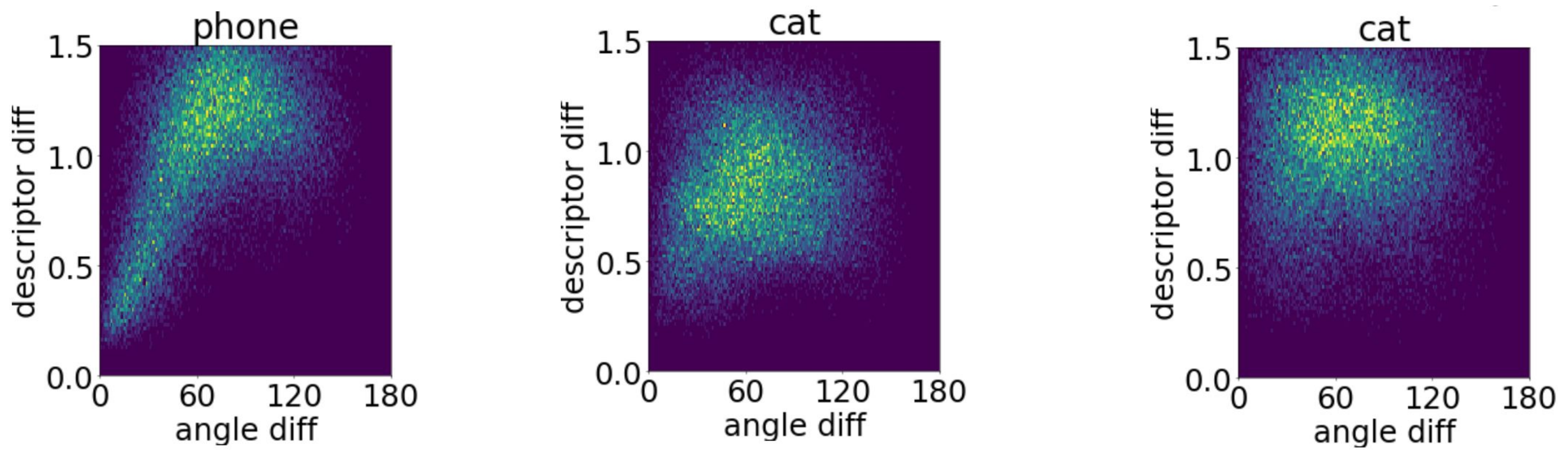
cat	duck
driller	eggbox

Seen objects | Unseen objects | After masking background



Object discrimination
(t-SNE visualization)

Seen objects | Unseen objects | After masking background



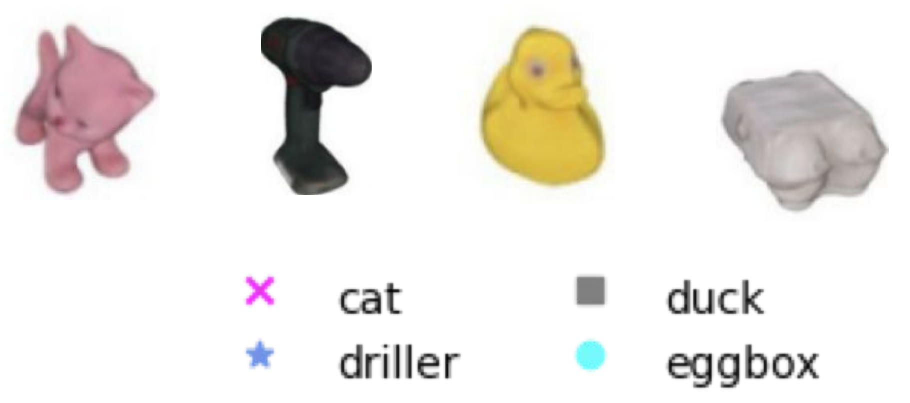
Pose discrimination

Learning Descriptors for Object Recognition and 3D Pose Estimation, Wolhart and Lepetit, CVPR 2015

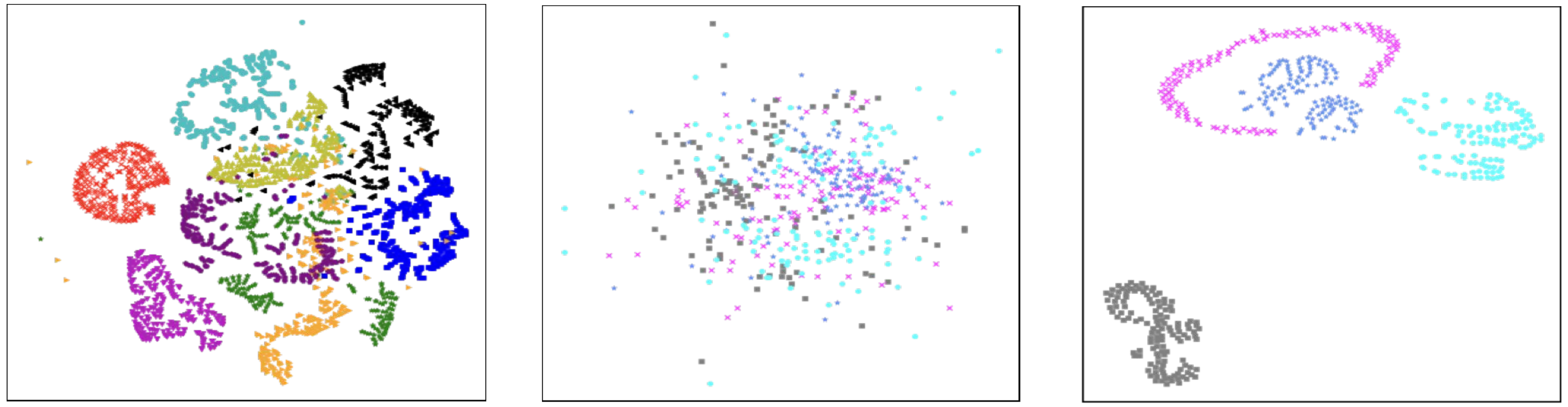
Failure cases of global representation

Training objects

Testing (unseen) objects

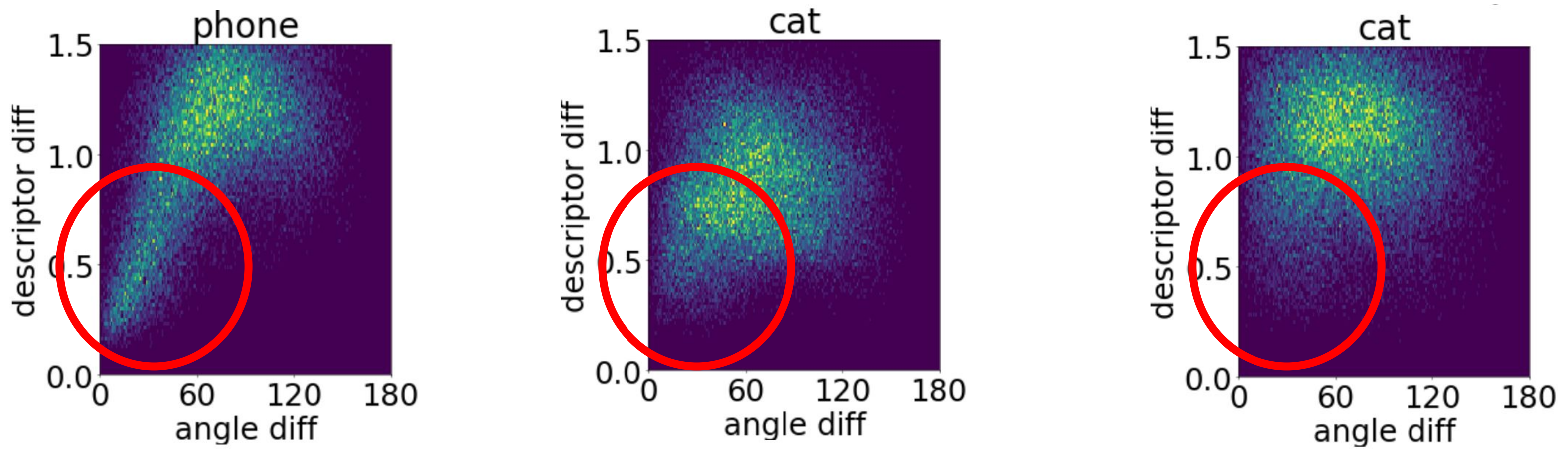


Seen objects | Unseen objects | After masking background



Object discrimination

Seen objects | Unseen objects | After masking background



Pose discrimination

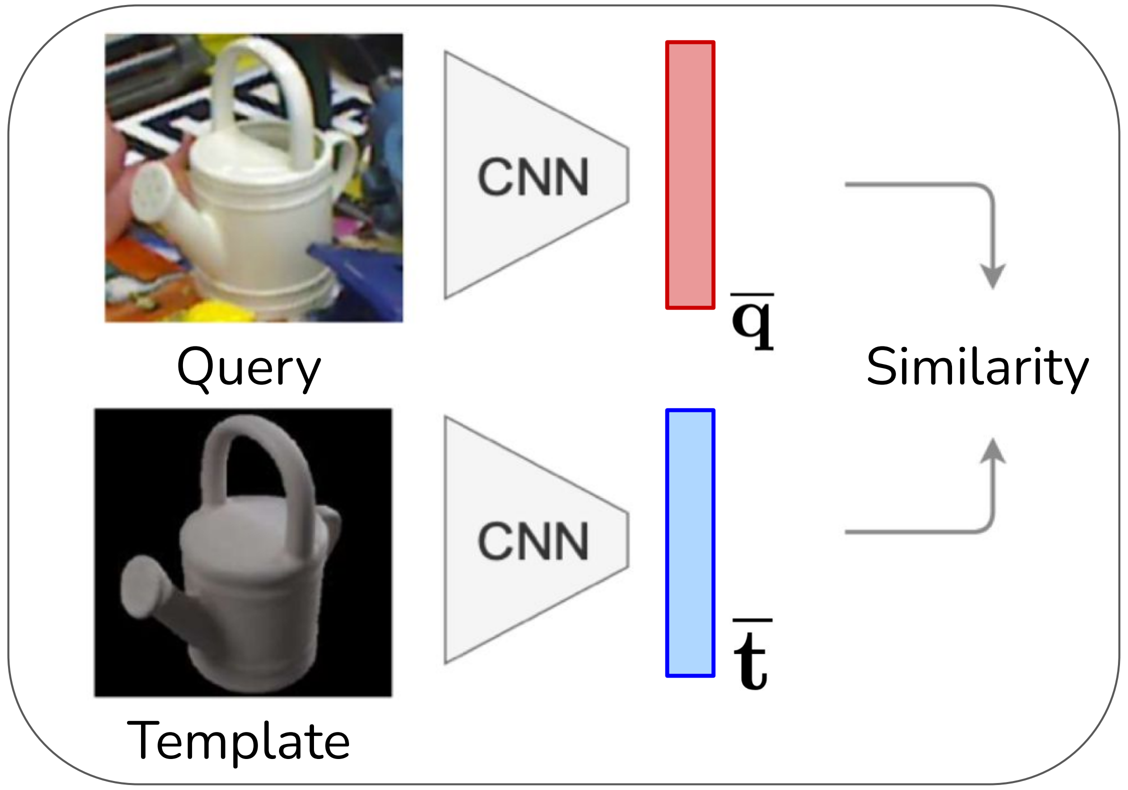
-> Fail on unseen objects in presence of cluttered background

-> No pose information in descriptors

Pose Guided RGB D Feature Learning for 3D Object Pose Estimation, Balntas et al., ICCV 2017

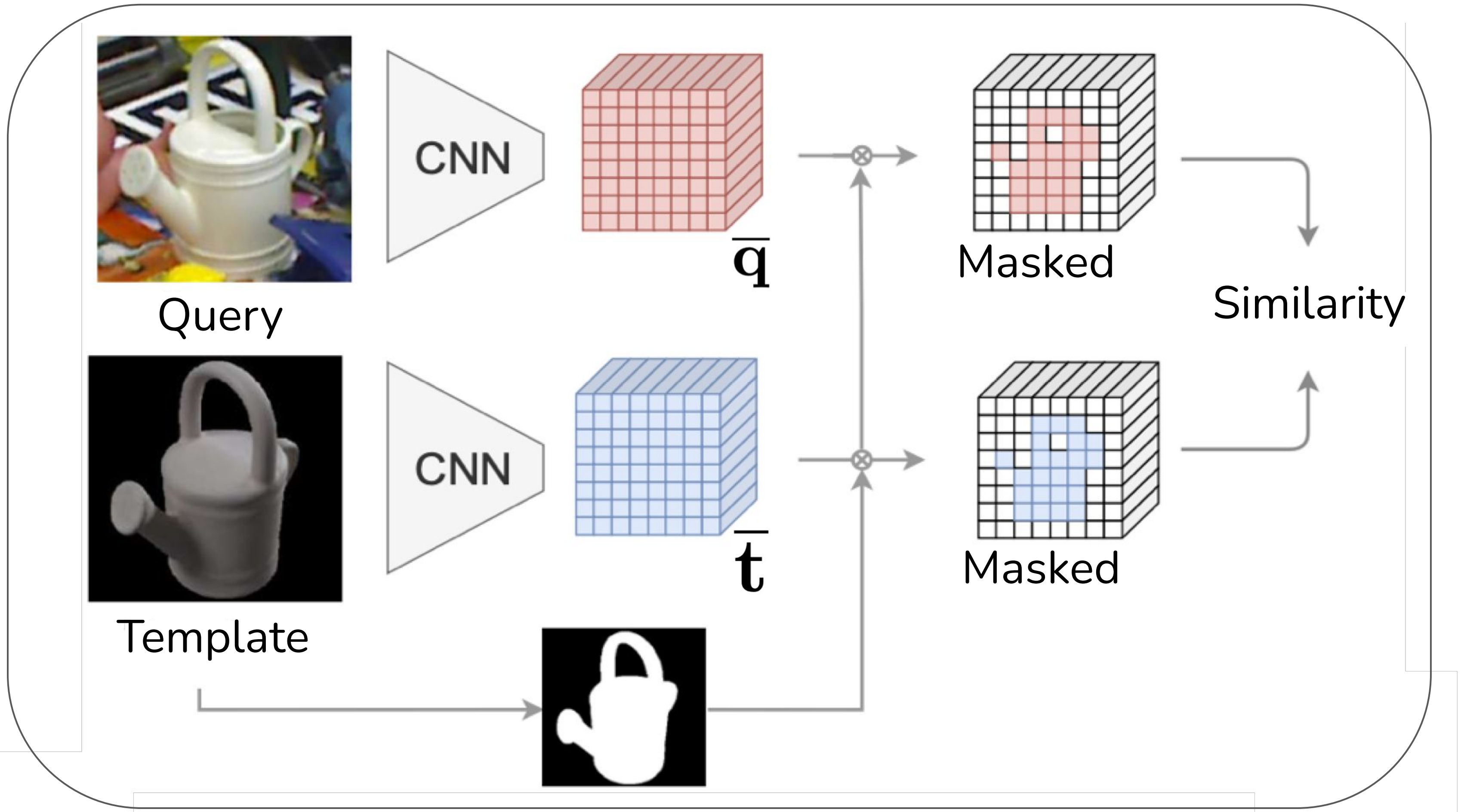
Our approach: Using local representation & template mask

Previous works [1,2]



$$\text{sim}\left(\overline{\mathbf{q}}, \overline{\mathbf{t}}\right) = \frac{\overline{\mathbf{q}}}{||\overline{\mathbf{q}}||} \cdot \frac{\overline{\mathbf{t}}}{||\overline{\mathbf{t}}||}$$

Our proposed method

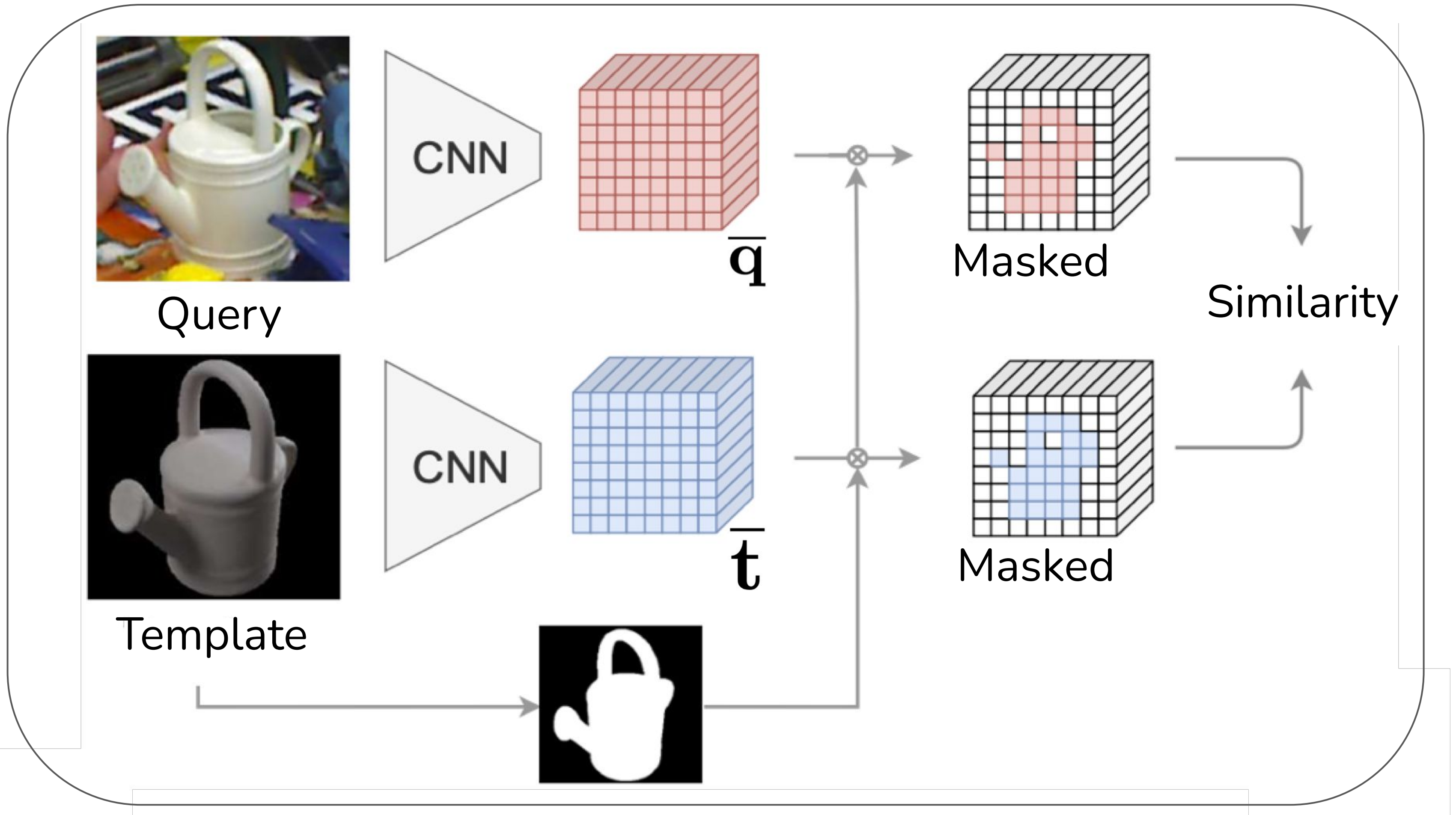


$$\text{sim}(\overline{\mathbf{q}}, \overline{\mathbf{t}}) = \frac{1}{|\mathcal{M}|} \sum_{l} \mathcal{M}^{(l)} \mathcal{S}\left(\overline{\mathbf{q}}^{(l)}, \overline{\mathbf{t}}^{(l)}\right)$$

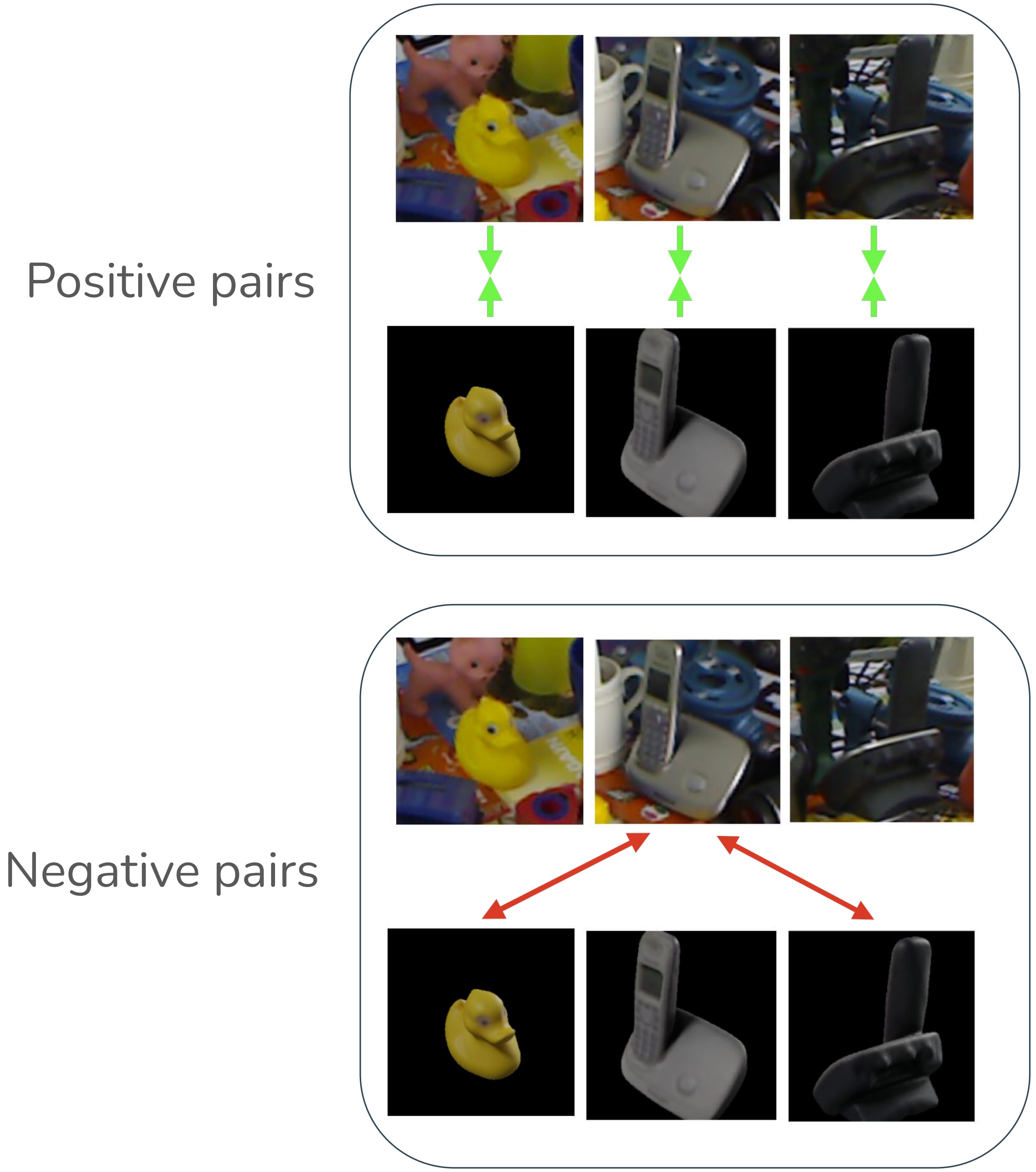
where $\mathcal{S}\left(\overline{\mathbf{q}}^{(l)}, \overline{\mathbf{t}}^{(l)}\right) = \frac{\overline{\mathbf{q}}^{(l)}}{||\overline{\mathbf{q}}^{(l)}||} \cdot \frac{\overline{\mathbf{t}}^{(l)}}{||\overline{\mathbf{t}}^{(l)}||}$

[1] Learning Descriptors for Object Recognition and 3D Pose Estimation, Wohlhart and Lepetit, CVPR 2015
[2] Pose Guided RGB D Feature Learning for 3D Object Pose Estimation, Balntas et al., ICCV 2017

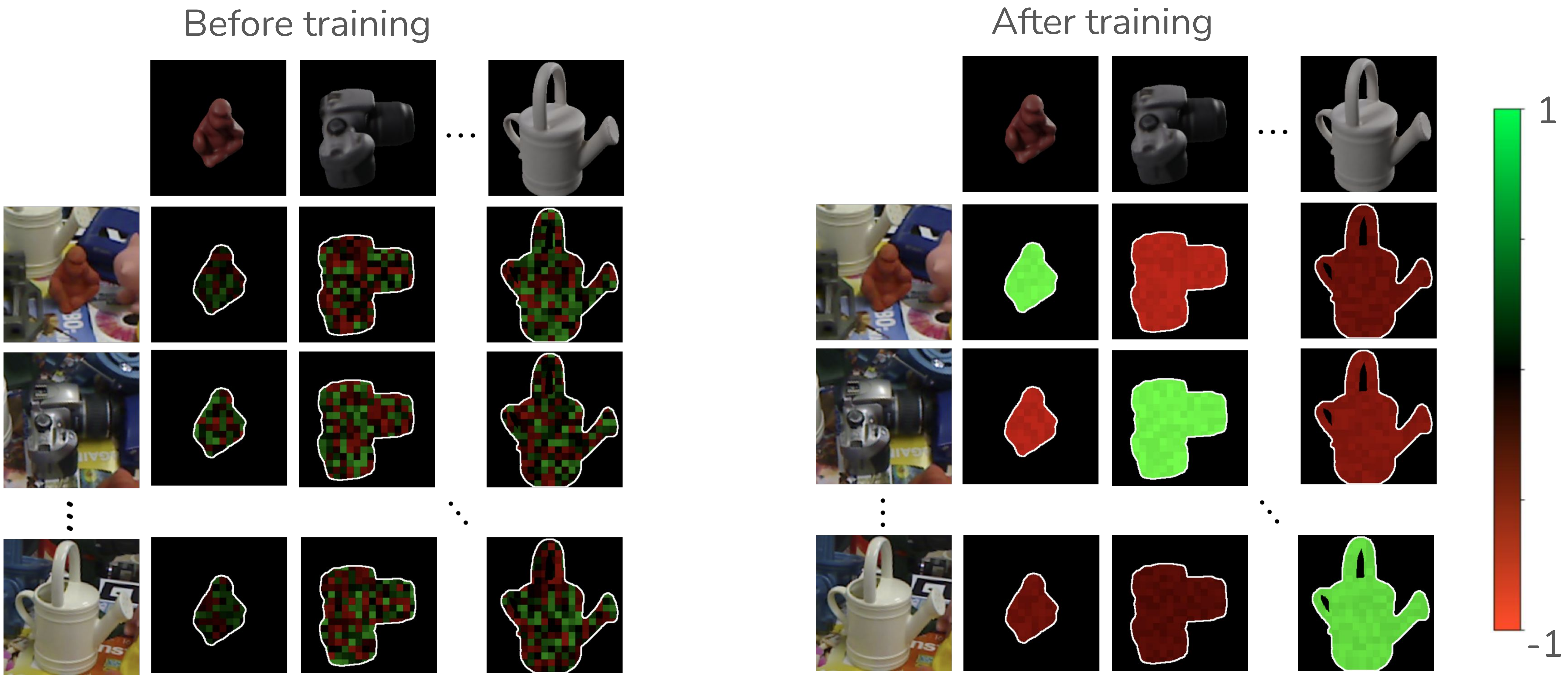
Training samples: positive pairs vs negative pairs



Our proposed method

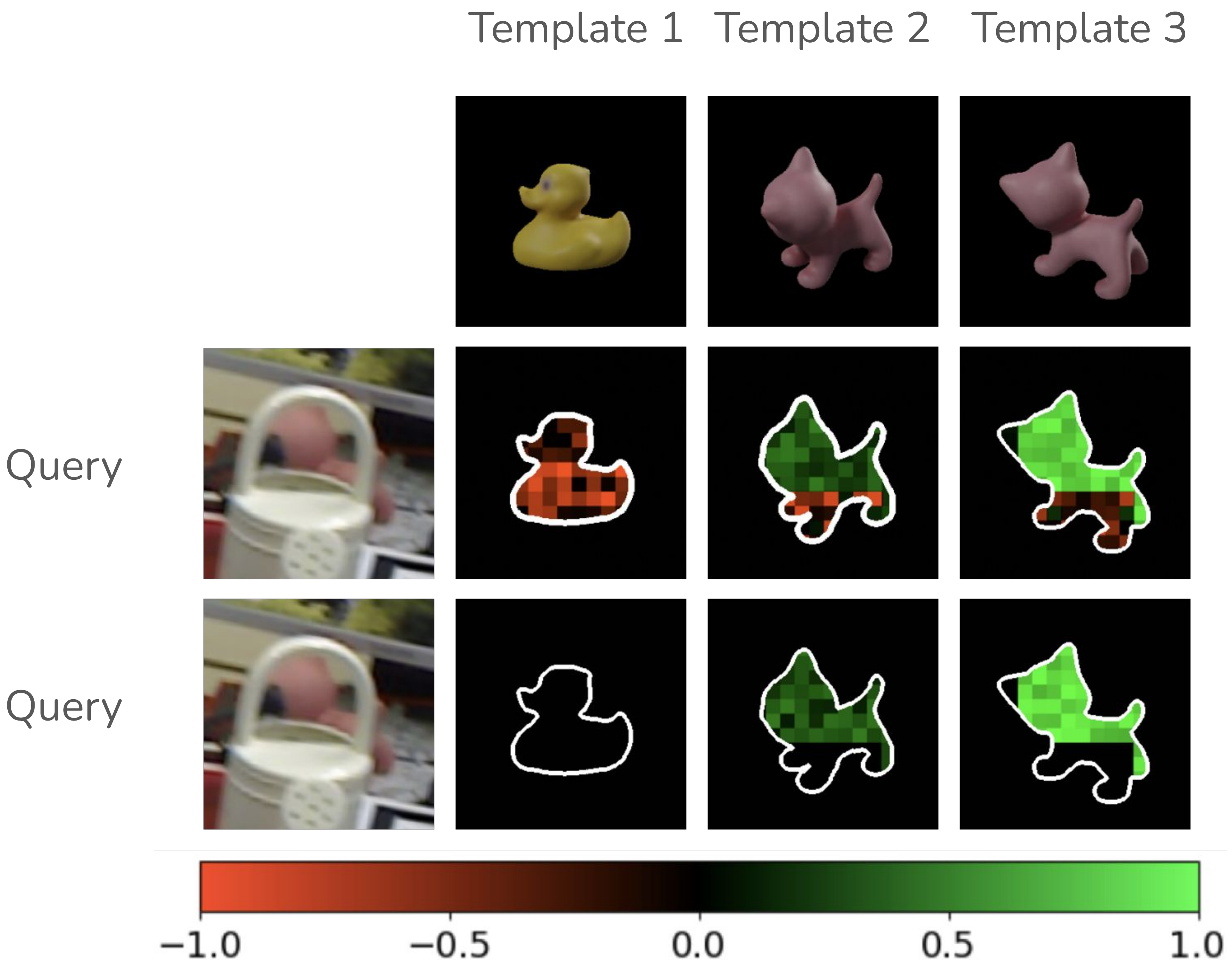


Training: contrastive loss InfoNCE



Representation learning with contrastive predictive coding, Äaron van den Oord et al., arXiV 2018

Robustness to occlusions

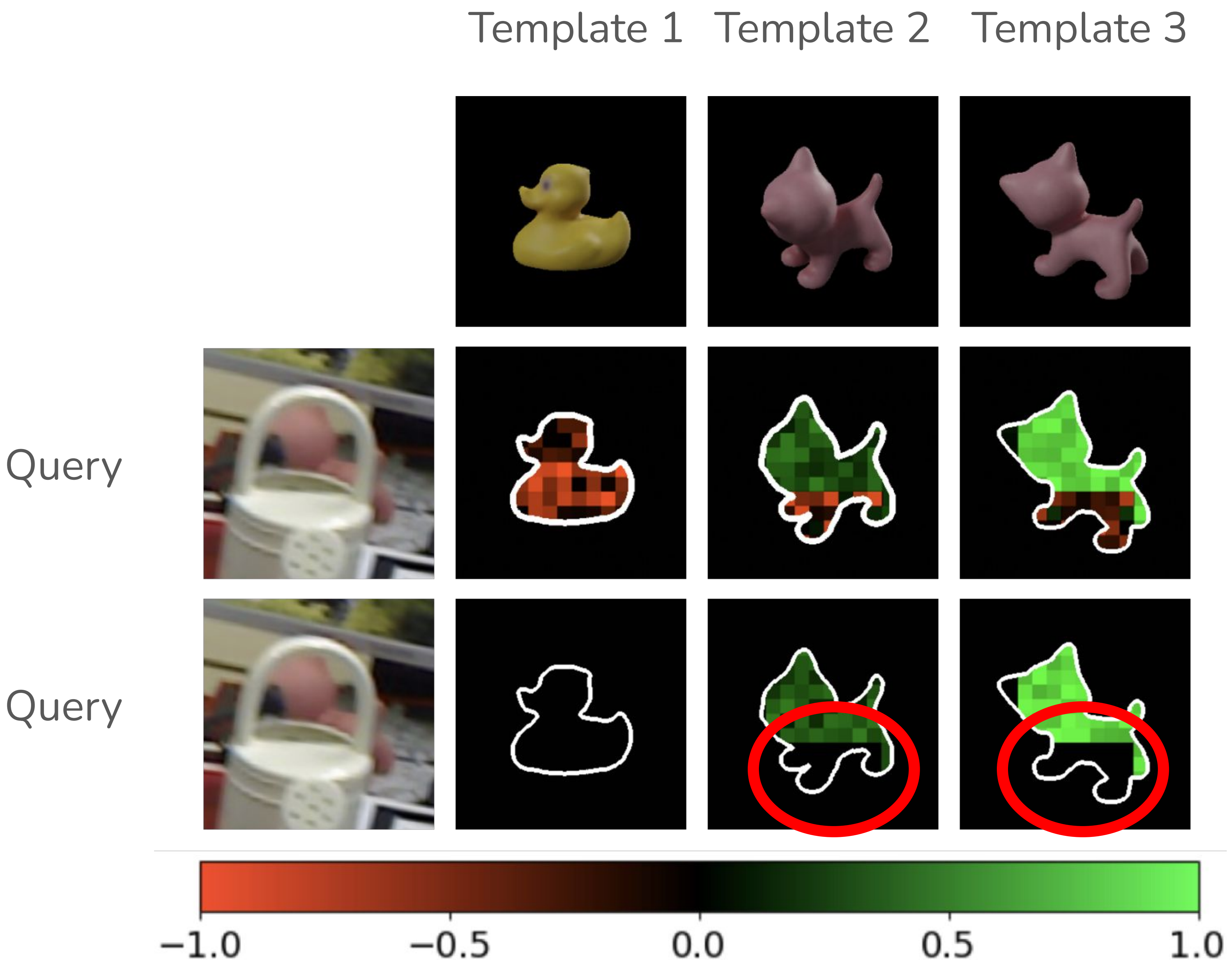


$$\mathrm{sim}(\overline{\mathbf{q}}, \overline{\mathbf{t}}) = \frac{1}{|\mathcal{M}|} \sum_{l} \mathcal{M}^{(l)} \mathcal{S}\left(\overline{\mathbf{q}}^{(l)}, \overline{\mathbf{t}}^{(l)}\right)$$

$$\mathrm{sim}^{*}(\overline{\mathbf{q}}, \overline{\mathbf{t}}) = \frac{1}{|\mathcal{M}|} \sum_{l} \mathcal{M}^{(l)} \mathcal{O}^{(l)} \mathcal{S}\left(\overline{\mathbf{q}}^{(l)}, \overline{\mathbf{t}}^{(l)}\right)$$

where $\mathcal{O}^{(l)} = 1_{\mathcal{S}(\overline{\mathbf{q}}^{(l)}, \overline{\mathbf{t}}^{(l)}) > \delta}$

Robustness to occlusions

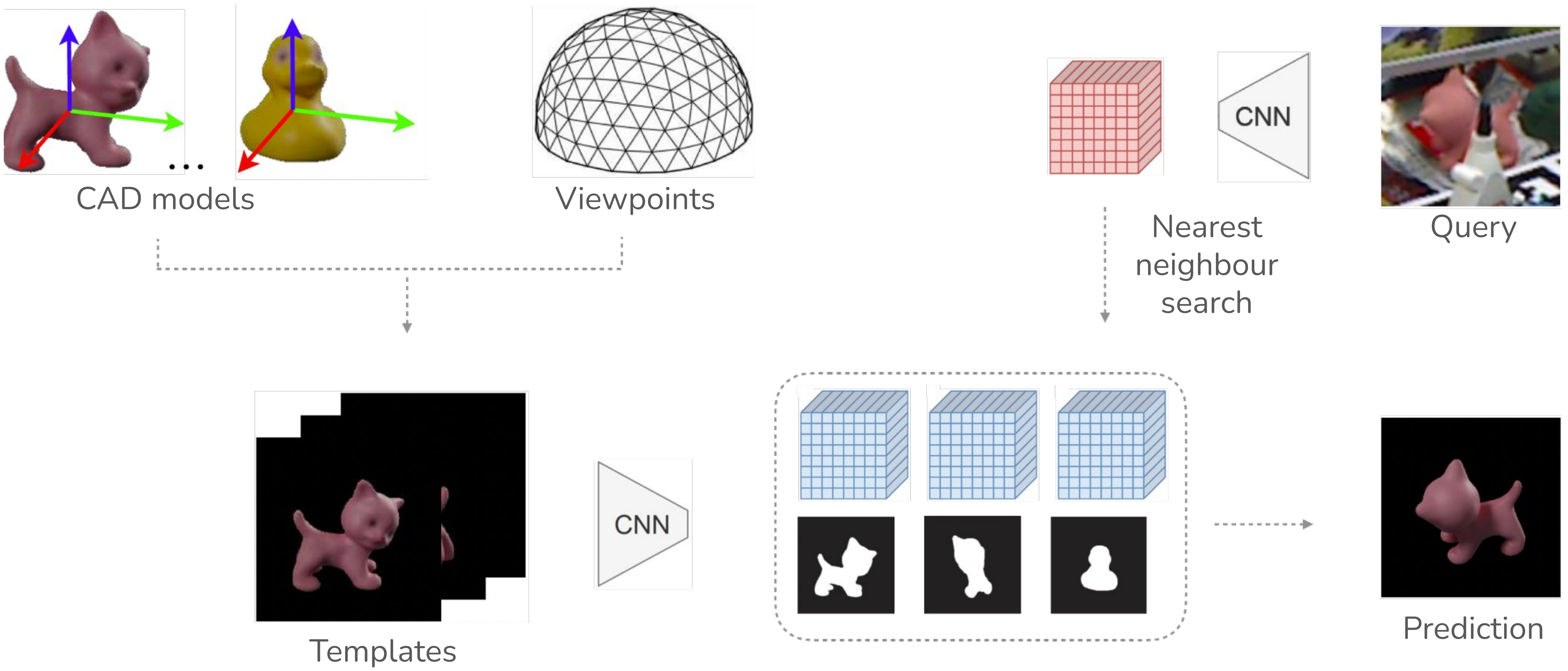


$$\mathrm{sim}(\overline{\mathbf{q}}, \overline{\mathbf{t}}) = \frac{1}{|\mathcal{M}|} \sum_{l} \mathcal{M}^{(l)} \mathcal{S}\left(\overline{\mathbf{q}}^{(l)}, \overline{\mathbf{t}}^{(l)}\right)$$

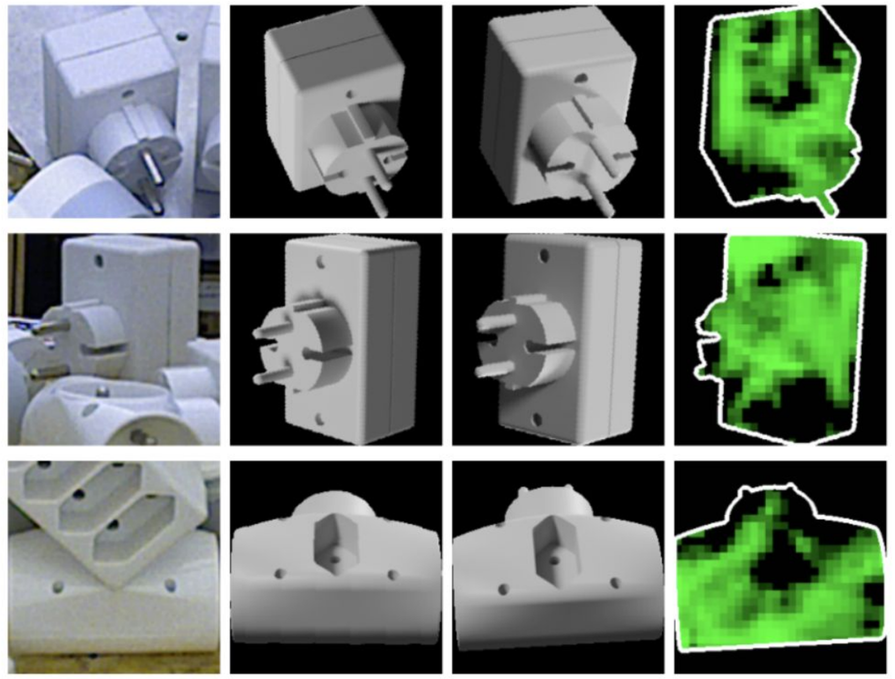
$$\mathrm{sim}^{*}(\overline{\mathbf{q}}, \overline{\mathbf{t}}) = \frac{1}{|\mathcal{M}|} \sum_{l} \mathcal{M}^{(l)} \mathcal{O}^{(l)} \mathcal{S}\left(\overline{\mathbf{q}}^{(l)}, \overline{\mathbf{t}}^{(l)}\right)$$

where $\mathcal{O}^{(l)} = 1_{\mathcal{S}(\overline{\mathbf{q}}^{(l)}, \overline{\mathbf{t}}^{(l)}) > \delta}$

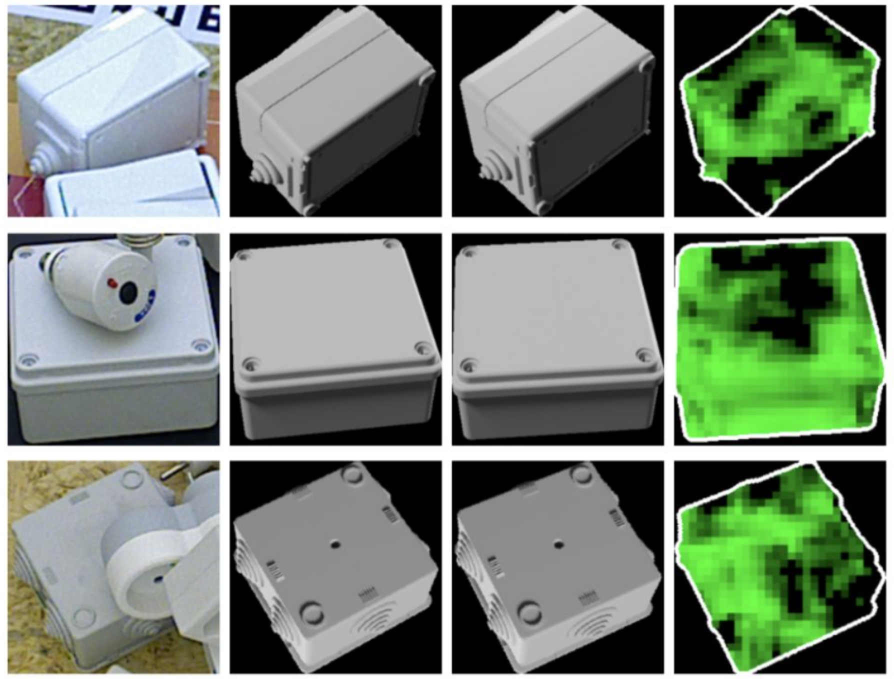
Inference on novel objects



Qualitative results

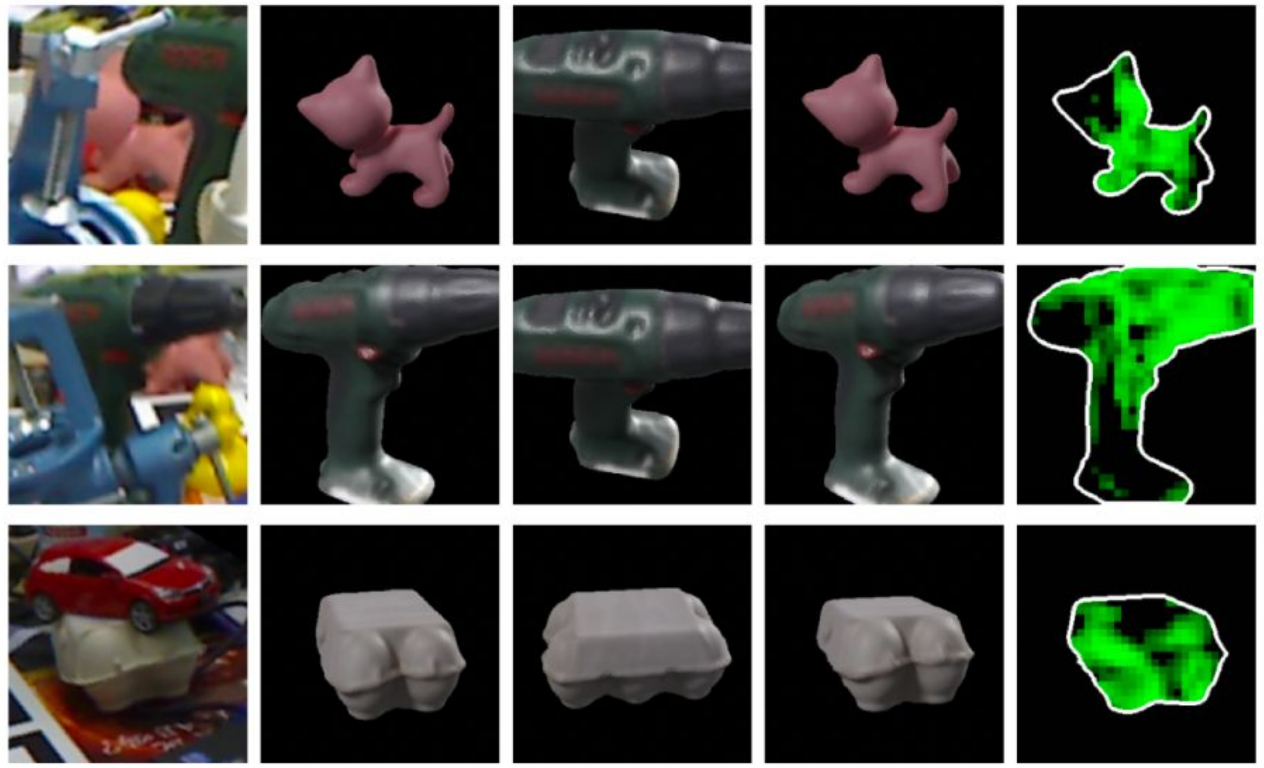


Query GT Ours Similarity



Query GT Ours Similarity

Query GT [1] Ours Similarity



Query GT [1] Ours Similarity

[1] Pose Guided RGB D Feature Learning for 3D Object Pose Estimation, Balntas et al., ICCV 2017

Evaluation protocol

- LINEMOD and LINEMOD-Occlusion:
  - Cross-validation style (i.e #1 for testing, #2 #3 for training)
  - Metric: Accuracy (object ID, angle difference <=15 degrees)

LINEMOD

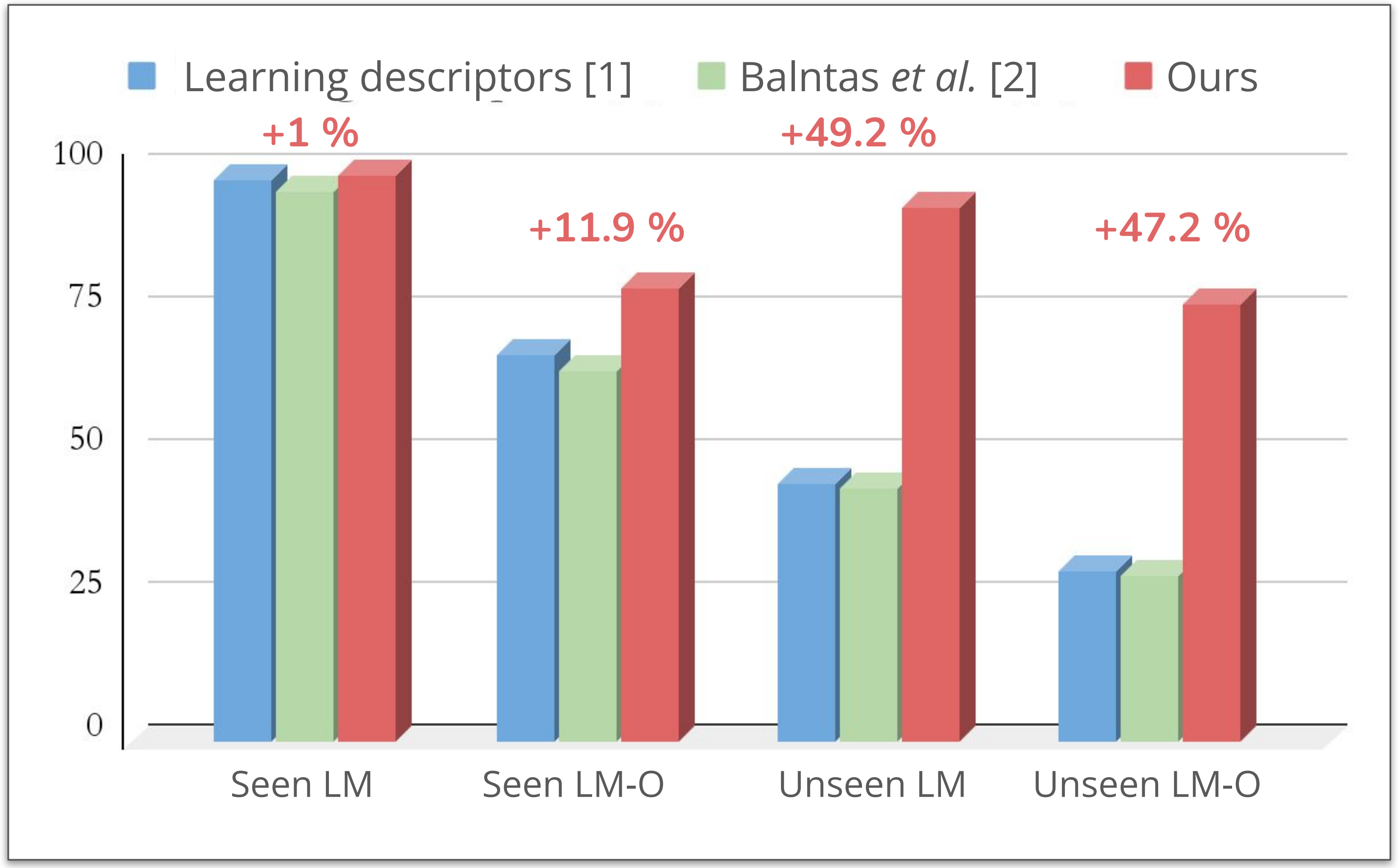
- T-LESS:
  - Training on objects 1-18, testing on objects 19-30
  - Metric: Average Recall with VSD

T-LESS

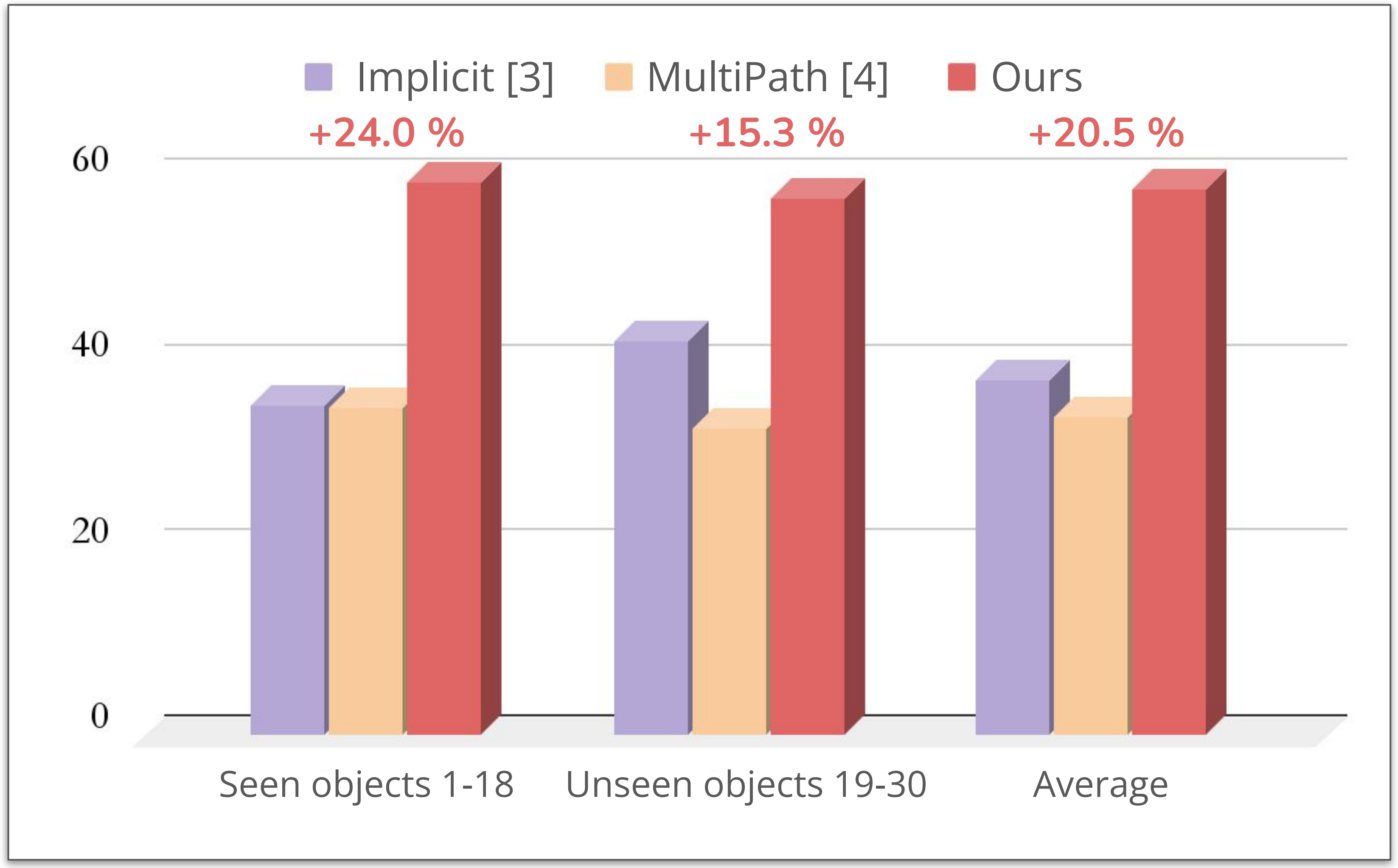
Learning Descriptors for Object Recognition and 3D Pose Estimation, Wolhart and Lepetit, CVPR 2015
MultiPath Learning for Object Pose Estimation Across Domains, Sundermeyer *et al.*, CVPR 2020

Results

- Our method significantly outperforms SOTA on both T-LESS, and LINEMOD(-Occlusion).



Quantitative results on LM and LM-O

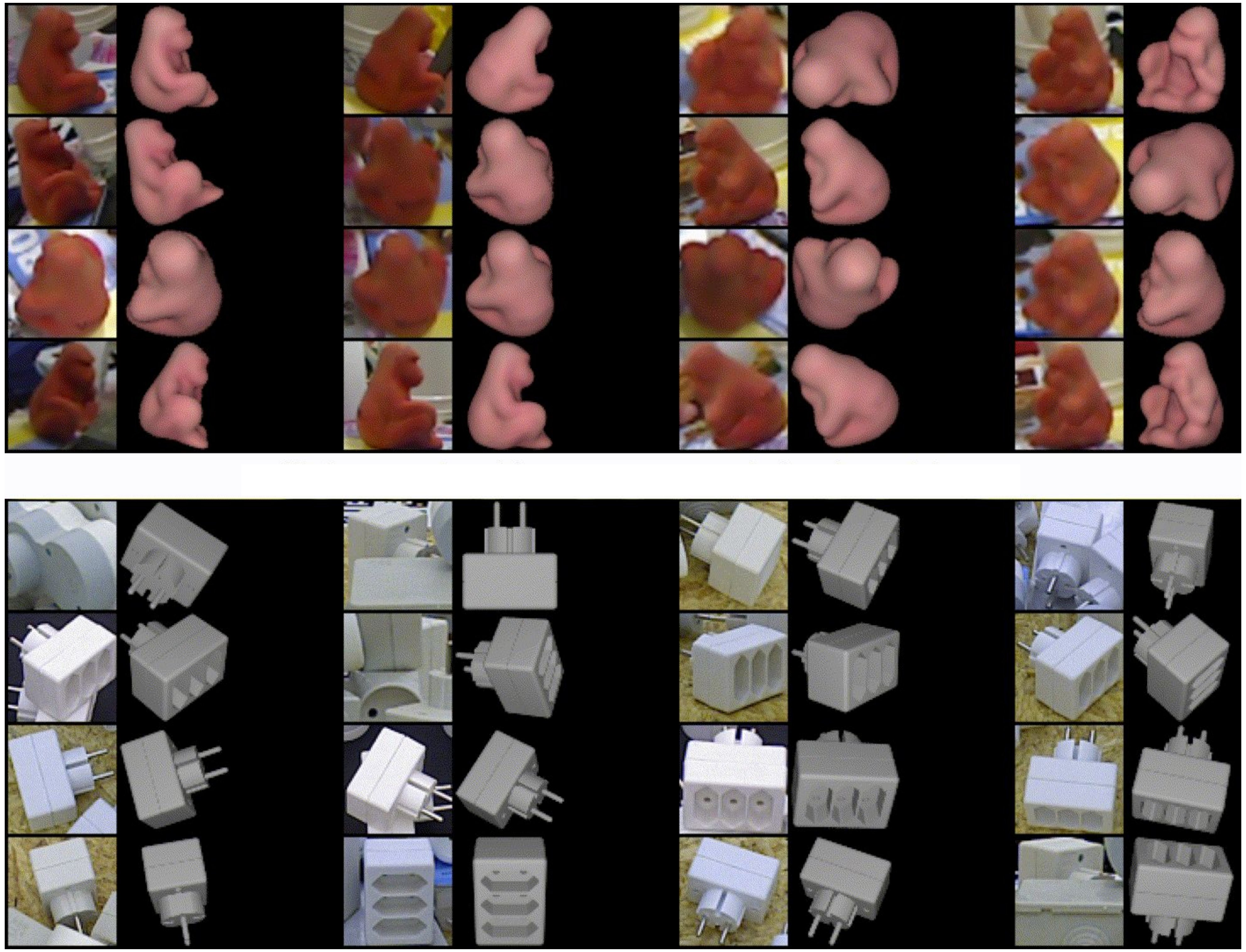


Quantitative results on T-LESS

[1] Learning Descriptors for Object Recognition and 3D Pose Estimation, Wohlhart and Lepetit, CVPR 2015
[2] Pose Guided RGB D Feature Learning for 3D Object Pose Estimation, Balntas et al., ICCV 2017
[3] Implicit 3D Orientation Learning for 6D Object Detection from RGB Images, Sundermeyer et al., ECCV 2018
[4] Multi-path Learning for Object Pose Estimation Across Domains, Sundermeyer et al., CVPR 2019

Summary

- **Failure cases** of global representation: cluttered background, pose discrimination;
- **Generalization**: ours is the first object pose method showing generalization on LINEMOD (same time as OSOP [1]);
- **Efficiency**: our method achieves 93.5% accuracy on unseen objects by training only on 7-8 reference objects.



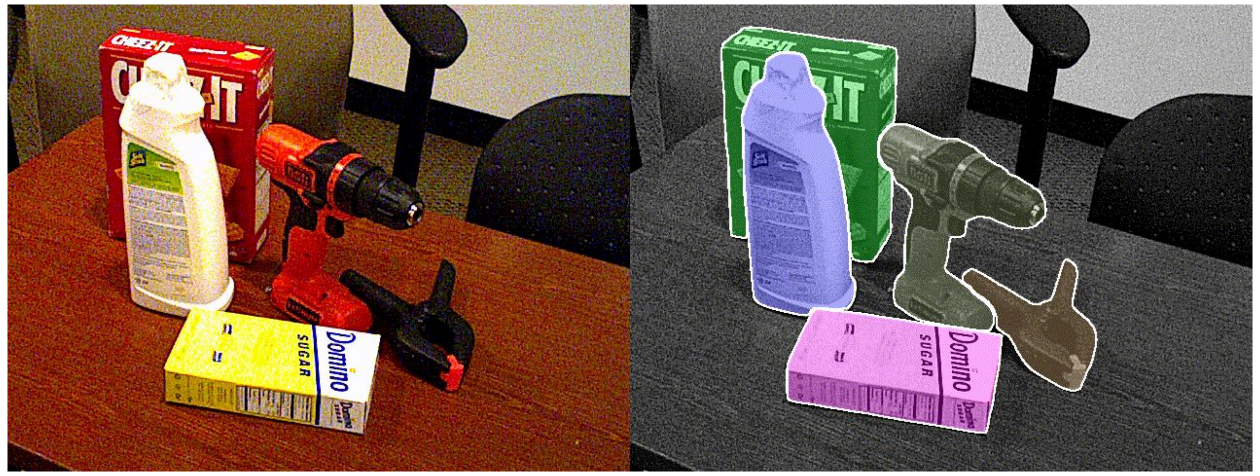
Quantitative results on novel objects

OSOP: A Multi-Stage One Shot Object Pose Estimation Framework, Shugurov *et al.*, CVPR 2022

Outline

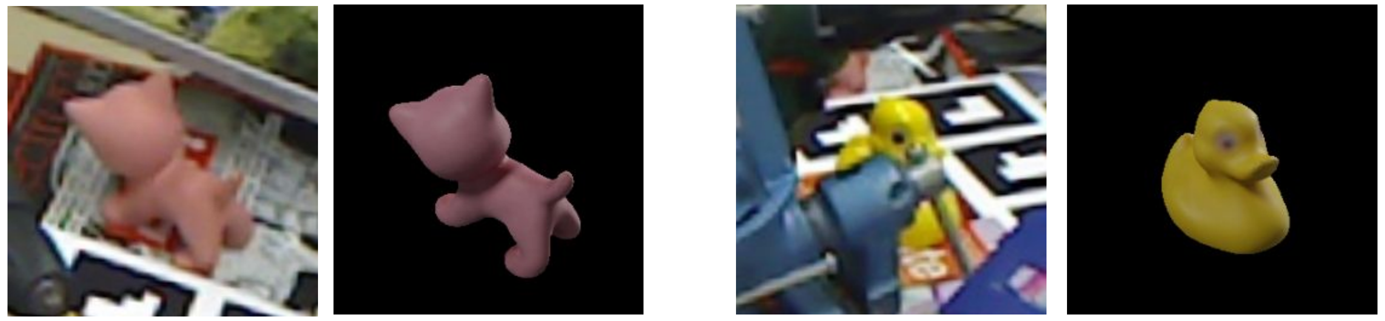
BOP 2023
4 awards

[ICCVW'23] CNOS: A Strong Baseline for CAD based Novel Object Segmentation



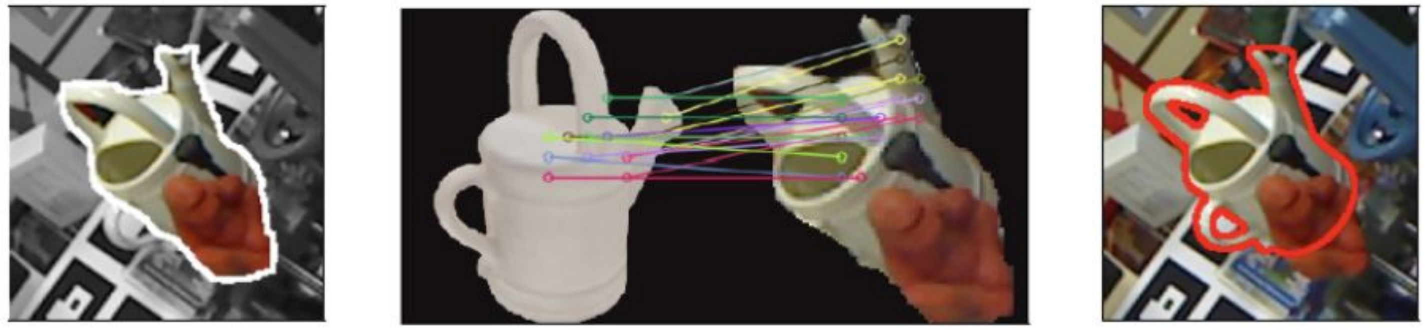
Input RGB Prediction

[CVPR'22] Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects and Robustness to Occlusions



Query Prediction Query Prediction

[CVPR'24] GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence



Query Predicted 2D-2D correspondences Prediction

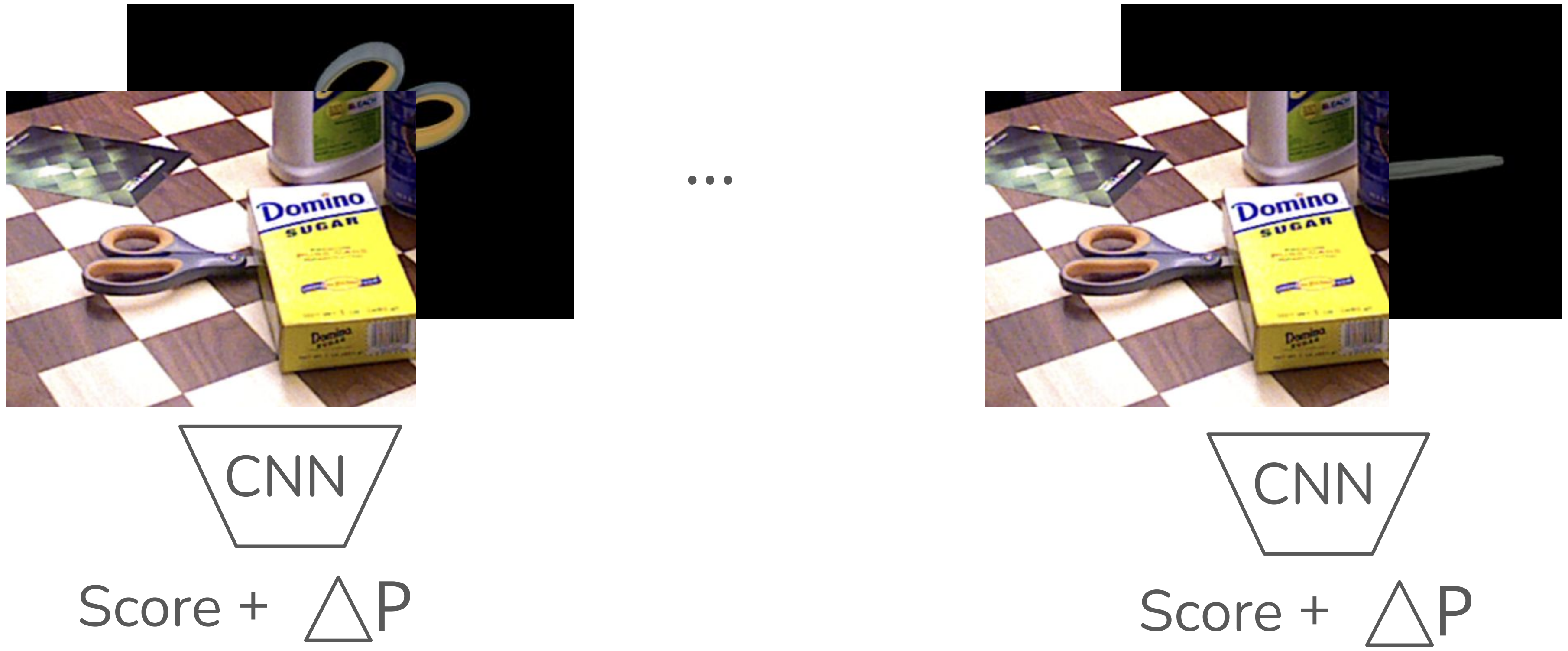
[CVPR'24] NOPE: Novel Object Pose Estimation from a Single Image

Reference Query Prediction Predicted pose distribution

Motivation

- CNOS detection/segmentation is noisy

- SOTA render-and-compare method MegaPose is slow

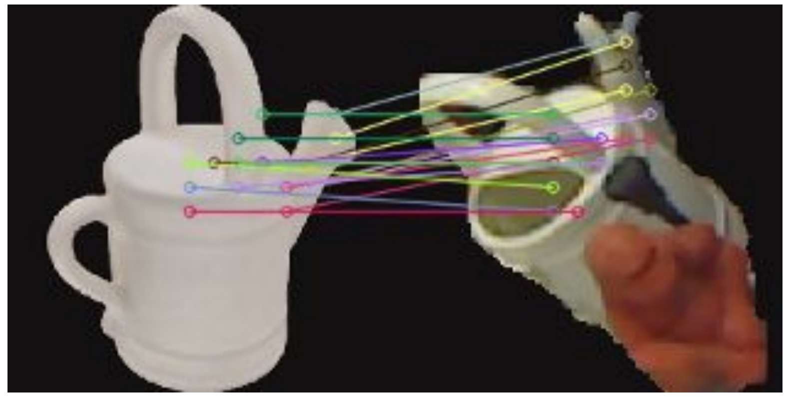


MegaPose 6D Pose Estimation of Novel Objects via Render & Compare, Labbé et al., CoRL 2022

Our approach: 6DoF from one 2D-to-2D correspondence

Segmentation

Nearest template

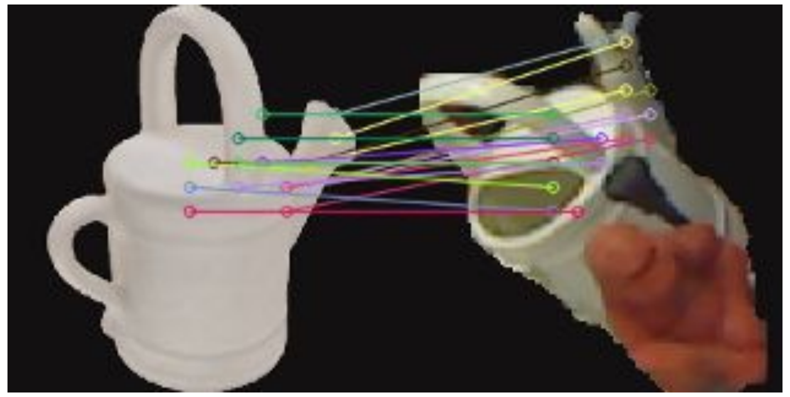


2D-to-2D correspondences

Our approach: 6DoF from one 2D-to-2D correspondence

Segmentation

Nearest template



2D-to-2D correspondences

Alignment

Prediction

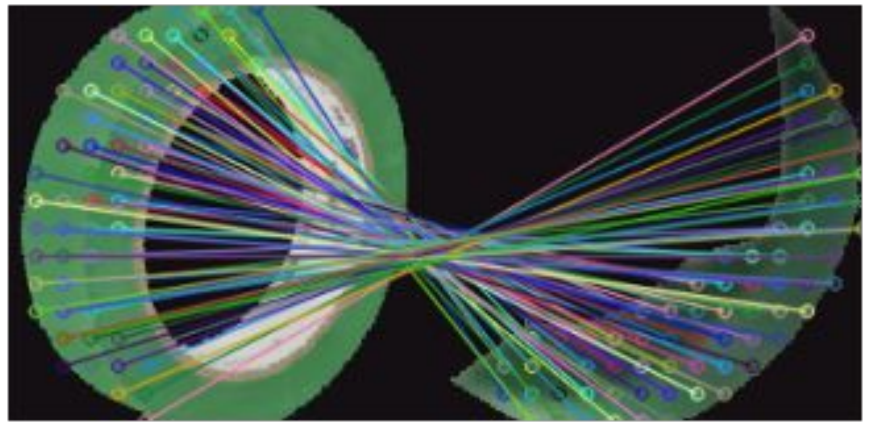
Out-of-plane rotation
2 DoF

Translation 2D
2 DoF

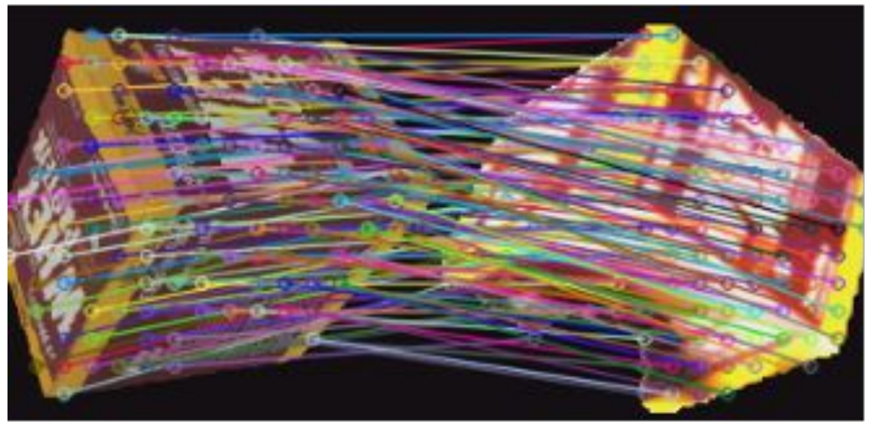
We predict 2 missing DoFs (in-plane, scaling) for each 2D-to-2D correspondence.

-> Reducing the number of templates required at inference.

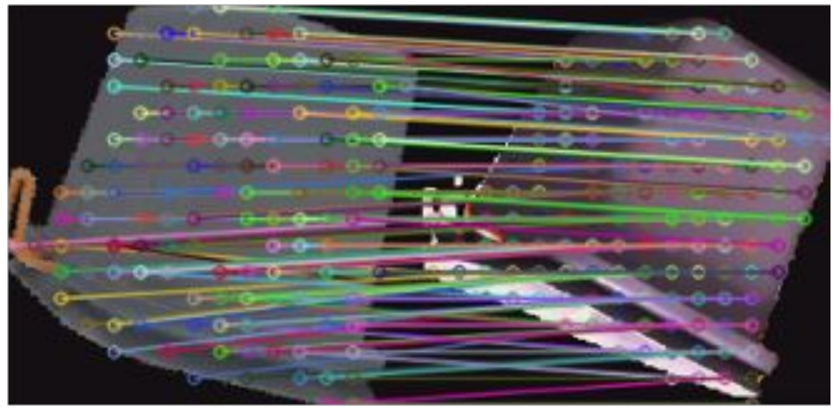
Training samples: 2D-to-2D correspondences



Template Query

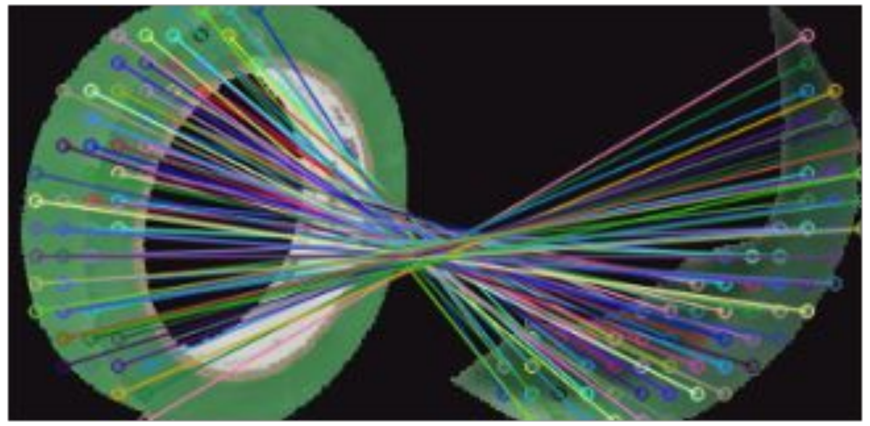


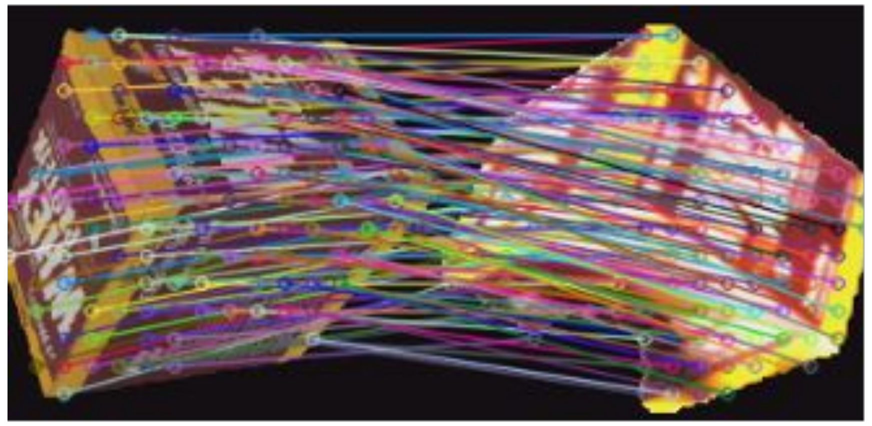
Template Query

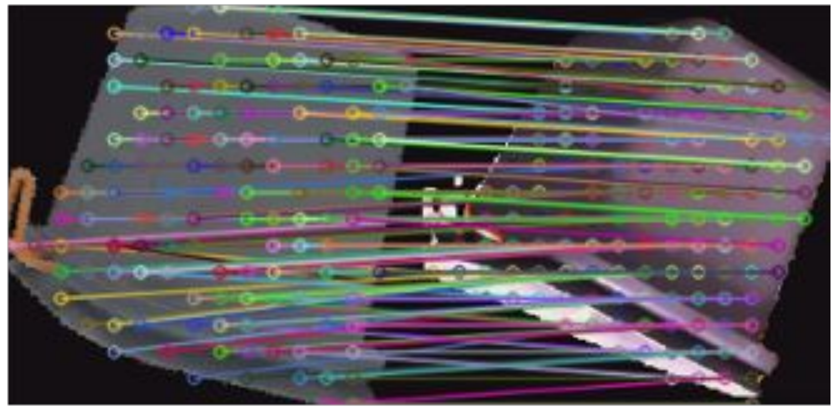


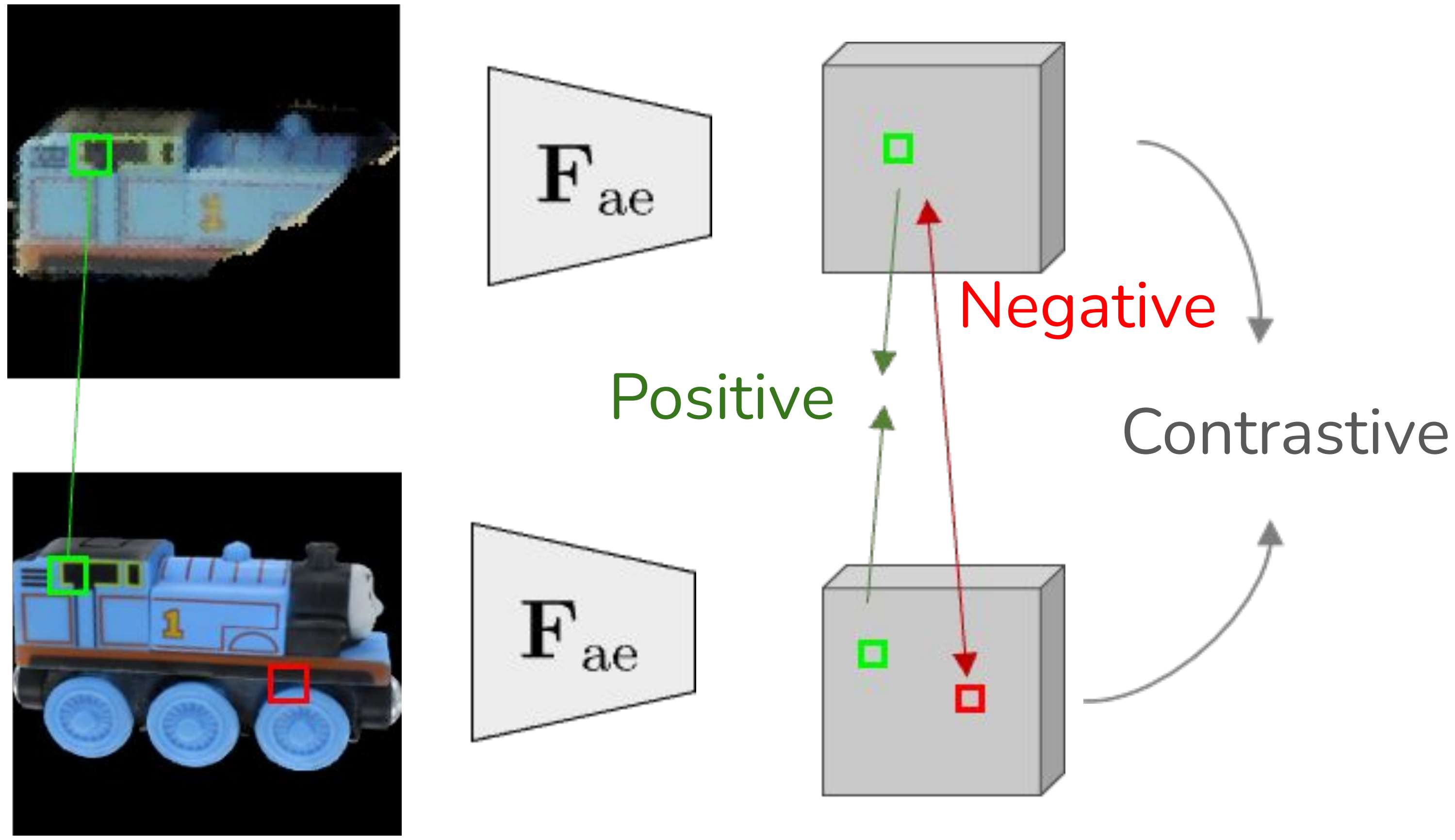
Template Query

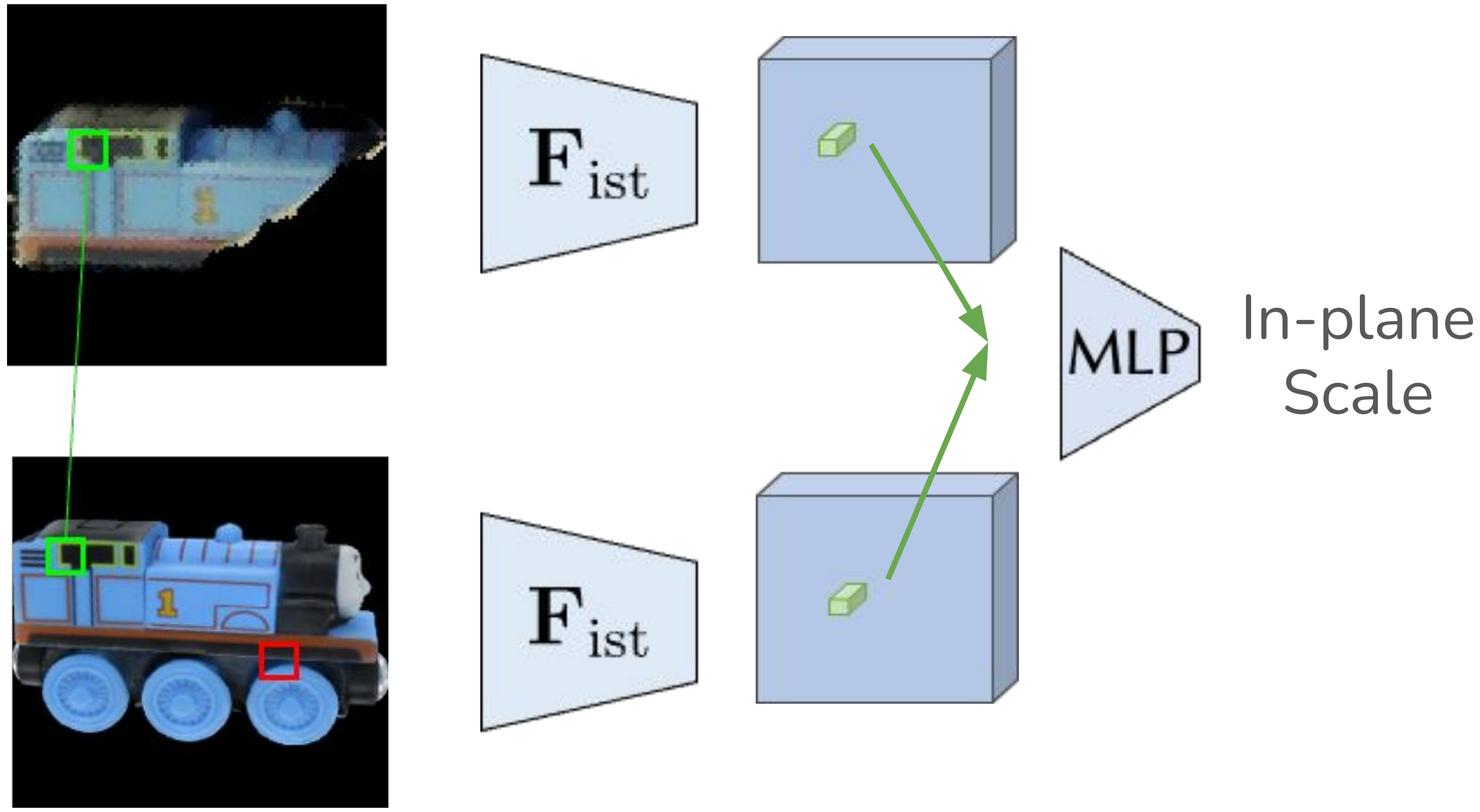
Training losses: contrastive InfoNCE & regression



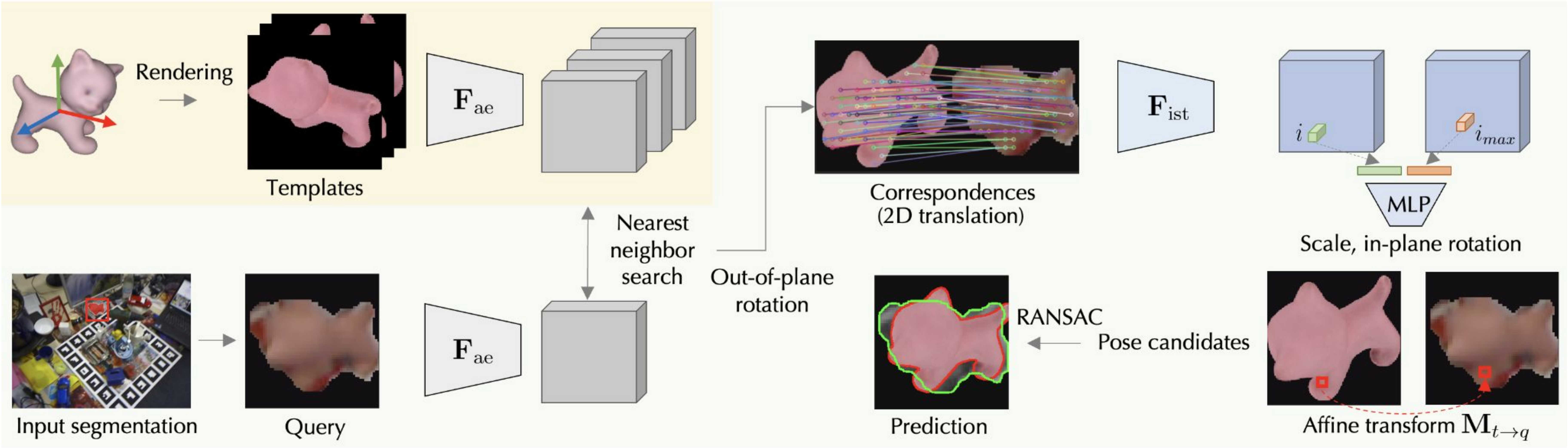






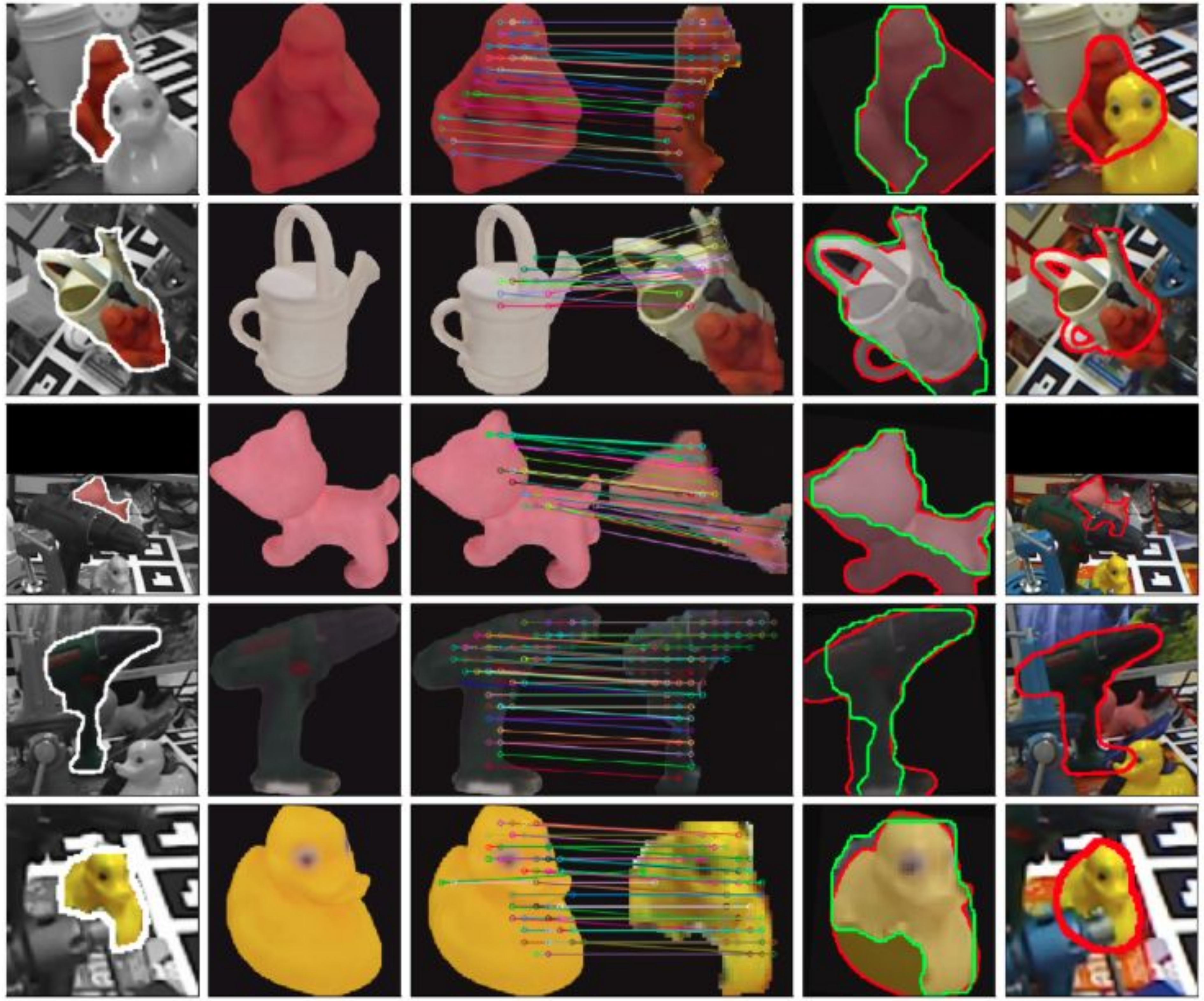


Inference



- Onboarding: rendering 162 templates from CAD model & extracting local features
- Processing: cropping input image and extract local features
- Retrieval: nearest templates and a set of 2D-to-2D correspondences
- Pose fitting: predicting 2D scale, in-plane rotation for each correspondence & RANSAC

Qualitative results



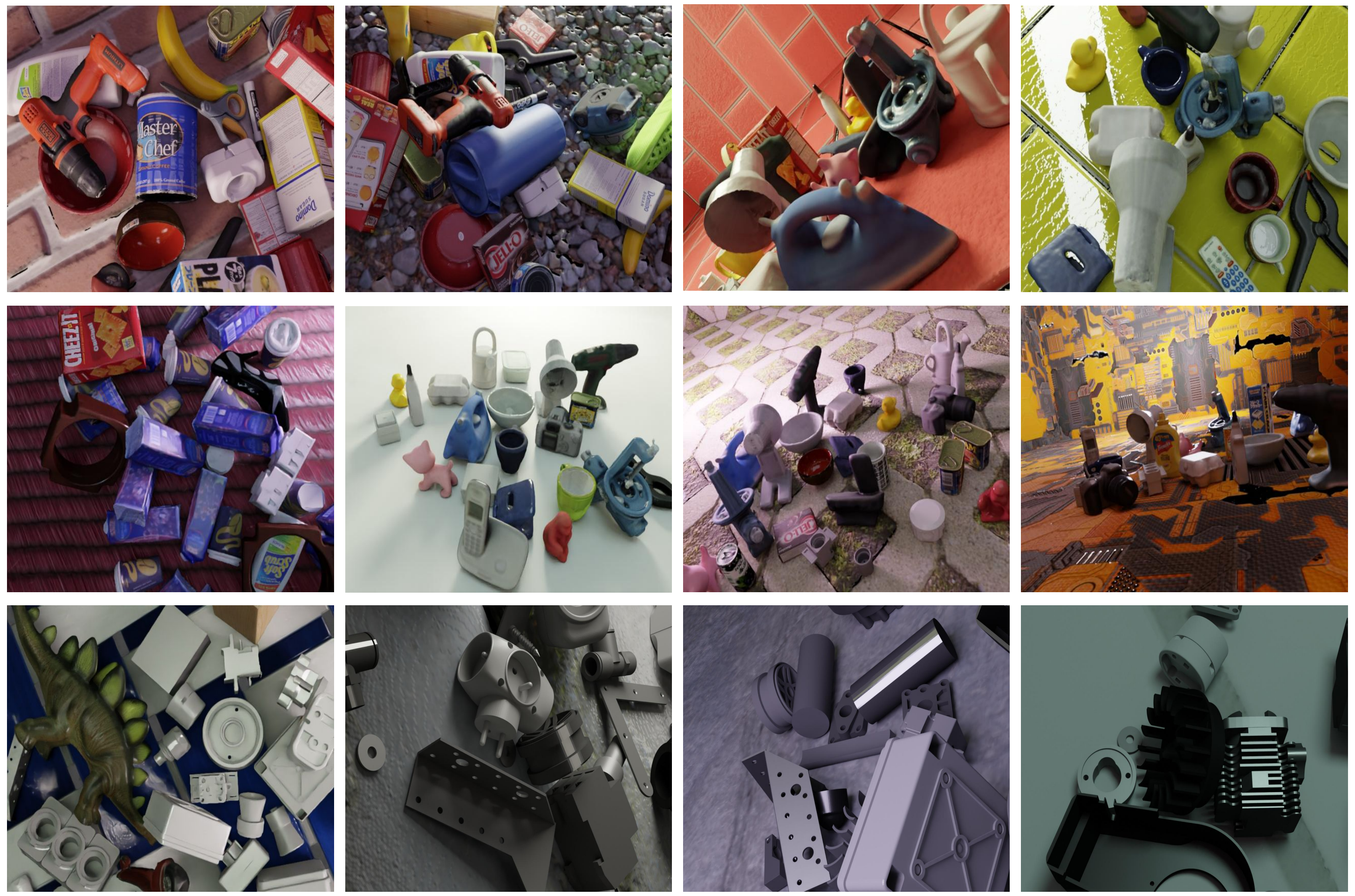
Segmentation | Nearest template | 2D-to-2D correspondences | Alignment | Prediction

Qualitative results

Segmentation | Nearest template | 2D-to-2D correspondences | Alignment | Prediction

Evaluation protocol

- Training datasets: 2M BlenderProc images of MegaPose-GSO, MegaPose-ShapeNet
- Testing datasets: 7 datasets of BOP-Classic-Core
- Evaluation metrics: BOP evaluation protocol (MSSD, MSPD, VSD)
- Input detections: CNOS

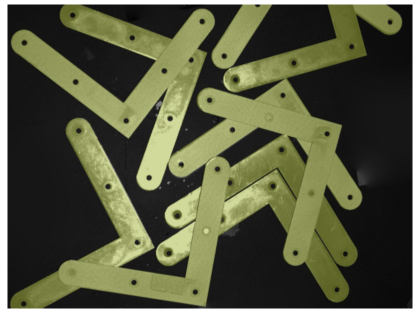


Synthetic images generated by BlenderProc
(MegaPose-GSO & MegaPose-ShapeNet)

LM-O

T-LESS

HB



ITODD

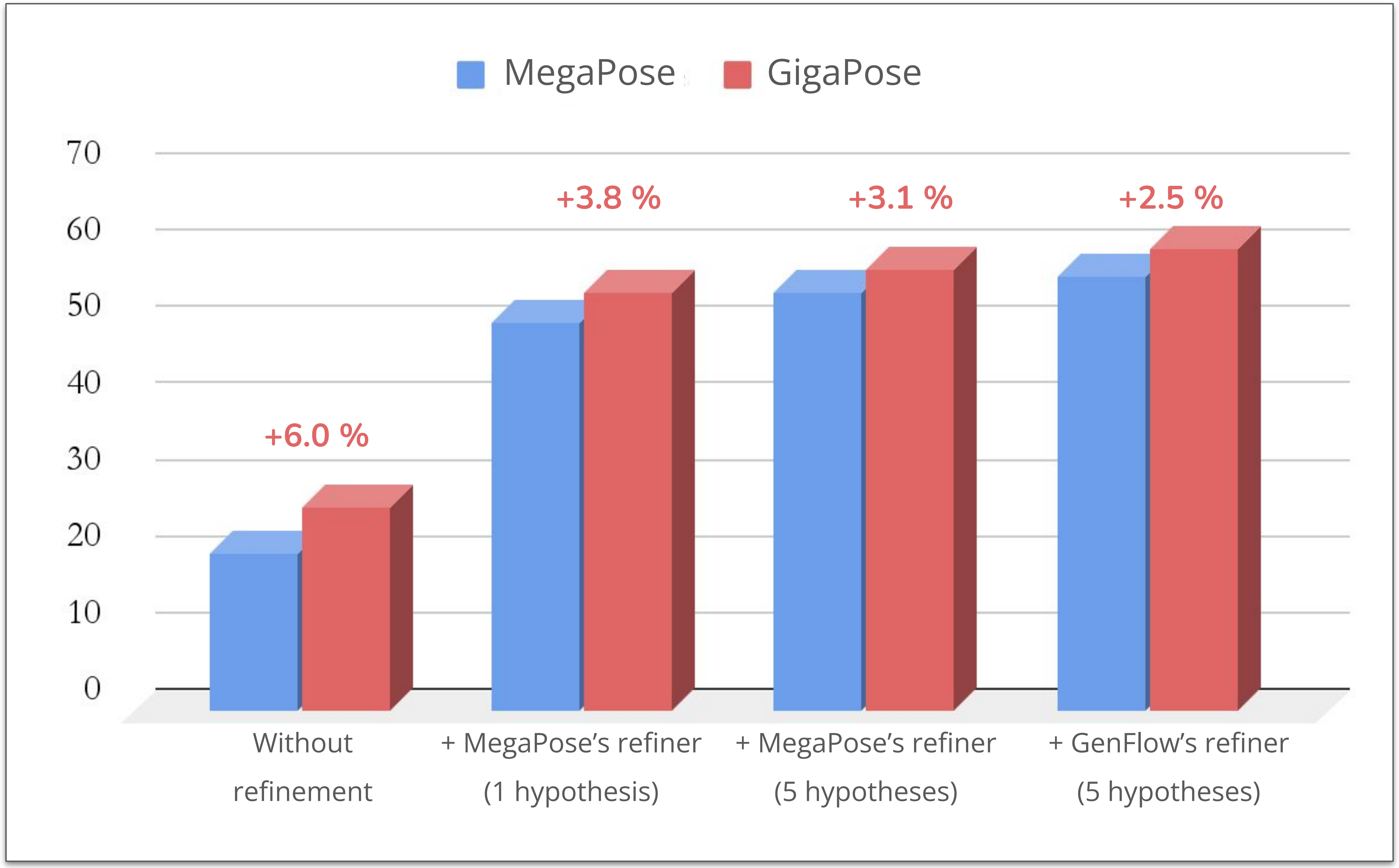
YCB-V

IC-BIN

TUD-L

Results

- GigaPose outperforms MegaPose in all settings while being **35x faster** for coarse pose stage.

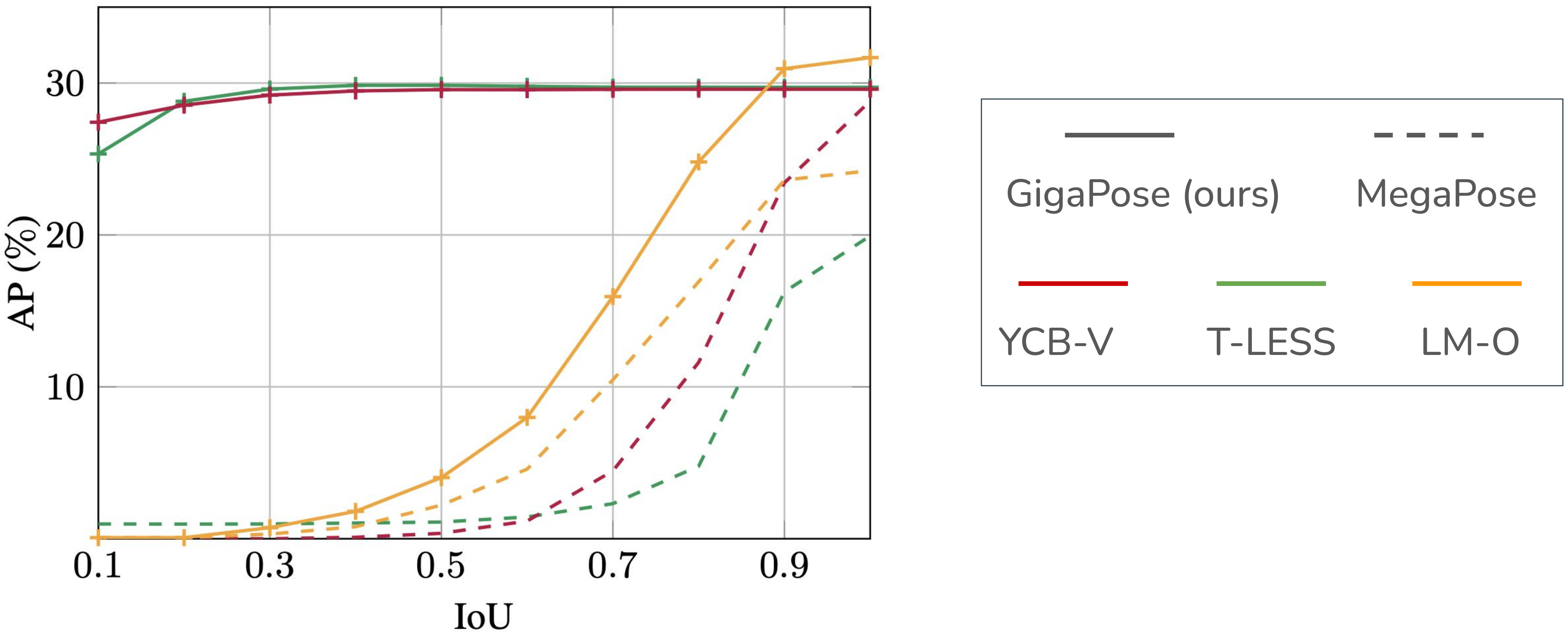


Quantitative results on seven core datasets of BOP challenge

[1] MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare, Labbé et al., CoRL 2022
[3] GenFlow: Generalizable Recurrent Flow for 6D Pose Refinement of Novel Objects, Moon et al., arXiv 2024

Results

- Evaluating robustness to segmentation errors:
  - X axis: IoU between the input masks vs GT masks
  - Y axis: performance at different IoU thresholds



[1] ZS6D: Zero-Shot 6D Object Pose Estimation Using Vision Transformers, arXiv, 2023
[2] MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare, Labbé et al., CoRL 2022
[3] GenFlow: Generalizable Recurrent Flow for 6D Pose Refinement of Novel Objects, Moon et al., arXiv 2024

Results using a single reference image

Reconstruction from a single image by Wonder3D [1]

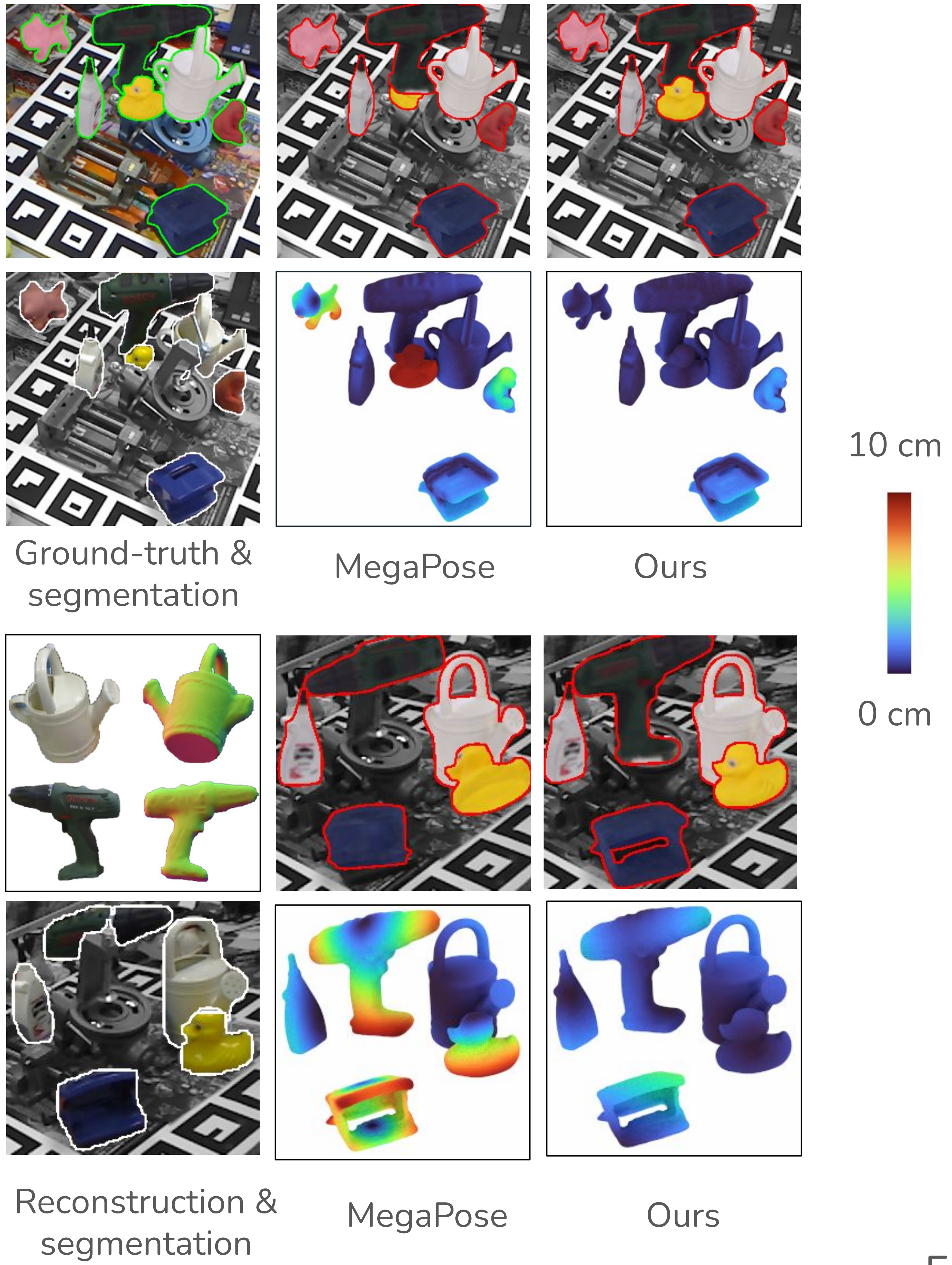
	Method	Detection	Single image Coarse	Single image Refined	GT 3D model w/o refinement
1	MegaPose	GT 3D model	16.3	25.6	22.9
2	GigaPose (ours)	GT 3D model	**19.5**	**29.1**	**29.9**
3	MegaPose	Single image	15.4	25.2	22.7
4	GigaPose (ours)	Single image	**18.5**	**28.2**	**29.8**

Table 2. Results with predicted 3D models on LM-O dataset

[1] Wonder3D: Single image to 3d using cross-domain diffusion, Long et al., ICLR 2024

Summary

- **Accuracy & run-time:** 2.5% AR improvement while 35x faster for coarse pose stage;
- **Robustness to segmentation errors:** GigaPose's performance is stable across input noisy masks;
- **Number of templates:** Requiring only 162 templates, depicting only out-of-plane rotation (3x less than MegaPose)
- **Pose estimation from a single image:** GigaPose + MegaPose + single image > MegaPose + GT model



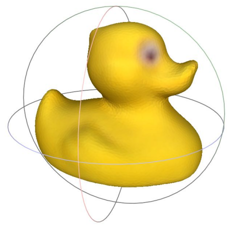
Conclusion

Training (hours/days) → **Onboarding** (sec/min) → **Inference** (online)

Task 1: Model-based 6D localization of novel objects: **Template-Pose, GigaPose, NOPE**

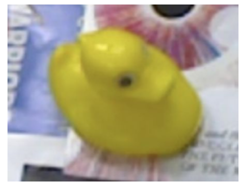
Input: No information about target objects

Input:



3D model

or

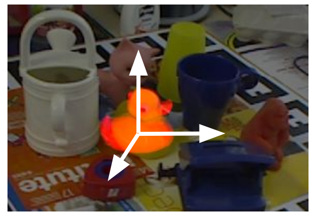


Reference image

Input:

RGB image

Output:

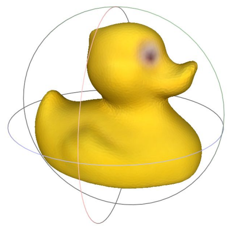


6D pose

Task 2: Model-based 2D detection of novel objects: **CNOS**

Input: No information about target objects

Input:



3D model

Input:

RGB image

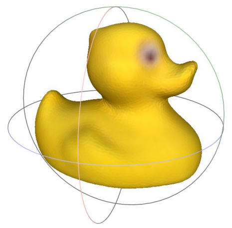
Output:

2D bounding box

Task 3: Model-based 2D segmentation of novel objects: **CNOS**

Input: No information about target objects

Input:



3D model

Input:

RGB image

Output:

2D segmentation

Open-source contributions on Github / HuggingFace

- All projects are open-source on Github:

★216	24	nv-nguyen/cnos
★187	11	nv-nguyen/nope
★179	13	nv-nguyen/template-pose
★153	14	nv-nguyen/gigaPose
★74	3	nv-nguyen/pizza
★35	3	nv-nguyen/bop_viz_kit

- BOP challenges:

>10K / month		huggingface.co/datasets/bop-benchmark/datasets
★417	141	thodan/bop_toolkit

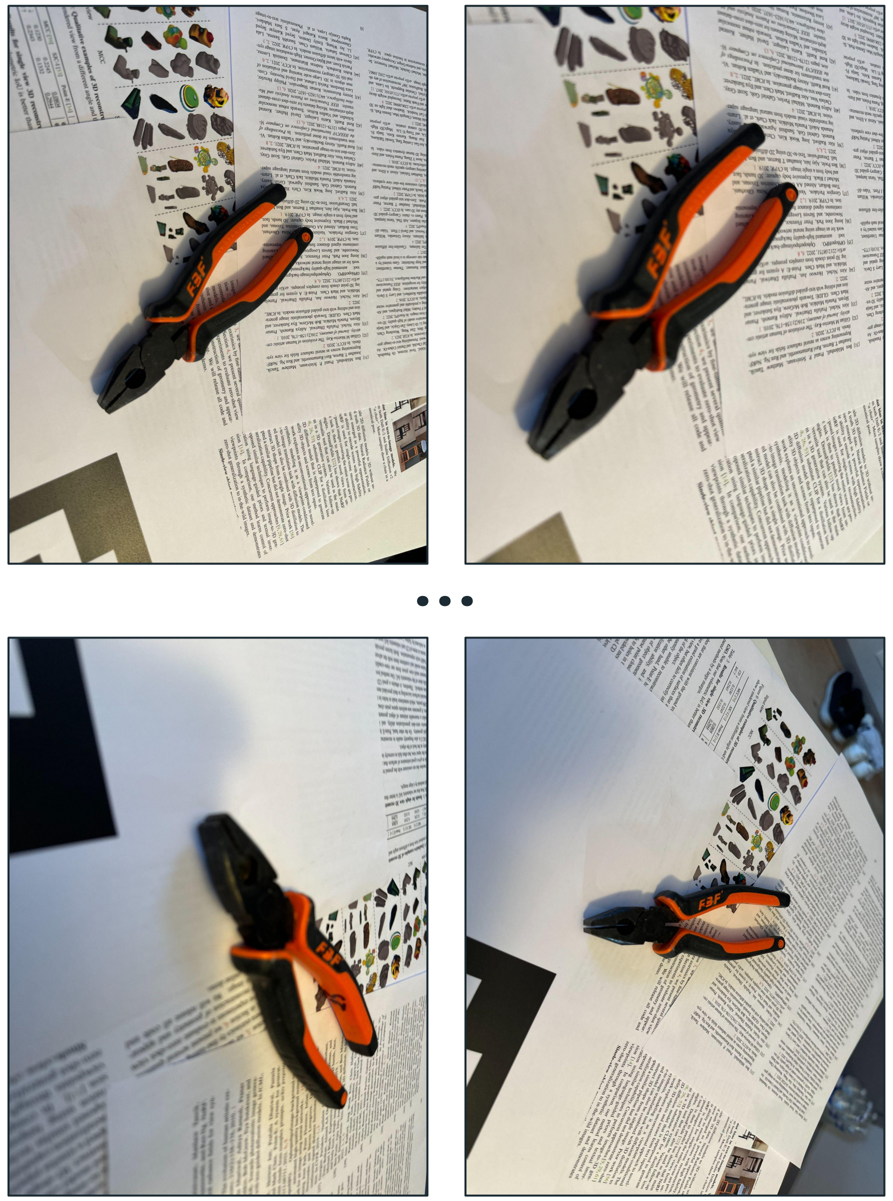
- Object pose paper summary:

★725	88	YoungXIAO13/ObjectPoseEstimationSummary

Number of stars, forks, downloads counted on 9th December 2024

Future work: model-free novel object pose estimation

- Replacing 3D model of test objects by a reference video (introduced in BOP challenge 2024)



Images captured by iPhone

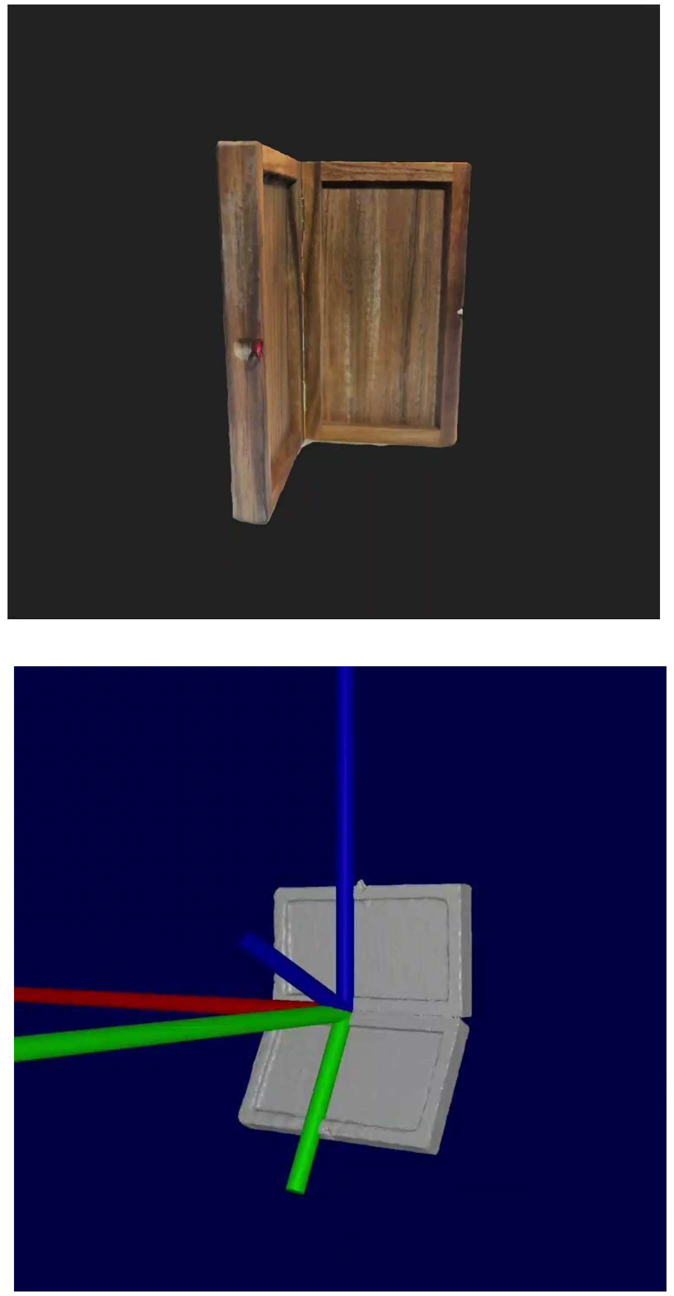
3D reconstruction by SuGaR [1]

[1] SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering, Guédon and Lepetit, CVPR 2024

Future work: articulated object pose estimation

- Extension to articulated object pose estimation

Examples of articulated objects



3D reconstruction + annotated joint

Thank you !!!

BOP

Benchmark Object Pose
Big Organized Party