EgoXtreme: A Dataset for Robust Object Pose Estimation in Egocentric Views under Extreme Conditions

Supplementary Material

A. Dataset specifications

Capture device The primary data streams were captured using Project Aria glasses [1], a lightweight platform designed for egocentric machine perception and providing tightly calibrated and time-synchronized streams crucial for our highly dynamic data capture. The core visual stream utilizes one rolling-shutter RGB camera, recording at 30 fps with 1408×1408 px resolution and an F-Theta fisheye lens ($110°$ FOV). Concurrently, dual integrated Inertial Measurement Units (IMUs) capture motion data at a high frequency (up to 1000 Hz), which is essential for precise temporal synchronization and SLAM trajectory reconstruction. To facilitate precise time synchronization and trajectory alignment with the external motion capture system, seven reflective markers were rigidly attached to the Aria glasses.

Figure A1. **Project Aria.** RGB capture device.

Test bed The experiments were conducted within a dedicated laboratory space measuring 2.4m $\times$ 2.6m (width $\times$ depth), with a ceiling height of 3.2m. The room was completely blacked out to eliminate external light interference, ensuring strict control over illumination conditions during testing. High-accuracy ground truth pose data was captured using a motion capture system consisting of four OptiTrack cameras mounted at the 3.2m height of the ceiling. This system offers high precision, featuring 1280×1024 resolution, 0.2mm 3D accuracy, a field of view of $56° \times 45°$, and supports frame rates up to 240fps, ensuring robust tracking even under rapid motion.

Ground Truth Validation We aligned SLAM and Mocap trajectories to leverage their complementary strengths: while SLAM provides continuous tracking, it is susceptible to low light and fast motion; Mocap (IR-based) remains robust in such conditions but suffers from occasional tracking loss due to occlusion. To validate this alignment, we randomly sampled 200 frames per scenario, selecting only those where keypoints were visibly verifiable. The resulting mean reprojection errors (1408×1408 res.) were 12.40 px (0.62%) for sports, 10.77 px (0.54%) for maintenance, and 19.08 px (0.96%) for emergency. The final average trajectory alignment error was $4.3\text{mm}/1.5°$, validating the reliability of our GT even under such extreme constraints.

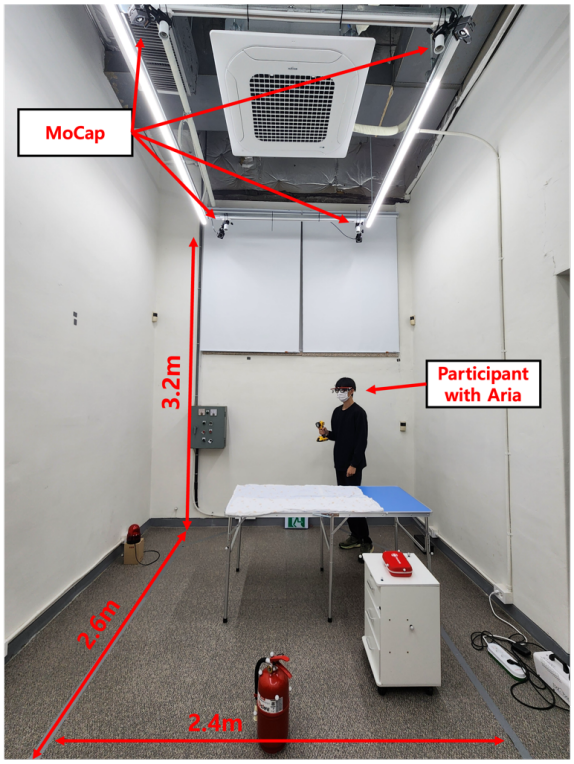

Figure A2. **Test bed.** Motion capture systems.

Data collection procedure A total of 15 participants took part in the study. Before commencing data capture, all subjects signed a consent form, and were thoroughly briefed by the principal investigator regarding the study's scope, motion capture boundaries, and specific behavioral guidelines. Crucially, participants were also required to wear protective masks during some sessions involving smoke injection to ensure safety and compliance with protocol. The data collection process required approximately three hours per participant, during which an average of 80 video sequences were captured per subject. In total, 845 video sequences were recorded.

Symmetry properties We identified *Bat*, *Brick*, and *Tennis* as symmetric objects within our dataset. Consequently, we utilized the ADD-S metric for these objects to account for their geometric symmetry during evaluation, ensuring a robust performance assessment.

B. Qualitative results of temporal dynamics

Figure B1 illustrates the diverse temporal dynamics within our dataset. The visualization is divided into three sections (top-to-bottom), illustrating the dataset's core challenges and temporal characteristics: rows 1–2 showcase severe motion blur (1-frame intervals); rows 3–4 capture the progression of visual obstruction caused by simulated smoke (10-frame intervals); and rows 5–6 demonstrate rapid egocentric perspective change (4-frame intervals).

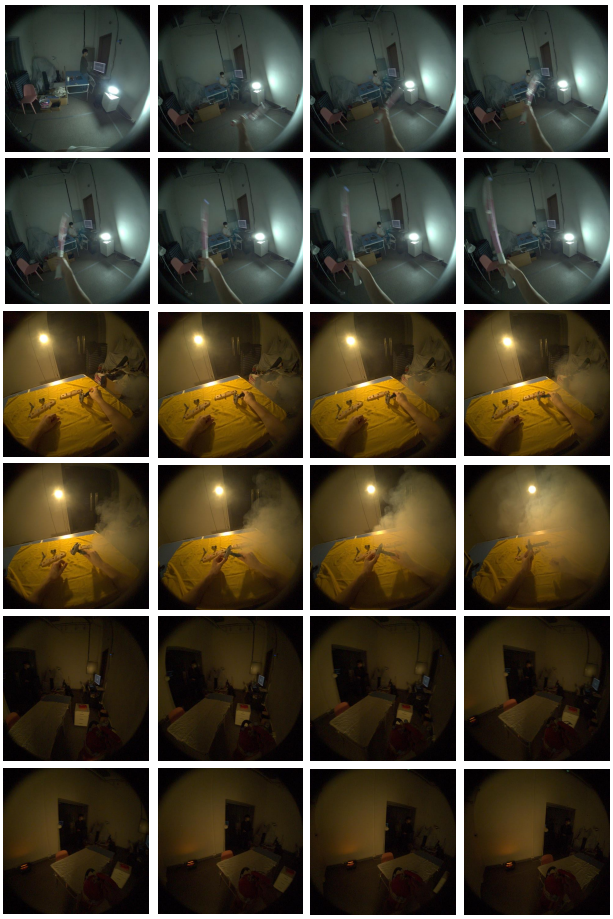

Figure B1. **Visualization of temporal dynamics in the *EgoXtreme* dataset.**

C. Extended baseline analysis

C.1. End-to-end 6D object pose estimation

In Table C1, we present the end-to-end evaluation results using CNOS [4] detections. As observed, the relatively low detection success (AP@0.5) leads to a significant drop in end-to-end pose accuracy (ADD-0.3d) compared to the GT baseline reported in the main paper (Table 3). These results justify our experimental design in the main paper, where we utilized GT bounding boxes to rigorously evaluate pose estimation performance decoupled from detection errors.

Table C1. **End-to-end 6D object pose estimation using CNOS detections.**

Scenario	Light	Smoke	Detection	Pose estimation (ADD-0.3d)		
			CNOS	FoundPose	GigaPose	PicoPose
Sports	Standard		17.85	2.12	9.38	8.68
	Extreme		8.49	0.69	3.44	3.75
Maintenance	Standard		41.76	22.77	33.78	41.30
	Extreme		22.97	11.67	15.32	21.13
	Standard	✔	30.86	14.43	19.94	23.36
	Extreme	✔	18.52	9.73	12.56	15.25
Emergency	Standard		42.10	16.88	34.51	18.21
	Extreme		16.37	4.08	8.16	1.10
	Standard	✔	35.71	12.25	28.78	31.88
	Extreme	✔	16.39	2.41	6.17	2.33

C.2. Instance-level and model-free baselines

Table C2 summarizes the instance-level (GDRNPP [3]) and model-free (OnePose++ [2]) methods on the *Tennis* sequence to assess annotation quality and task difficulty. GDRNPP achieves high accuracy across both conditions, serving as a fully-supervised upper bound that validates the reliability of our ground truth annotations. In contrast, the reconstruction-based OnePose++ fails significantly due to rapid motion and frequent occlusions. These results highlight the challenging nature of our dataset and suggest that model-based approaches remain a necessary prerequisite for robustness in this domain.

Table C2. **Additional baseline results on the *Tennis* sequence.**

Method	Condition	0.1d	0.2d	0.3d
GDRNPP (Instance-level)	Standard	84.96	95.06	96.50
	Extreme	74.15	90.14	93.64
OnePose++ (Model-free)	Standard	0.46	8.09	20.96
	Extreme	0.11	4.76	14.15

D. Extended analysis of 6D pose tracking

D.1. Evaluation of GoTrack baseline

Table D1 presents the evaluation of GoTrack [5] on the *Sports* scenario. The results indicate that the 'Direct' tracking mode suffers significant degradation compared to the

per-frame baseline. This aligns with our main findings (Table 5), confirming that rapid egocentric motion renders the previous frame's pose unreliable for initialization.

Table D1. **6D object pose tracking using GoTrack.**

Object	Method	GoTrack(GigaPose)		
		0.1d	**0.2d**	**0.3d**
Pingpong	Per-frame	1.42	3.51	17.33
	Direct	0.37	0.85	4.95
Tennis	Per-frame	13.81	35.17	44.29
	Direct	7.14	9.74	11.67
Bat	Per-frame	14.43	29.78	46.66
	Direct	3.94	8.35	11.62
Golf	Per-frame	0.45	1.56	3.87
	Direct	0.48	0.61	0.66
Hockey	Per-frame	0.57	4.95	14.24
	Direct	0.53	2.30	4.95

D.2. Evaluation under all light conditions

Table D2 summarizes the 6D object pose tracking results in the sports and emergency scenarios under all light conditions. Our findings indicate that the hybrid tracking approach provided performance gains even in the lower-motion Emergency scenario. Furthermore, analyzing the failure cases under extreme conditions suggests that performance enhancement in low-light environments requires feature restoration pre-processing to be successfully applied before the tracking step.

D.3. Qualitative results

Figure D1 visualizes the comparative performance of four distinct 6D pose temporal strategies. In the high-motion sports scenario, prediction often failed using the simple direct temporal approach due to large inter-frame displacement, leading to frequent failures. While the fusion temporal approach performed better, it still struggled to accurately stabilize rotation. The hybrid temporal method ultimately showed the most significant and reliable improvement in this difficult setting. Conversely, in the emergency scenario, where object movement was relatively minimal, the fusion temporal approach already provided a satisfactory performance level. However, the hybrid method still yielded the greatest overall stability gain, demonstrating its superior ability to fuse reliable measurements over simple propagation.

E. Extended analysis of image restoration

E.1. Pre-processing results under all conditions

Table E1 details the impact of image restoration pre-processing for 6D object pose estimation, analyzing the ef-

Table D2. **6D object pose tracking for GigaPose.**

Object	Method	Lighting condition					
		Standard			Extreme		
		0.1d	**0.2d**	**0.3d**	**0.1d**	**0.2d**	**0.3d**
Pingpong	Per-frame	2.17	5.36	17.84	2.77	5.94	15.15
	Direct	0.33	1.25	5.22	0.36	1.03	2.01
	Fusion	0.76	2.61	11.43	0.92	2.30	8.34
	Hybrid	2.44	4.59	19.07	3.32	5.84	16.51
Tennis	Per-frame	9.29	42.64	59.20	6.13	35.59	52.03
	Direct	2.92	14.52	23.64	0.57	7.79	14.84
	Fusion	4.98	38.12	57.49	5.56	28.37	47.62
	Hybrid	10.83	42.61	58.72	6.36	34.84	50.77
Bat	Per-frame	20.35	42.67	63.93	14.13	34.93	57.25
	Direct	7.55	18.20	24.57	0.0	0.72	1.92
	Fusion	14.28	38.95	64.92	11.36	35.47	60.82
	Hybrid	23.08	49.05	67.86	16.99	39.49	59.75
Golf	Per-frame	0.11	1.48	8.30	0.03	1.08	8.36
	Direct	0.63	1.86	3.92	0.0	0.50	2.33
	Fusion	0.66	5.53	14.07	0.65	2.10	7.13
	Hybrid	0.30	3.71	15.47	0.20	3.18	14.66
Hockey	Per-frame	0.29	4.46	16.26	0.13	2.33	7.50
	Direct	1.77	9.34	15.38	1.86	4.19	7.48
	Fusion	1.38	10.16	20.44	0.67	5.04	11.74
	Hybrid	0.91	8.56	22.43	0.40	3.08	10.18
Fire extinguisher	Per-frame	9.35	34.85	51.54	6.89	15.73	24.95
	Direct	8.75	21.12	25.55	3.78	6.62	8.21
	Fusion	13.45	34.25	42.79	5.36	10.43	15.72
	Hybrid	13.34	40.95	56.19	9.40	20.31	28.40
Kit	Per-frame	18.04	23.89	26.62	7.22	8.80	12.11
	Direct	19.31	23.31	24.52	2.60	4.38	4.83
	Fusion	20.92	26.38	29.16	7.96	10.44	11.62
	Hybrid	30.81	37.98	40.47	13.32	16.53	17.85
Flashlight	Per-frame	6.37	22.71	25.19	2.43	6.67	7.52
	Direct	6.99	16.52	17.20	0.17	1.59	1.66
	Fusion	8.17	20.11	21.29	0.32	2.16	2.85
	Hybrid	9.72	26.55	28.77	2.54	7.01	7.68

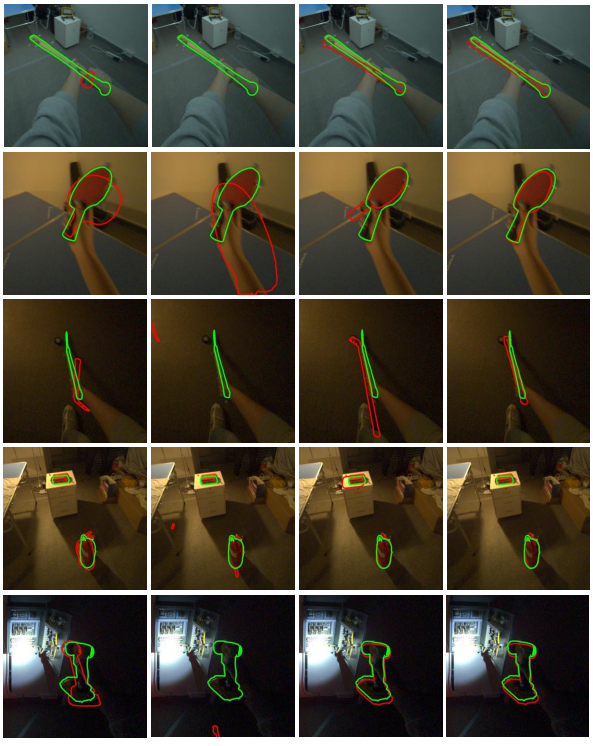

Figure D1. **Visualization of pose tracking performance.** The panels, arranged from left to right, showcase the results of the per-frame baseline, direct temporal, fusion temporal, and hybrid temporal methods.

ficacy across specific illumination conditions and the presence of simulated smoke. Across all tested sub-conditions, we observe a general performance degradation when applying pre-processing, underscoring its inherent limitations for downstream pose estimation tasks.

Table E1. **6D object pose estimation with pre-processing under conditions**.

Scenario	Conditions		Pre-processing			PicoPose		
	Light	Smoke	Deblur	Dehaze	Light enhance	0.1d	0.2d	0.3d
Sports	Standard					3.13	9.48	24.61
			✔			2.77	9.12	24.42
					✔	3.18	9.77	23.49
			✔		✔	2.81	8.91	23.12
Sports	Extreme					1.80	6.61	17.86
			✔			2.08	6.59	15.20
					✔	1.87	6.71	17.38
			✔		✔	1.84	5.99	14.13
Maintenance	Standard					39.27	62.42	76.84
			✔			37.72	57.40	71.21
				✔		34.15	54.47	69.60
					✔	37.62	59.07	73.18
			✔		✔	36.08	54.69	68.14
Maintenance	Extreme					26.44	47.18	64.09
			✔			26.83	45.11	60.08
				✔		22.93	43.40	60.65
					✔	22.13	41.02	57.58
			✔		✔	22.59	39.17	54.01
Maintenance	Standard	✔				26.37	46.05	59.87
			✔			20.63	36.69	49.94
				✔		21.52	39.54	52.94
					✔	22.95	41.02	54.88
			✔		✔	18.28	33.86	46.69
Maintenance	Extreme	✔				20.97	38.50	52.30
			✔			20.15	35.73	48.63
				✔		17.92	33.69	47.51
					✔	17.31	32.41	46.12
			✔		✔	17.03	31.47	44.03
Emergency	Standard					22.67	59.11	67.83
			✔			22.59	57.74	66.23
				✔		5.13	25.67	43.33
					✔	20.47	57.08	65.60
			✔		✔	21.62	55.92	64.60
Emergency	Extreme					9.18	27.59	36.23
			✔			8.62	21.74	27.59
				✔		4.65	17.26	28.42
					✔	8.62	24.46	31.87
			✔		✔	8.41	19.85	25.62
Emergency	Standard	✔				19.66	61.35	72.82
			✔			18.19	53.38	64.84
				✔		7.27	34.49	54.20
					✔	18.38	53.60	65.04
			✔		✔	17.75	49.82	61.83
Emergency	Extreme	✔				9.45	24.19	31.54
			✔			7.57	19.51	25.02
				✔		2.02	13.29	26.12
					✔	8.07	20.24	25.98
			✔		✔	7.55	18.04	24.12

E.2. Qualitative results

Figure E1, E2 and E3 visualizes the impact of image restoration pre-processing on 6D pose estimation. The upper row showcases the visual input across four pre-processing conditions (deblurring, dehazing, light enhancement, and deblurring + light enhancement). The lower row displays the corresponding predicted pose versus the ground truth. Critically, while the processed images show clear visual enhancement, the predicted masks are either not improved or, in some cases, are observed to even worsen.

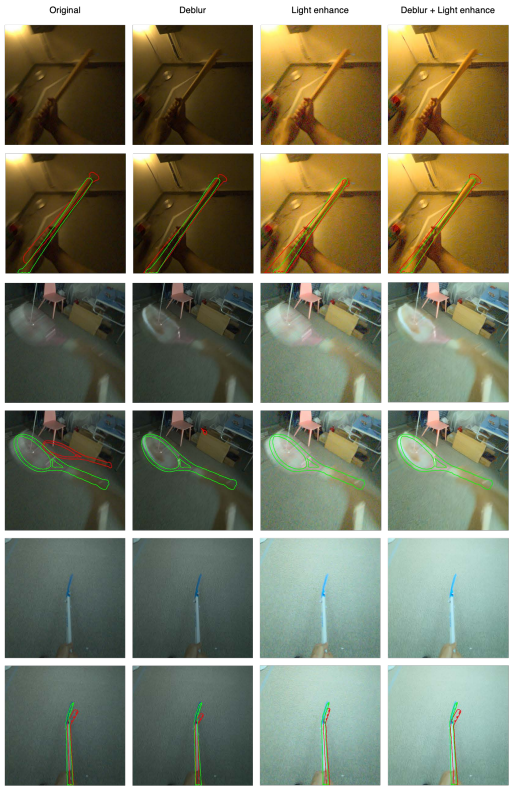

Figure E1. **Visualization of 6D Pose estimation results with pre-processing on sports scenario.**

Figure E2. **Visualization of 6D Pose estimation results with pre-processing on maintenance scenario.**

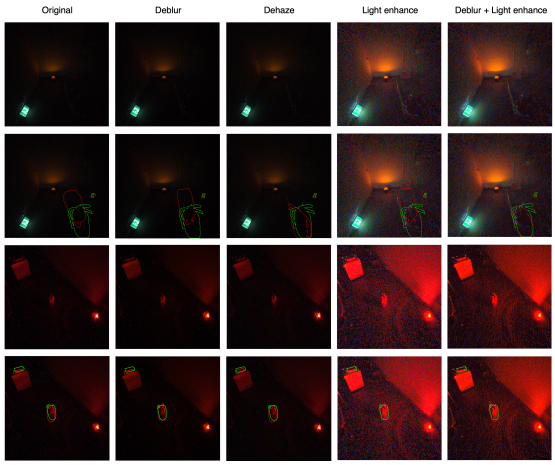

Figure E3. **Visualization of 6D Pose estimation results with preprocessing on emergency scenario.**

F. Institutional Review Board

The data collection protocol for this study was approved by the Institutional Review Board (IRB) of Seoul National University (IRB No. 2511_004-017). All participants provided informed consent prior to participating in the study.

References

[1] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 1

[2] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. In *Advances in Neural Information Processing Systems*, 2022. 2

[3] Xingyu Liu, Ruida Zhang, Chenyangguang Zhang, Gu Wang, Jiwen Tang, Zhigang Li, and Xiangyang Ji. Gdrnpp: A geometry-guided and fully learning-based object pose estimator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[4] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2134–2140, 2023. 2

[5] Van Nguyen Nguyen, Christian Forster, Bugra Tekin, Sindi Shkodrani, Vincent Lepetit, Cem Keskin, and Tomáš Hodaň. Gotrack: Generic 6dof object pose refinement and tracking. *Computer Vision and Patern Recognition Workshops (CVPRW)*, 2025. 2