CVPR
JUNE 17-21, 2024
SEATTLE, WA

HouseCat6D – A Large-Scale Multi-Modal Category Level 6D Object Perception Dataset with Household Objects in Realistic Scenarios

HyunJun Jung[1,*], Shun-Cheng Wu[1,*], Patrick Ruhkamp[1,*], Guangyao Zhai[1,2,†,*], Hannah Schieber[1,3,7,*], Giulia Rizzoli[4], Pengyuan Wang[1], Hongcheng Zhao[1], Lorenzo Garattoni[5], Sven Meier[5], Daniel Roth[1,7], Nassir Navab[1], Benjamin Busam[1,2,6]

hyunjun.jung@tum.de guangyao.zhai@tum.de shuncheng.wu@tum.de b.busam@tum.de

Poster ID12086 , Friday, 10:30 - 12:00

[1] Technical University of Munich, [2] Munich Center for Machine Learning, [3] FAU Erlangen-Nüremberg, [4] University of Padova, [5] Toyota Motor Europe, [6] 3dwe.ai, [7] Klinikum rechs der Isar, * Equal Contributions, † Corresponding Author

Munich Center for Machine Learning
Friedrich-Alexander-Universität Erlangen-Nürnberg
3Dwe
Klinikum rechts der Isar
Technische Universität München

HouseCat6D Dataset : Motivation

Category Level 6D Pose Estimation

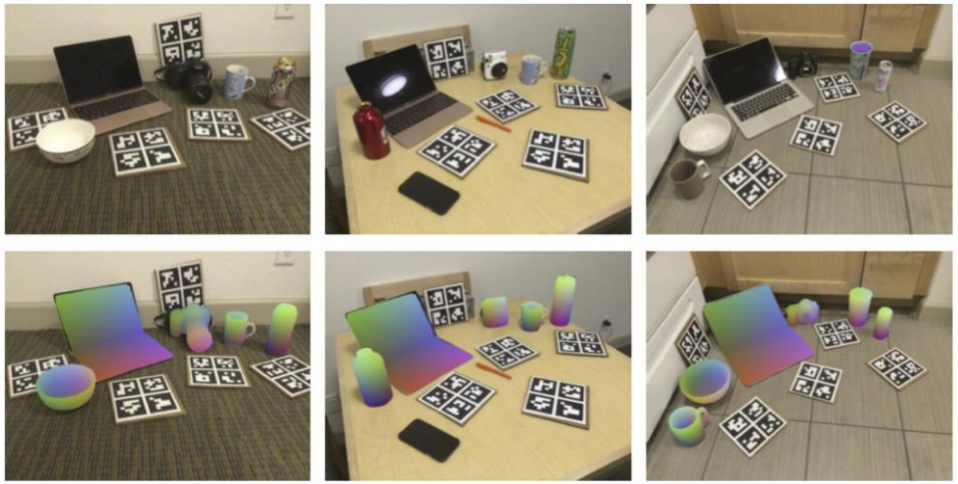
Train Set :

Test Set :

Test on **unknown instances from known categories**

Only Existing Real-World Category Level 6D Pose Dataset : NOCS dataset

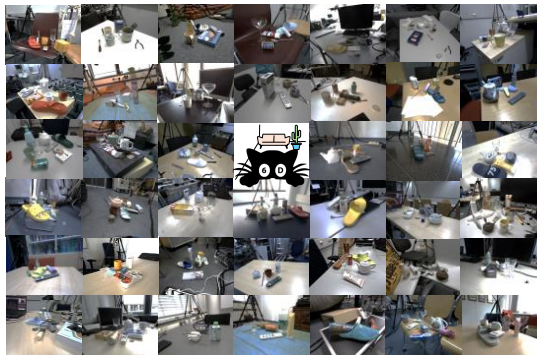
x42 Objects



x18 Scenes

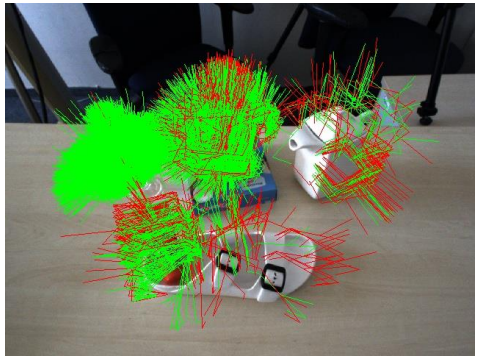
Normalized object co- ordinate space for category-level 6d object pose and size estimation. H.Wang, S.Sridhar, J.Huang, J.Valentin, S.Song, LJ.Guibas. (CVPR 2019)

HouseCat6D Dataset : Intro



x41

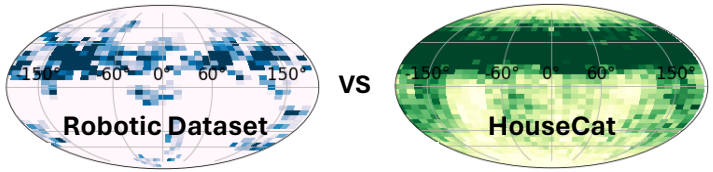
x160K



x10M

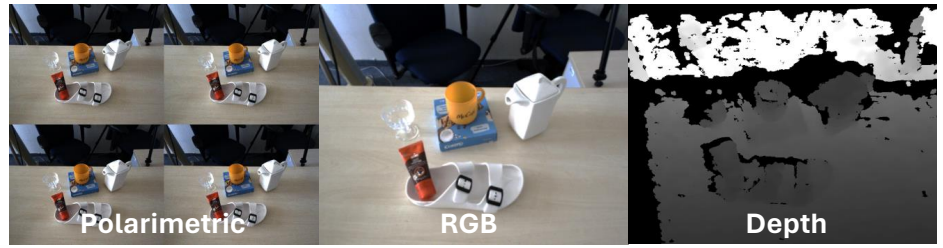
Reflective, Transparent

194 Objects from **10x** Household Categories



Spherical Density of Camera Poses

Extensive Viewpoint Coverage

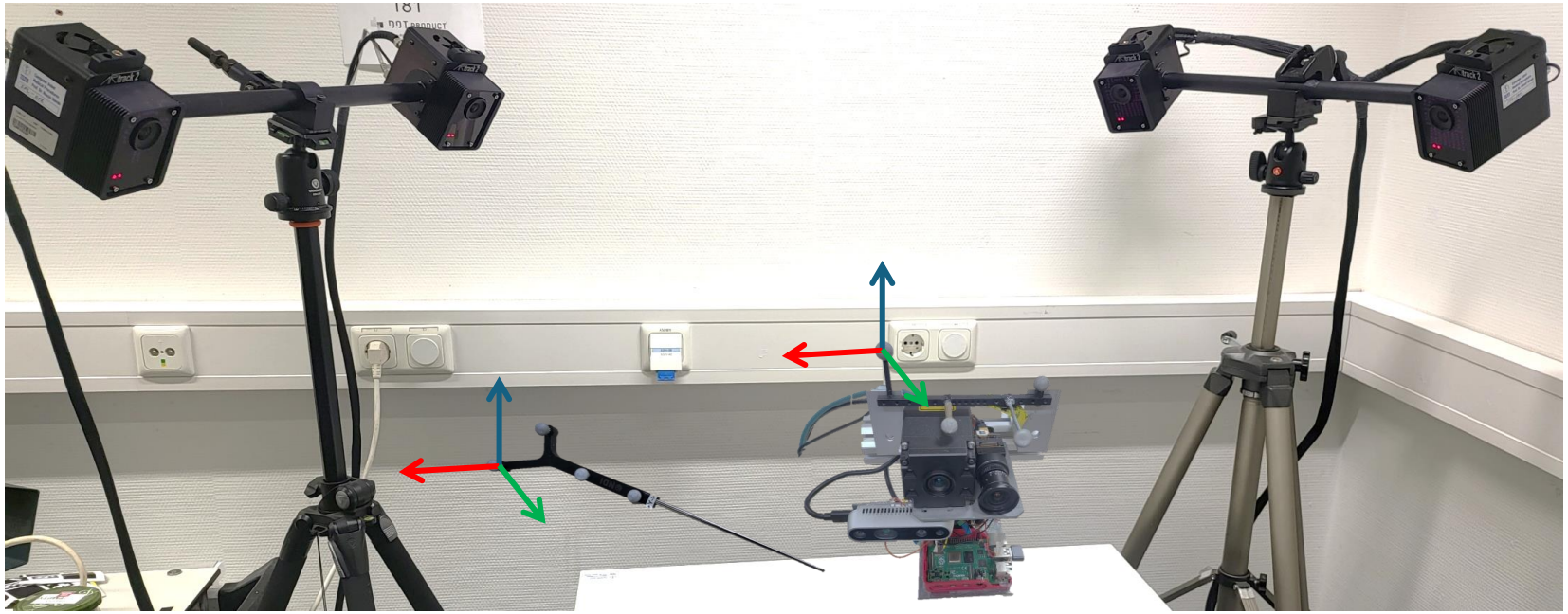


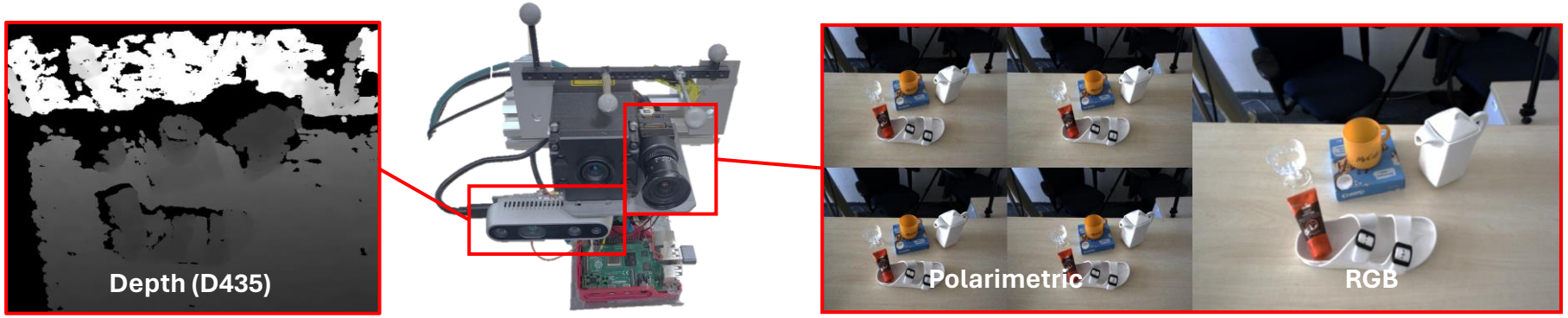
3 Modalities

Dataset	RGBD based	TOD [1]	StereOBJ [2]	PhoCal [3]	Ours
3D Labeling	Depth Map	Multi-View	Multi-View	Robot	IR tracker
Point RMSE	≥ 17 mm	3.4 mm	2.3 mm	0.80 mm	$1.35\text{ mm} \leq \epsilon \leq 1.73\text{ mm}$

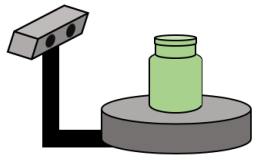
[1] **Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects**, X.Liu, R.Jonschkowski, A.Angelova, K.Konolige (CVPR 2020)
[2] **Stereobj-1m : Large-scale stereo image dataset for 6d object pose estimation**, X.Liu, S.Iwase, K.M.Kitani (ICCV 2021)
[3] **PhoCal : A Multi-Modal Dataset for Category-Level Object Pose Estimation with Photometrically Challenging Objects**, P.Wang, HJ.Jung, Y.Li, S.Shen, RP.Srikanth, L.Garattoni, S.Meier, N.Navab, B.Busam (CVPR 2022)

HouseCat6D Dataset : Hardware

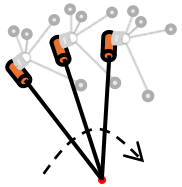




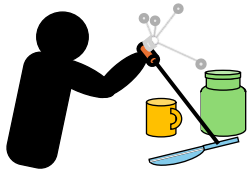
HouseCat6D Dataset : Pipeline



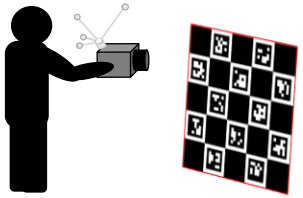
(a) Object Scanning



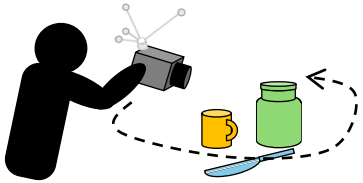
(b) Tip Calibration



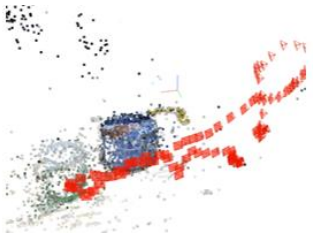
(c) Pose Annotation



(d) Hand-Eye Calibration

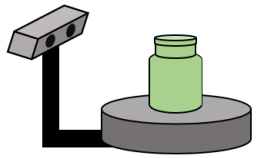


(e) Trajectory Recording

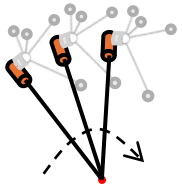


(f) Trajectory Refinement

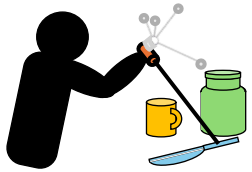
HouseCat6D Dataset : Pipeline



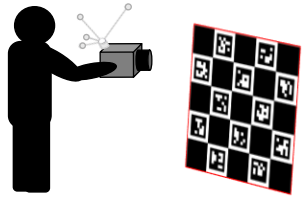
(a) Object Scanning



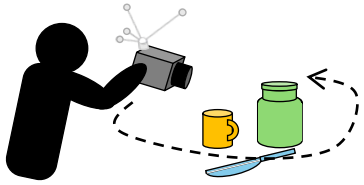
(b) Tip Calibration



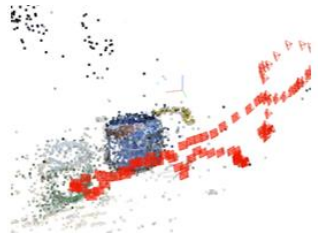
(c) Pose Annotation



(d) Hand-Eye Calibration



(e) Trajectory Recording



(f) Trajectory Refinement

Shining 3D EinScan

Glass

Bottle

Can

Tube

Box

x10 Categories
x194 Objects

Cutlery

Cup

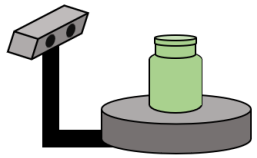
Shoe

Remote

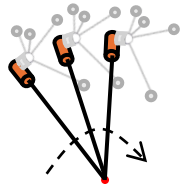
Teapot

(a) Object Scanning

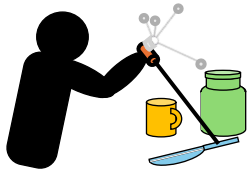
HouseCat6D Dataset : Pipeline



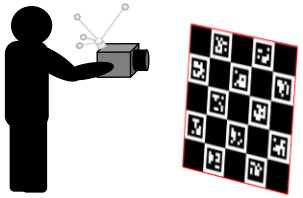
(a) Object Scanning



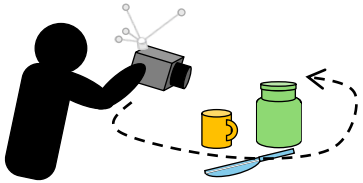
(b) Tip Calibration



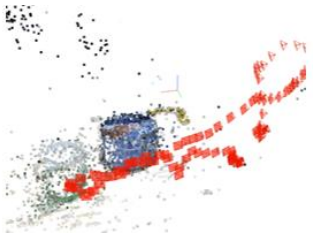
(c) Pose Annotation



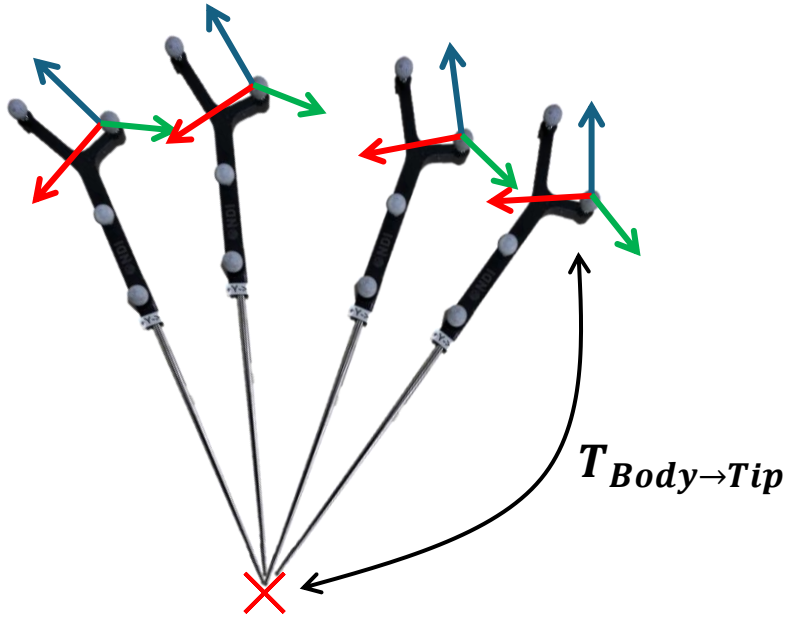
(d) Hand-Eye Calibration



(e) Trajectory Recording

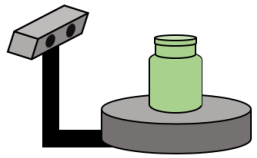


(f) Trajectory Refinement

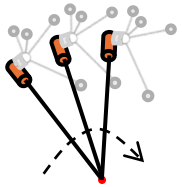


(b) Tip Calibration

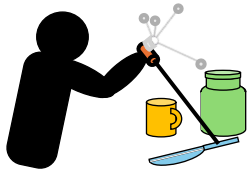
HouseCat6D Dataset : Pipeline



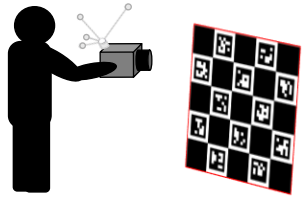
(a) Object Scanning



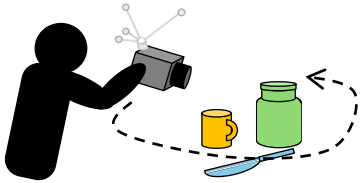
(b) Tip Calibration



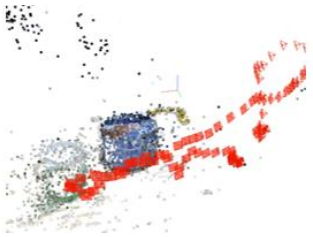
(c) Pose Annotation



(d) Hand-Eye Calibration



(e) Trajectory Recording

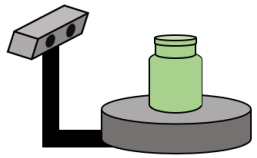


(f) Trajectory Refinement

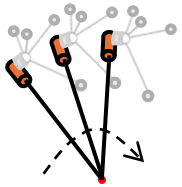
Collect
~25pts

(c) Pose Annotation

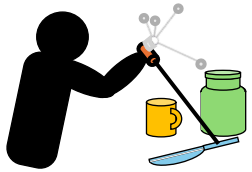
HouseCat6D Dataset : Pipeline



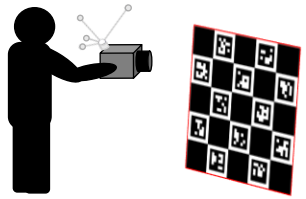
(a) Object Scanning



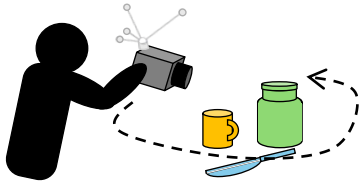
(b) Tip Calibration



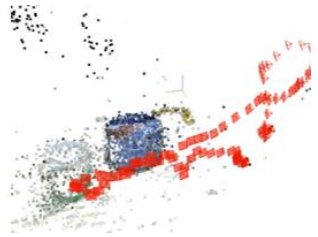
(c) Pose Annotation



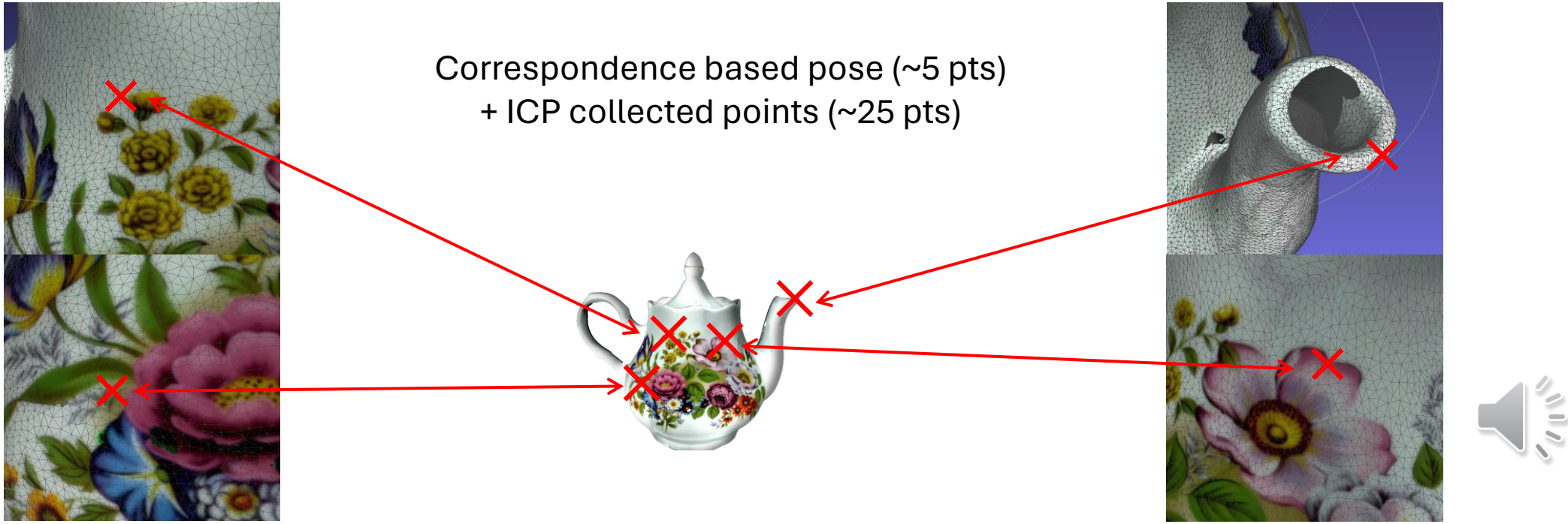
(d) Hand-Eye Calibration



(e) Trajectory Recording

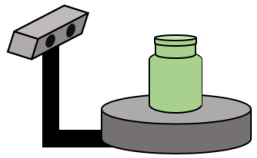


(f) Trajectory Refinement

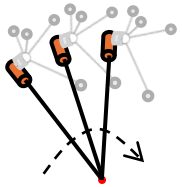


(c) Pose Annotation

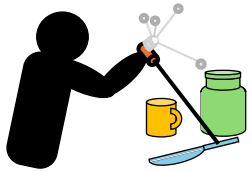
HouseCat6D Dataset : Pipeline



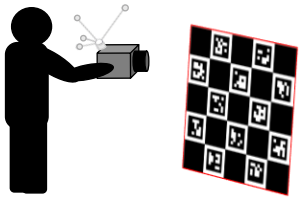
(a) Object Scanning



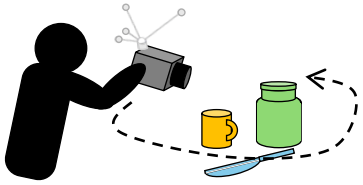
(b) Tip Calibration



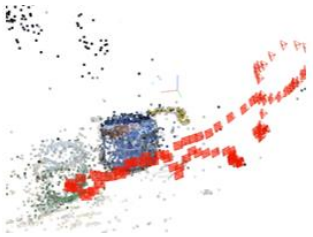
(c) Pose Annotation



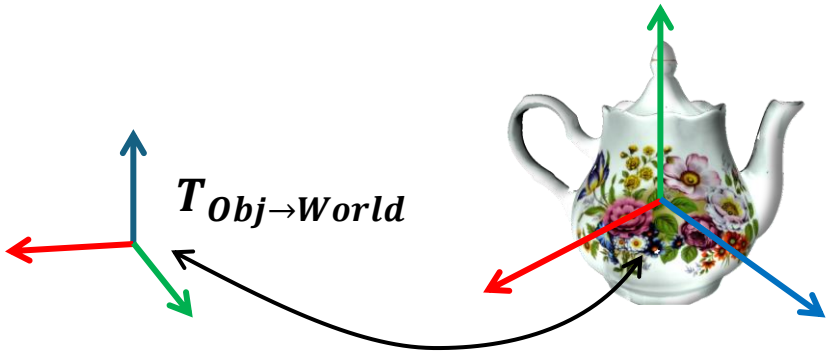
(d) Hand-Eye Calibration



(e) Trajectory Recording

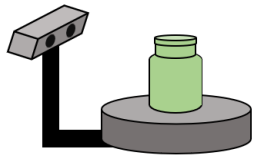


(f) Trajectory Refinement

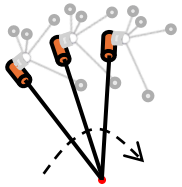


(c) Pose Annotation

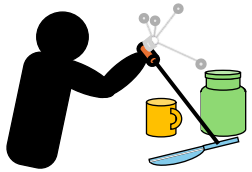
HouseCat6D Dataset : Pipeline



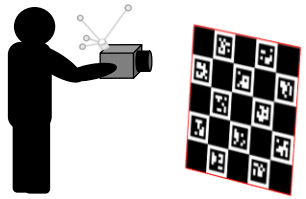
(a) Object Scanning



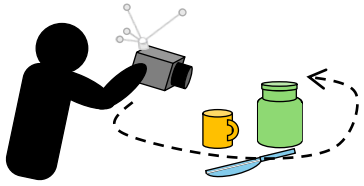
(b) Tip Calibration



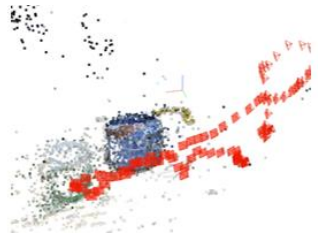
(c) Pose Annotation



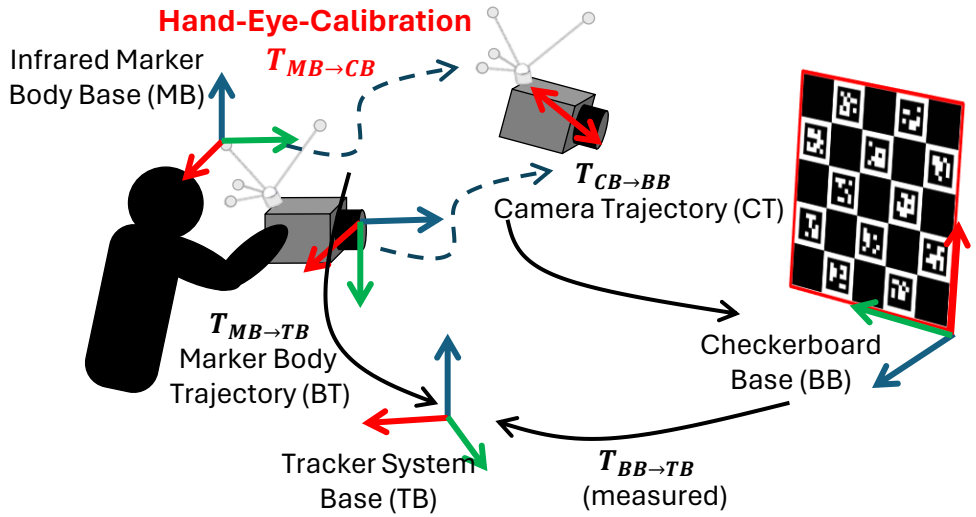
(d) Hand-Eye Calibration



(e) Trajectory Recording



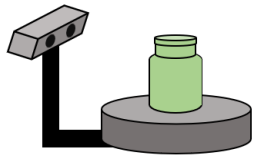
(f) Trajectory Refinement



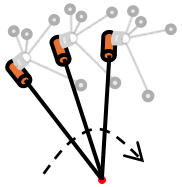
$$T_{MB\to CB} = Align(T_{MB\to TB}, T_{BB\to TB} \cdot T_{CB\to BB})$$

(d) Hand-Eye Calibration

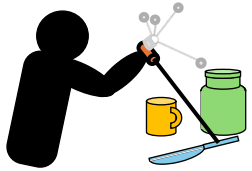
HouseCat6D Dataset : Pipeline



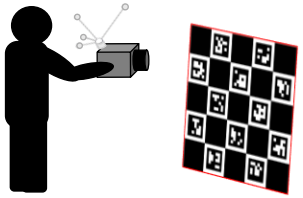
(a) Object Scanning



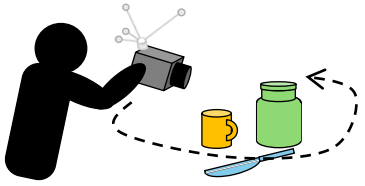
(b) Tip Calibration



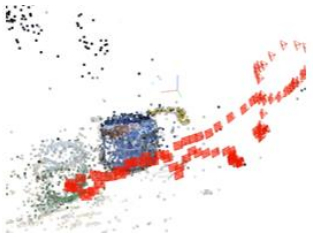
(c) Pose Annotation



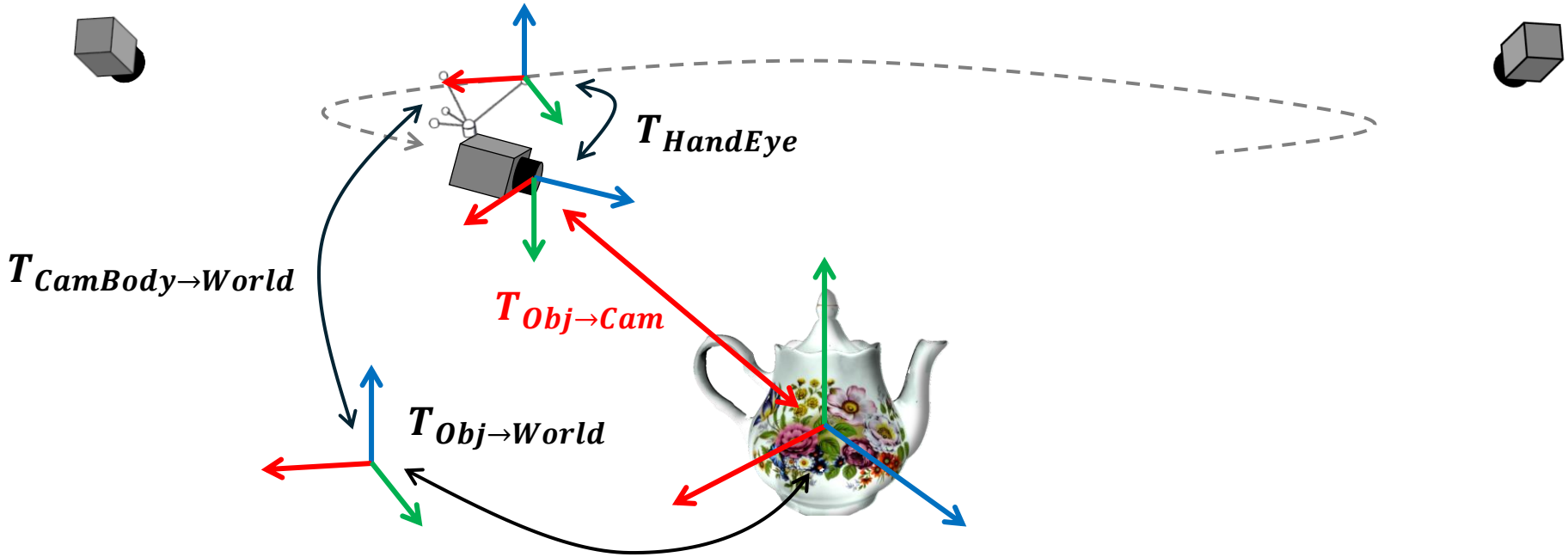
(d) Hand-Eye Calibration



(e) Trajectory Recording

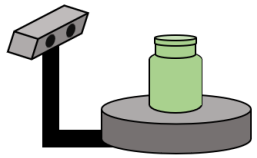


(f) Trajectory Refinement

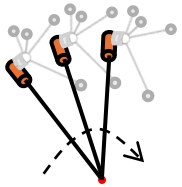


(e) Trajectory Recording

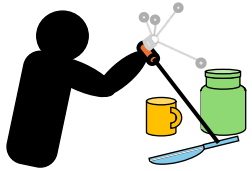
HouseCat6D Dataset : Pipeline



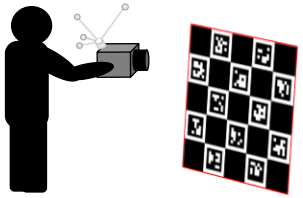
(a) Object Scanning



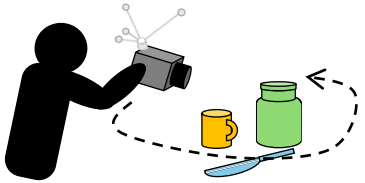
(b) Tip Calibration



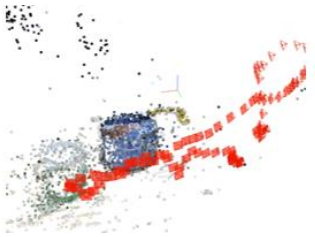
(c) Pose Annotation



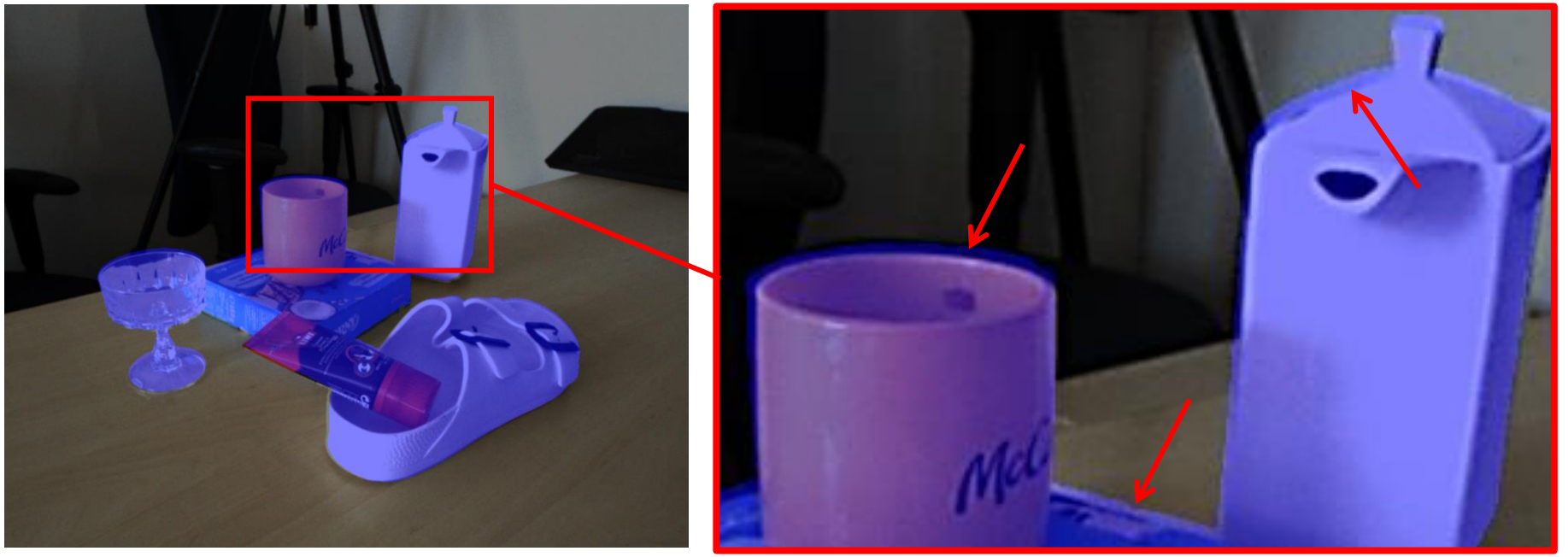
(d) Hand-Eye Calibration



(e) Trajectory Recording

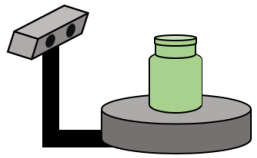


(f) Trajectory Refinement

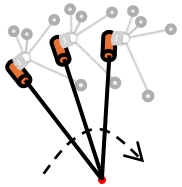


(f) Trajectory Refinement

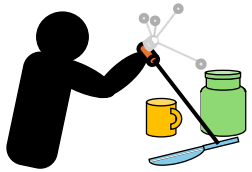
HouseCat6D Dataset : Pipeline



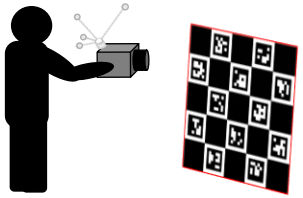
(a) Object Scanning



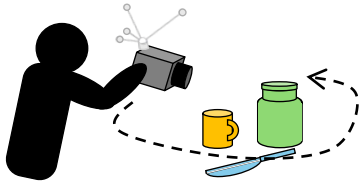
(b) Tip Calibration



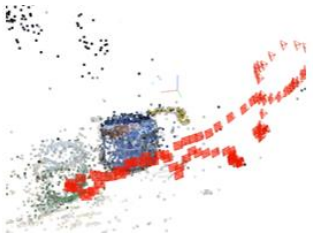
(c) Pose Annotation



(d) Hand-Eye Calibration

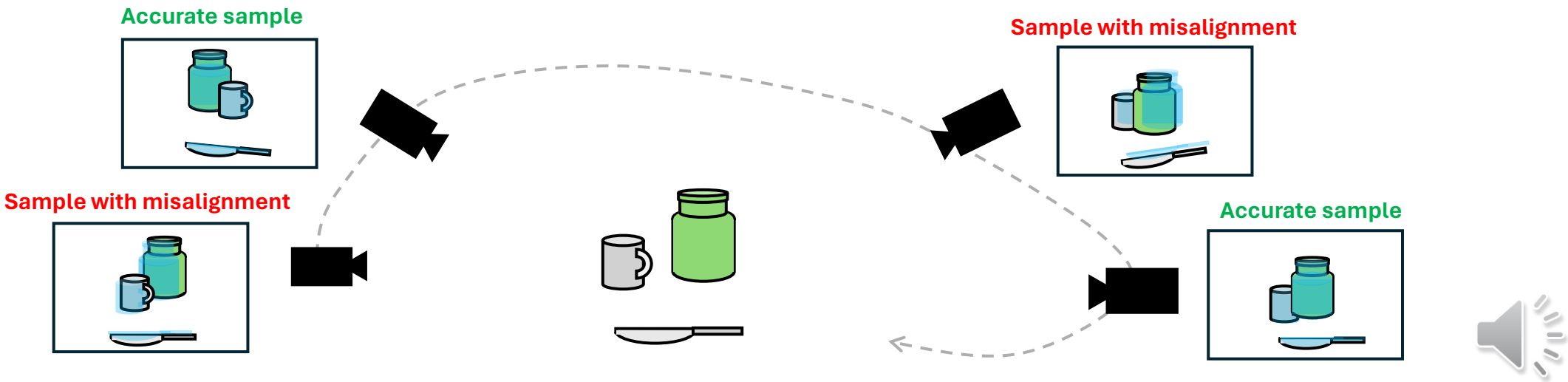


(e) Trajectory Recording



(f) Trajectory Refinement

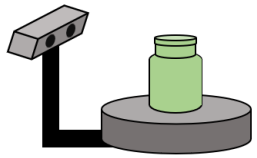
Run COLMAP [1,2] with initial pose, and fix few accurate samples to keep the scale correct



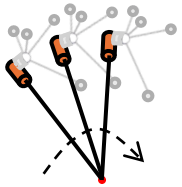
(f) Trajectory Refinement

[1] **Structure-from-Motion Revisited** J.L.Schoenberger, J.M.Frahm (CVPR 2016)
[2] **Pixelwise View Selection for Unstructured Multi-View Stereo** J.Schoenberger, L.Johannes E.Zheng, M.Pollefeys, J.M.Frahm (ECCV 2016)

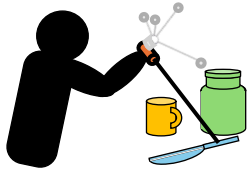
HouseCat6D Dataset : Pipeline



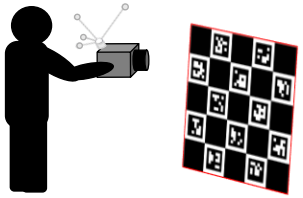
(a) Object Scanning



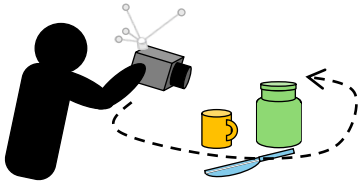
(b) Tip Calibration



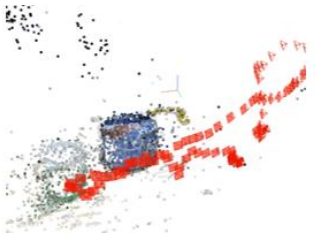
(c) Pose Annotation



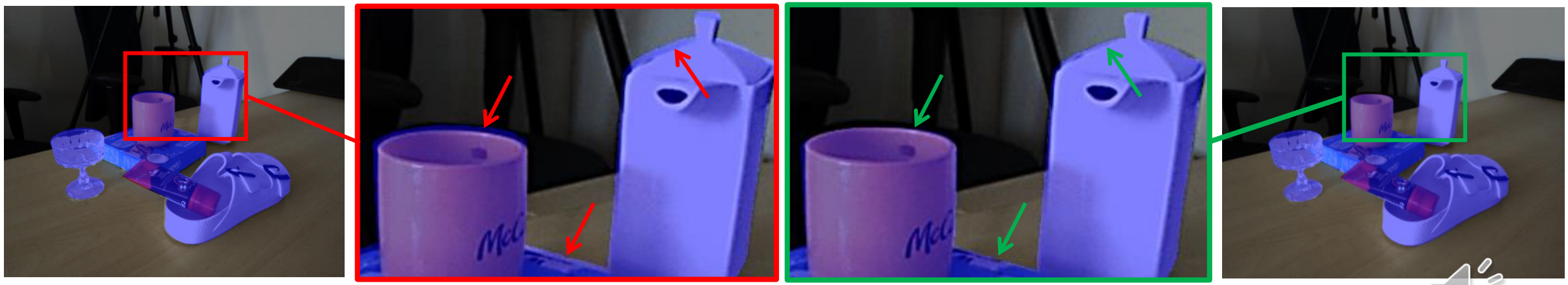
(d) Hand-Eye Calibration



(e) Trajectory Recording



(f) Trajectory Refinement



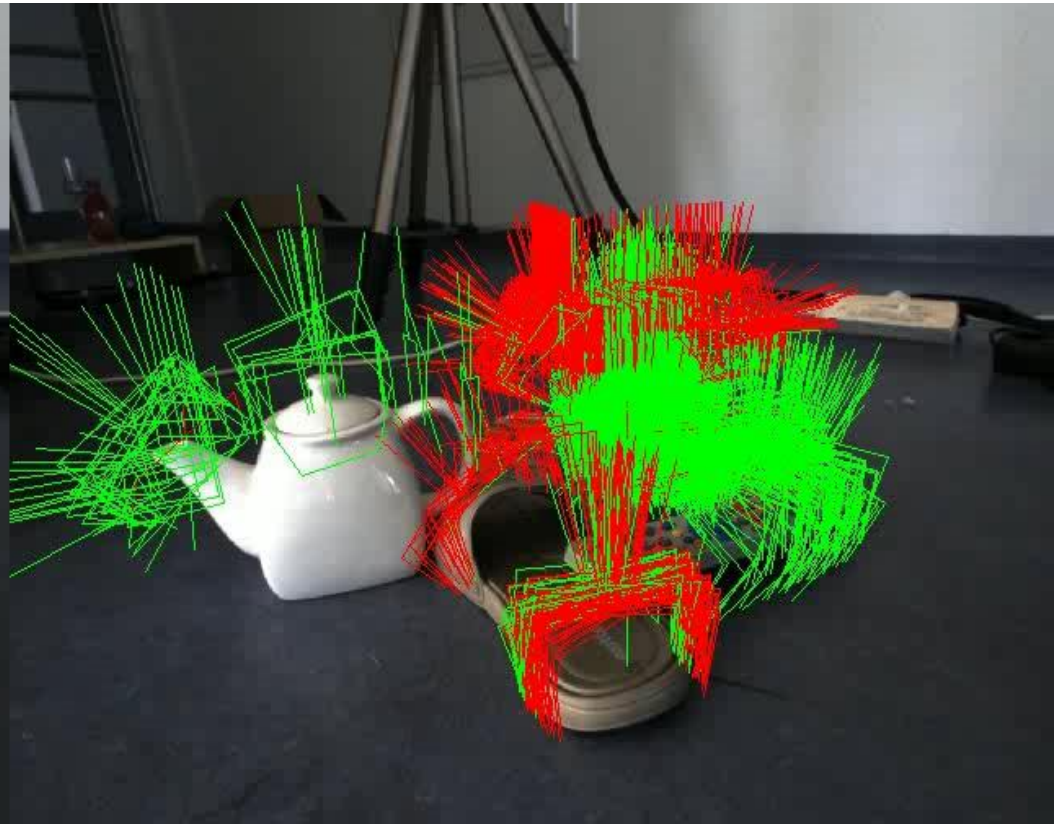
(a) Rendered object mask on camera trajectory without COLMAP refinement step

(b) Rendered object mask on camera trajectory with COLMAP refinement step

(f) Trajectory Refinement

HouseCat6D Dataset : Example

(a) Rendered Masks from Annotated Objects



(b) Rendered Grasping Poses from Annotations

Dataset	RGBD based	TOD [1]	StereOBJ [2]	PhoCal [3]	Ours
3D Labeling	Depth Map	Multi-View	Multi-View	Robot	IR tracker
Point RMSE	$\geq$ 17 mm	3.4 mm	2.3 mm	0.80 mm	1.35 mm $\leq \epsilon \leq$ 1.73 mm

[1] **Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects**, X.Liu, R.Jonschkowski, A.Angelova, K.Konolige (CVPR 2020)

[2] **Stereobj-1m : Large-scale stereo image dataset for 6d object pose estimation**, X.Liu, S.Iwase, K.M.Kitani (ICCV 2021)

[3] **PhoCal : A Multi-Modal Dataset for Category-Level Object Pose Estimation with Photometrically Challenging Objects**,P.Wang, HJ.Jung, Y.Li, S.Shen, RP.Srikanth, L.Garattoni, S.Meier, N.Navab, B.Busam (CVPR 2022)

HouseCat6D Dataset : Experiments

Pose Estimation Benchmarks :

Approach	$3D_{25}$ / $3D_{50}$	Bottle	Box	Can	Cup	Remote	Teapot	Cutlery	Glass	Tube	Shoe
NOCS [1]	50.0 / 21.2	41.9 / 5.0	43.3 / 6.5	81.9 / 62.4	68.8 / 2.0	81.8 / 59.8	24.3 / 0.1	14.7 / 6.0	95.4 / 49.6	21.0 / 4.6	26.4 / 16.5
FS-Net [2]	74.9 / 48.0	65.3 / 45.0	31.7 / 1.2	98.3 / 73.8	96.4 / 68.1	65.6 / 46.8	69.9 / 59.8	71.0 / 51.6	99.4 / 32.4	79.7 / 46.0	71.4 / 55.4
GPV-Pose [3]	74.9 / 50.7	66.8 / 45.6	31.4 / 1.1	98.6 / 75.2	**96.7** / 69.0	**65.7 / 46.9**	**75.4 / 61.6**	70.9 / 52.0	**99.6 / 62.7**	76.9 / 42.4	67.4 / 50.2
VI-Net [4]	**80.7 / 56.4**	**90.6 / 79.6**	**44.8 / 12.7**	**99.0 / 67.0**	**96.7 / 72.1**	54.9 / 17.1	52.6 / 47.3	**89.2 / 76.4**	99.1 / 93.7	**94.9 / 36.0**	**85.2 / 62.4**

Ablation Study on the Smaller Scale Training Set :

Approach	Train Set	$3D_{25}$ / $3D_{50}$	Bottle	Box	Can	Cup	Remote	Teapot	Cutlery	Glass	Tube	Shoe
VI-Net [4]	Full	**80.7 / 56.4**	90.6 / **79.6**	44.8 / 12.7	**99.0** / 67.0	**96.7 / 72.1**	54.9 / 17.1	**52.6 / 47.3**	**89.2 / 76.4**	**99.1 / 93.7**	**94.9** / 36.0	85.2 / **62.4**
	RV	74.2 / 46.8	**91.0** / 76.6	**59.1 / 23.5**	98.9 / **67.2**	76.0 / 36.6	**59.4 / 34.3**	22.7 / 18.8	79.4 / 57.3	97.7 / 85.3	66.3 / **47.8**	**91.4** / 20.4
	RS	67.7 / 35.8	90.1 / 68.7	49.0 / 9.8	96.9 / 53.6	87.2 / 48.5	40.2 / 16.3	28.8 / 15.8	67.4 / 49.0	98.5 / 73.6	86.6 / 7.9	32.4 / 14.9

[1] **Normalized object co- ordinate space for category-level 6d object pose and size estimation.** H.Wang, S.Sridhar, J.Huang, J.Valentin, S.Song, LJ.Guibas. (CVPR 2019)
[2] **Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism.** W.Chen, X.Jia, HJ.Chang, J. Duan, L.Shen, A.Leonardis. (CVPR 2021)
[3] **Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting.** Y.Di, R.Zhang, Z.Lou, F.Manhardt, X.Ji, N.Navab, F.Tombari. (CVPR 2022)
[4] **Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations.** J.Lin, Z.Wei, Y.Zhang, K.Jia. (CVPR 2023)

HouseCat6D Dataset : Experiments

CVPR
JUNE 17-21, 2024
SEATTLE, WA

HouseCat6D – A Large-Scale Multi-Modal Category Level 6D Object Perception Dataset with Household Objects in Realistic Scenarios

HyunJun Jung[1,*], Shun-Cheng Wu[1,*], Patrick Ruhkamp[1,*], Guangyao Zhai[1,2,†,*], Hannah Schieber[1,3,7,*], Giulia Rizzoli[4], Pengyuan Wang[1], Hongcheng Zhao[1], Lorenzo Garattoni[5], Sven Meier[5], Daniel Roth[1,7], Nassir Navab[1], Benjamin Busam[1,2,6]

hyunjun.jung@tum.de guangyao.zhai@tum.de shuncheng.wu@tum.de b.busam@tum.de

Poster ID12086 , Friday, 10:30 - 12:00

[1] Technical University of Munich, [2] Munich Center for Machine Learning, [3] FAU Erlangen-Nüremberg, [4] University of Padova, [5] Toyota Motor Europe, [6] 3dwe.ai, [7] Klinikum rechs der Isar, * Equal Contributions, † Corresponding Author

Munich Center for Machine Learning
Friedrich-Alexander-Universität Erlangen-Nürnberg
3Dwe
Klinikum rechts der Isar
Technische Universität München