This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Taegyoon Yoon[1] Yegyu Han[1] Seojin Ji[1] Jaewoo Park[1]
Sojeong Kim[1] Taein Kwon[2*] Hyung-Sin Kim[1*]

[1]Seoul National University [2]VGG, University of Oxford

{taegyoun88, yegyuhan, seojinji23, 1qkrwodn1, kvia2230, hyungkim}@snu.ac.kr
taein@robots.ox.ac.uk



Figure 1. ***EgoXtreme*, an egocentric dataset for robust 6D object pose estimation in extreme environments.** The dataset provides 775.5 minutes of egocentric RGB video from 15 participants using Aria glasses. As illustrated, it spans three challenging scenarios—Sports, Maintenance, and Emergency—featuring significant real-world visual degradations such as low light, smoke, and motion blur.

Abstract

Smart glass is emerging as an useful device since it provides plenty of insights under hands-busy, eyes-on-task situations. To understand the context of the wearer, 6D object pose estimation in egocentric view is becoming essential. However, existing 6D object pose estimation benchmarks fail to capture the challenges of real-world egocentric applications, which are often dominated by severe motion blur, dynamic illumination, and visual obstructions. This discrepancy creates a significant gap between controlled lab data and chaotic real-world application. To bridge this gap, we introduce EgoXtreme, a new large-scale 6D pose estimation dataset captured entirely from an egocentric perspective. EgoXtreme features three challenging scenarios—industrial maintenance, sports, and emergency rescue—designed to introduce severe perceptual ambiguities through extreme lighting, heavy motion blur, and smoke. Evaluations of state-of-the-art generalizable pose estimators on EgoXtreme indicate that their generalization fails to hold in extreme conditions, especially under low light. We further demonstrate that simply applying image restoration (e.g., deblurring) offers no positive improvement for extreme conditions. While performance gain has appeared in tracking-based approach, implying using temporal information in fast-motion scenarios is meaningful. We conclude that EgoXtreme is an essential resource for developing and evaluating the next generation of pose estimation models robust enough for real-world egocentric vision. The dataset and code are available at https://taegyoun88.github.io/EgoXtreme/

*Joint supervision and corresponding authors.

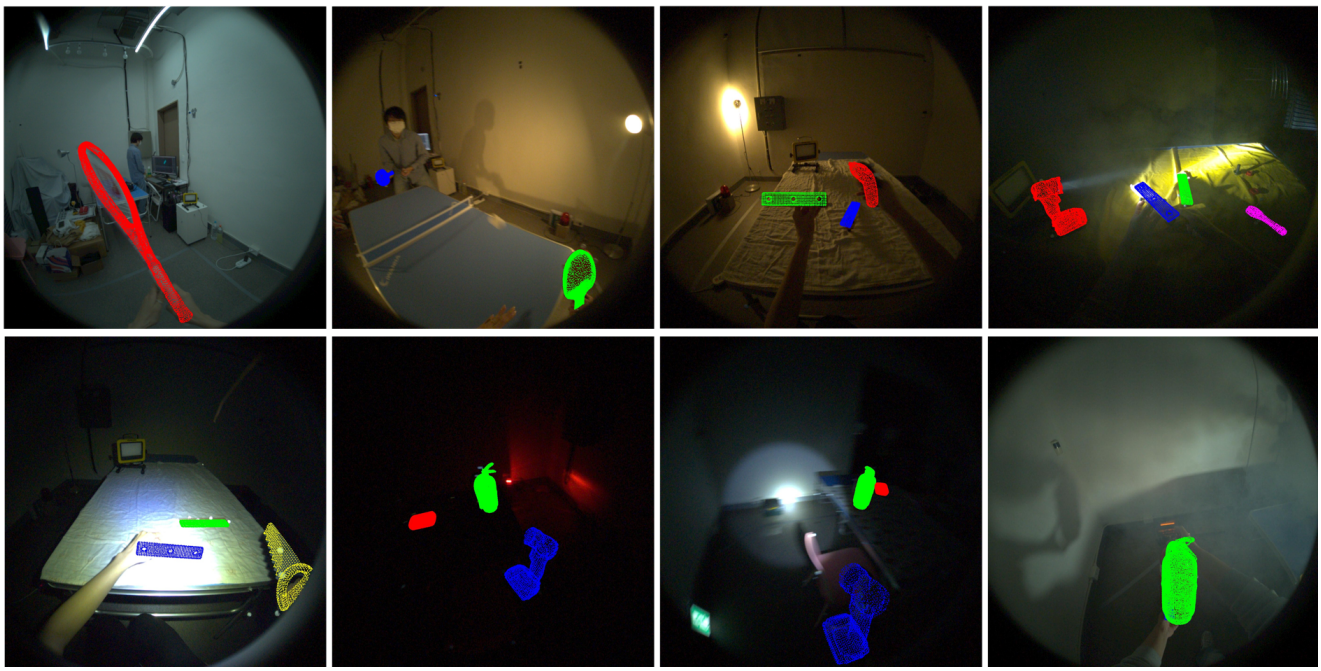


Figure 2. **Visualization of ground truth 6D pose annotations from our dataset, arranged sequentially from top-left to bottom-right.** The frames include the sports scenario (normal and middle light), the industrial maintenance scenario (low light, flashlight with smoke, and headlight), and the emergency rescue scenario (warning light, exit green light, and high light with smoke).

1. Introduction

Lightweight smart glasses [7] provide a powerful interface that can continuously capture and interpret the wearer's *egocentric* view, enabling vision applications precisely when users *cannot free both hands or divert their gaze to a phone*. Examples include tightening a bolt while following a torque sequence, navigating a dim corridor to locate an extinguisher, or executing a high-speed swing. In these moments, the camera is rigidly mounted to a moving head; objects appear at close range and are often truncated by hands or tools; and illumination is supplied or modulated by the wearer (headlamp or flash) or by the environment (warning beacons, exit-sign LEDs).

While 6D object pose estimation is fundamental to understanding physical activities and environments—and has achieved strong results on static third-person RGB datasets [2, 5, 6, 14–16, 19, 39, 43]—egocentric settings introduce unique challenges. In **hands-busy, eyes-on-task, always-moving** scenarios, the data distribution departs from stationary, uniformly lit third-person scenes and instead shifts toward rapid and frequent head motion; a near-field perspective at very short working distances; severe truncation at image boundaries; and extreme visual conditions (dynamic or narrow-spectrum illumination, smoke, and clutter) that induce motion blur and visibility loss. The very advantages of smart glasses as an interface therefore create the **sensing extremes** we must target.

Despite these realities, the field of 6D object pose estimation largely relies on controlled, third-person datasets such as YCB-Video [43], Honnotate [11], LineMod [14] , T-LESS [15], all collected under stable lighting and limited motion. While the H2O [21] and HOT3D [1] introduced egocentric benchmarks, they do not fully capture the diversity and severity of real-world smart-glasses conditions—particularly rapid object manipulations and dynamic lighting. In such extremes, severe motion blur significantly degrades the generalization of existing models. Attempts to bridge the gap via simulation—synthetic data generation [4], frame averaging to simulate blur [35], and artificial brightness variations [23]—remain insufficient: a substantial discrepancy persists between simulated and real egocentric scenes. The absence of an egocentric dataset that *directly captures* these extreme conditions is therefore a critical bottleneck for advancing 6D object pose estimation in egocentric vision.

This work aims to step toward bridging the gap between controlled laboratory settings and extreme cases that can occur in real-world usage, analyzing where current pose estimators fail under challenging egocentric conditions. To this end, we introduce *EgoXtreme*, a novel dataset specifically designed to evaluate robustness under dynamic lighting, smoke, and severe motion blur—conditions that induce significant perceptual ambiguity. The dataset comprises video sequences collected from 15 participants over 775.5 minutes, and Figure 2 provides representative sample images from *EgoXtreme*.

Table 1. **Datasets for object pose estimations.**

Dataset	Extreme Condition				Frames	Egocentric	Subjects	Objects	Annotation
	Instance	Speed	Light	Smoke					
LM [14]	No	-	1	No	18.2k	No	-	15	RGB-D
IC-BIN [43]	Yes	-	1	No	177	No	-	2	RGB-D
T-LESS [15]	No	-	1	No	49k	No	-	30	RGB-D
YCB-V [43]	No	Slow	1	No	0.1M	No	-	21	RGB-D
TUD-L [16]	No	-	8	No	62k	No	-	3	RGB-D
HOPE [39]	No	-	5	No	238	No	-	28	RGB-D
IPD [18]	Yes	-	3	No	30k	No	-	20	RGB-D
H2O [21]	No	Slow	1	No	572k	Yes	4	8	RGB-D
HOT3D [1]	No	Slow	1	No	1.5M	Yes	19	33	Mocap
EgoXtreme (Ours)	Yes	Fast	8	Yes	1.3M	Yes	15	13	Mocap

Our dataset presents three distinct scenarios:

- **Industrial maintenance:** Fine manipulation using tools like hammers and drills, testing robustness to subtle, precise motions.
- **Sports:** Rapidly swung objects (e.g., table tennis rackets, baseball bats), testing resilience to extreme motion blur from high-speed rotational and translational movement.
- **Emergency rescue:** Recreating urgent situations, users search for and interact with emergency items (e.g., first-aid kits, extinguishers), testing object detection and tracking amid intense camera shake and visual obstructions.

Our extensive evaluation of 6D object pose on *EgoXtreme* reveals that current state-of-the-art pose estimators [27, 31, 42]—despite large-scale pretraining aimed at generalizing to unseen objects in cluttered scenes—are highly fragile under egocentric extremes. This indicates that the challenges captured by *EgoXtreme* are largely orthogonal to those addressed by existing generalizable methods, exposing a gap in current benchmarks for egocentric applications. Specifically, we identify two primary failure modes: (1) distribution shift induced by dynamic illumination and smoke, and (2) feature loss caused by severe truncation and motion blur. Critically, the degradation persists even when perceptual image quality improves: visual restoration (e.g., deblurring, dehazing, and low-light enhancement) offers little benefit, and model performance remains significantly reduced. In contrast, incorporating temporal information via pose tracking improves accuracy in dynamic, high-motion scenarios, underscoring the importance of temporal modeling for 6D pose in egocentric video.

The core contributions of this paper are as follows:

- **A large-scale egocentric benchmark for robust pose estimation:** We release *EgoXtreme*, the first large-scale egocentric 6D pose estimation dataset collected across versatile scenarios under extreme conditions, including severe motion blur, dynamic lighting, and visual obstructions. This dataset can serve as a challenging benchmark for smart-glasses applications.
- **Comprehensive baseline analysis:** We evaluate state-of-the-art generalizable 6D object pose estimators on *EgoXtreme*, analyze their limitations and failure modes in extreme egocentric environments, and study two mitigation strategies—image restoration and temporal tracking—demonstrating the critical importance of the latter for robust performance under severe motion.

2. Related work

This section reviews existing research pertinent to our work, covering 6D object pose estimation models, relevant benchmark datasets, and methodologies aimed at improving performance under adverse conditions such as blur, haze, and low light.

2.1. Benchmarks for 6D object pose estimation

Standard benchmarks like YCB-Video [43] and LM/LM-O [2, 14] advanced the field but they only controlled limited conditions like lighting and motion. Datasets focusing on specific challenges like multiple instances/symmetry (IC-BIN [5], IC-MI [38], T-LESS [15]) or static lighting variations (TUD-L [16], HOPE [39], SenseShift6D [12]) exist, but lack the dynamic or extreme lighting. Table 1 indicates the difference between *EgoXtreme* and the other 6D pose estimation benchmarks.

6D pose estimation methods are broadly divided into instance-specific models (e.g., GDRN [41], ZebraPose [37], HiPose [26]) and zero-shot models. Instance-specific models are impractical for our egocentric scenario as they require object-specific retraining for novel objects. While zero-shot models offer generalization, RGB-D approaches (e.g., SAM6D [25], FoundationPose [42]) are often incompatible with lightweight smart glasses that lack depth sensors. We therefore focus on RGB-only zero-shot models. Two main strategies exist within this category. The first one is a two-step pipeline where a coarse estimator like GigaPose [31] and FoundPose [33] retrieves template

Table 2. ***EgoXtreme* datasets configuration**

Scenario	Lighting condition								Smoke	Object	Speed (m/s)	
	Standard			Extreme							Camera	Object
	normal	middle	high	low	head	flash	warning	green				
Maintenance	✔	✔	✔	✔	✔	✔			✔	5	0.03	0.09
Sports	✔	✔	✔	✔						5	0.40	1.37
Emergency	✔	✔	✔	✔			✔	✔	✔	3	0.47	0.10

candidates via feature matching for PnP-RANSAC [9, 24], and a separate refinement module (e.g., MegaPose [22], Gen-Flow [28]) iteratively improves the best hypothesis by aligning dense features. The second strategy uses integrated models like PicoPose [27], which perform the entire pipeline internally: they first find a single best-matched template and then use dedicated multi-stage blocks to compute 2D transformations and refine the final pose. Since our egocentric scenario requires an RGB-only unseen approach, we analyze the robustness of these state-of-the-art coarse-to-fine pipelines on our proposed extreme environment benchmark.

2.2. Egocentric pose estimation datasets

Recently, datasets capturing the first-person perspective have become crucial for smart glass applications. However, a significant portion of this research has centered on human-centric aspects, such as full-body 3D pose (e.g., You2Me [29], EgoBody [45]) or detailed hand pose and hand-object interactions (e.g., H2O [21]), rather than the 6D pose of general objects in the environment. HOT3D [1], the most prominent large-scale egocentric object tracking benchmark, was a significant step in addressing this. However, it was captured under relatively well-lit conditions and similarly lacks the extreme motion, challenging illumination, or visual obstructions like smoke that we target. Consequently, a significant gap remains in evaluating object pose estimation robustness under the severe motion blur, dynamic/extreme lighting, and drastic movements typical of real-world smart glass usage.

2.3. Methodologies for adverse conditions

Various methods aim to enhance image quality under adverse conditions. Deblurring techniques range from GAN-based approaches [20] to recent Transformer [44] and efficient architectures [3]. Dehazing methods include classic approaches [13] and deep learning models [34] or Transformer-based [36]. Low-light enhancement research offers specialized solutions, ranging from efficient zero-reference methods to multi-task restoration models. For instance, Zero-DCE [10] formulates enhancement as a fast, image-specific curve estimation task, while DarkIR [8] jointly addresses low-light, noise, and blurring



Figure 3. **3D models.** This image shows the 13 object models in the *EgoXtreme* dataset. From top to bottom five maintenance scenarios, five sports scenarios and three emergency scenarios.

issues by employing advanced attention mechanisms within efficient CNNs. While potentially beneficial for detection/segmentation, the impact of these restoration methods and potential artifacts on downstream pose estimation accuracy, especially in extreme conditions, is not well-studied. We investigate the effectiveness of applying representative enhancement techniques as preprocessing for robust 6D pose estimation on our benchmark.

3. *EgoXtreme* dataset

EgoXtreme is a novel, large-scale dataset designed for robust egocentric 6D object pose estimation under extreme conditions. Specifically, 8 illumination conditions are used across three scenarios, and smoke is included in specific scenes. The configurations of whole scenario is summarized in Table 2. These conditions, combined with severe motion blur, make accurate 6D object pose estimation extremely challenging.

Comprising approximately 1.3 million frames (775.5 minutes total) captured at 30fps (1650 frames/55 seconds per sequence), the dataset is split into training (518.8 min), validation (80.7 min), and test (176 min) sets. The dataset includes three challenging scenarios: industrial maintenance (319 min), emergency rescue (165 min), and sports (291.5 min). Videos were recorded using Aria glasses [7], providing raw 1408×1408 fisheye RGB images and the

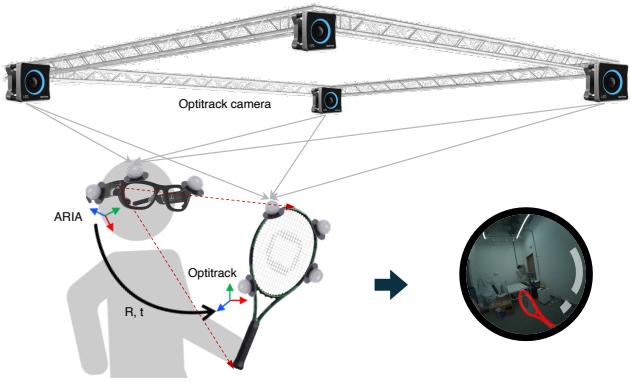


Figure 4. **Diagram for data collection.**

corresponding undistorted version, enabling evaluation on both data types. 15 participants were recorded performing diverse actions and object manipulations. The dataset features 13 objects, including sports equipment, assembly blocks, and emergency supplies, with corresponding 3D CAD models as in Figure 3. Some scenarios involve multiple instances of the same object (assembly blocks and pingpong racket), requiring instance-level disambiguation.

3.1. Scenario details

As shown in Table 2, the average speeds of the camera and the object differ significantly depending on the scenario which demonstrates the characteristics of each scenario.

Industrial maintenance scenario focuses on estimating object pose during seated precision tasks involving manipulating tools, including hammer, drill, saw, wrench and brick, and assembling identical multiple blocks in various industrial lighting conditions. 6 lighting conditions include normal brightness, low light, medium brightness, high brightness, headlamp, and flashlight. To mimic hazy workplaces, smoke was introduced in a subset of the data. This scenario allows for benchmarking model robustness against fine-grained object movements and diverse illumination changes, particularly the movable light sources are included.

Emergency rescue scenario involves searching for and picking up three types of emergency supplies, such as first aid kit, extinguisher, and flashlight, while moving through different lighting environments. 6 lighting conditions comprise normal brightness, low light, medium brightness, high brightness, emergency beacon, and exit sign lighting. Smoke was added in a portion to simulate fire situations. This setup evaluates the ability to detect objects and estimate object poses under intense camera shake and restricted visibility.

Sports scenario features five types of sports equipment, including baseball bat, hockey stick, pingpong racket, tennis racket, and golf club, being swung rapidly. To make it more realistic, actual pingpong gameplay or self-practice are included. 4 lighting conditions are used in this scenario: normal brightness, low light, medium brightness, and high brightness. This scenario evaluates the model's ability to handle extreme motion blur caused by high-speed rotational and translational object movements.

3.2. Data collection and processing

RGB images were captured at 30fps using the Aria glasses, synchronized with 1000fps SLAM data. A reflective marker was attached to the Aria glasses, and rigid body marker clusters were attached to all scenario objects, allowing simultaneous recording with a 120fps OptiTrack motion capture system. This system provides sub-millimeter accuracy for the 6D pose of both the headset and the objects, serving as our ground truth (GT). In this setup, both the SLAM trajectory and the motion capture marker trajectory were collected concurrently. The trajectories were aligned using the Umeyama method [40] to unify coordinate frames and achieve precise temporal synchronization between the Aria SLAM and the motion capture system. The whole process is illustrated in Figure 4.

To compensate for SLAM drift that occurred during motion, a Kalman filter [17] was applied to refine the trajectory alignment. Once the coordinate frames were unified through this process, various objects tracked by the motion capture system could be represented in the same global coordinate system, enabling accurate object pose annotation. Finally, to correct minor temporal misalignments, on the order of a few milliseconds, between the synchronized Aria timestamp and the true physical capture time of the RGB image caused by the RGB camera's exposure delay, a manual time-offset adjustment was performed based on visual inspection of projection results before final interpolation.

3.3. Dataset configuration

The lighting and environmental conditions within the dataset are structured for multi-dimensional evaluation of model robustness, as detailed in Table 2. The conditions are classified into two main categories: **Standard** (common illumination) and **Extreme** (limited illumination), comprising a total of 8 unique lighting types. The Standard category includes environments with consistent, ample light sources such as normal fluorescent light (normal), bright lamp light (middle), and floodlight illumination (high). Conversely, the extreme category features setups designed to challenge pose estimation: low-intensity lighting (low), dynamic illumination coupled with head movement (head, flash), and critical emergency sources (warning - rotating beacons and green - emergency exit signs). Furthermore, to accurately model visual degradation in maintenance and emergency scenarios, simulated smoke was intentionally introduced using a fog machine. This smoke was dispersed to maximize

Table 3. **6D object pose estimation on *EgoXtreme*.** Performance is evaluated using ADD(S) recall at thresholds of $0.1d$, $0.2d$, $0.3d$ (↑) and spatial accuracy metrics MSSD/MSPD (↓), where d denotes the object diameter.

Scenario	Light	Smoke	FoundPose [33]					GigaPose [31]					PicoPose [27]				
			0.1d	0.2d	0.3d	MSSD	MSPD	0.1d	0.2d	0.3d	MSSD	MSPD	0.1d	0.2d	0.3d	MSSD	MSPD
Sports	Standard		0.53	1.55	4.72	0.59	7.65	4.12	11.77	24.64	5.21	10.61	3.13	9.48	24.61	5.00	12.87
	Extreme		0.18	0.78	2.42	0.23	6.87	3.11	9.19	19.04	4.15	8.31	1.80	6.61	17.86	3.07	10.10
Maintenance	Standard		21.02	30.53	37.61	12.94	18.78	33.64	48.84	62.77	15.97	24.17	39.27	62.42	76.84	24.76	33.20
	Extreme		13.78	22.94	30.03	9.92	16.40	19.78	32.20	45.52	10.44	19.89	26.44	47.18	64.09	18.37	28.12
	Standard	✔	14.44	22.49	30.00	8.23	13.11	23.01	37.87	52.86	12.24	19.83	26.37	46.05	59.87	18.26	25.48
	Extreme	✔	11.19	18.57	25.63	7.01	13.07	17.56	30.35	45.11	8.92	18.38	20.97	38.50	52.30	14.25	25.53
Emergency	Standard		6.31	11.96	12.88	9.73	13.08	22.03	40.29	46.34	32.08	43.52	22.67	59.11	67.83	46.40	50.81
	Extreme		0.10	0.29	0.56	8.27	9.61	9.40	14.75	21.30	12.44	29.02	9.18	27.59	36.23	22.01	31.24
	Standard	✔	3.52	9.62	12.12	0.21	1.66	16.25	35.28	44.91	28.34	35.99	19.66	61.35	72.82	48.10	50.64
	Extreme	✔	0.11	0.60	0.76	0.44	2.05	7.07	15.26	21.54	12.00	32.81	9.45	24.19	31.54	19.37	29.62

atmospheric scattering and visual obstruction, ensuring that objects were not contaminated directly, thereby isolating the effects of airborne particulates on visibility.

4. Experiments

In this section, we evaluate the three state-of-the-art RGB-based models and analyze the performance of 6D object pose estimation under the proposed extreme conditions. Since real-world scenarios demand high generalizability, we focus our evaluation on state-of-the-art zero-shot 6D object pose estimation models. Specifically, we define the testing pipelines for each as follows: FoundPose [33] is used as a coarse alignment module with its output refined by MegaPose [22]. GigaPose [31] also utilize the MegaPose as a refinement module. Conversely, PicoPose [27] is employed as an integrated coarse-to-fine model performing its own internal refinement. For evaluation, we explicitly utilize the official test set. First, we establish baselines on the *EgoXtreme* dataset using state-of-the-art 6D object pose estimation models (Sec. 4.1). Second, we investigate how image restoration affects 6D object pose estimation performance (Sec. 4.2). Finally, we benchmark and evaluate tracking of the object pose under severe motion blur (Sec. 4.3).

4.1. Baseline evaluation

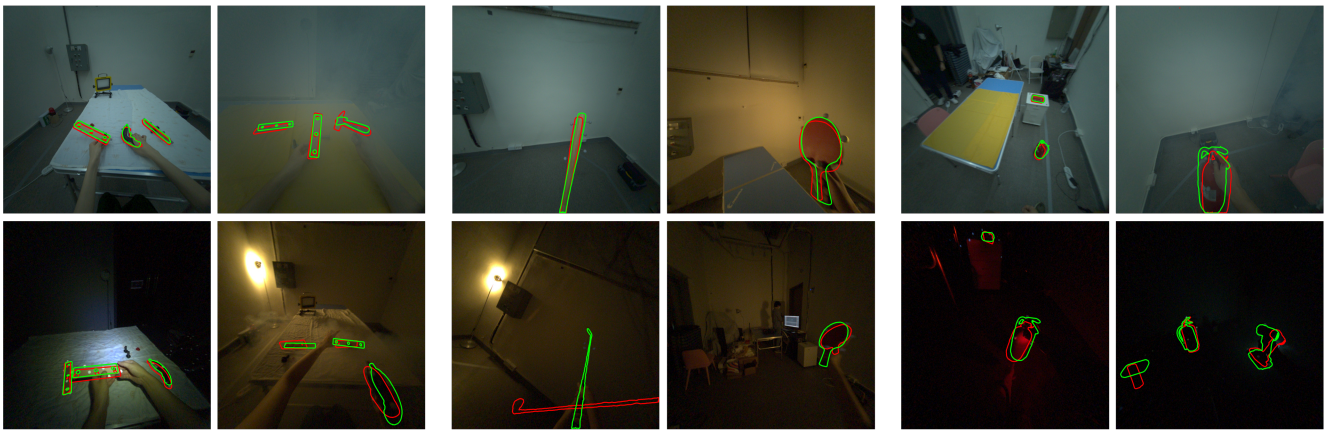
We first establish a baseline with recent state-of-the-art RGB-only zero-shot models, FoundPose [33], GigaPose [31] and PicoPose [27], on the three scenarios of the *EgoXtreme* dataset. To rigorously evaluate the pose estimation performance decoupled from detection errors, we utilized GT bounding boxes. For evaluation metrics, we employ the ADD(-S) [2, 43] recall, along with the standard BOP metrics [32]: Maximum Symmetry-Aware Surface Distance (MSSD) and Maximum Symmetry-Aware Projection Distance (MSPD). End-to-end evaluation results using CNOS [30] detections are provided in Appendix C1.

As shown in Table 3, the results demonstrate a significant performance degradation in extreme lighting and smoke conditions compared to standard light without smoke conditions. Specifically, PicoPose showed a $31.6\%p$ performance degradation in the emergency scenario (measured $@0.3d$) when comparing the extreme lighting condition to the standard lighting condition. Furthermore, in the industrial maintenance scenario, GigaPose's performance degraded by $9.91\%p$ with the addition of smoke. The sports scenario characterized by severe object cropping and motion blur proved to be the most challenging with models failing to achieve meaningful recall overall. FoundPose exhibited the lowest recall, which we attribute to the brittleness of its direct sparse feature matching approach. Feature extraction and matching under egocentric cropping and motion blur frequently failed to secure the minimum number of correspondences required for PnP-RANSAC [9, 24], resulting in a high rate of failure where no 6D pose output was generated. These baselines show that current state-of-the-art models are highly vulnerable to real-world challenges, particularly motion blur and extreme lighting that are not captured in existing datasets. It highlights the necessity of the *EgoXtreme* dataset in filling the gap caused by these real-world challenges. Figure 5 visualizes pose estimation results in standard and extreme cases, respectively.

4.2. Object pose estimation with pre-processing

We investigate whether image restoration techniques can mitigate the performance drop under adverse conditions such as motion blur, smoke, and dynamic lighting. To this end, we apply representative preprocessing methods for deblurring [3], dehazing [36], and low-light enhancement [8] directly to the images before running the baseline pose estimator.

The recall values under full lighting and smoke conditions are detailed in Table 4. The result of image restoration techniques ultimately fail to help, and often hurt performance. Applying a single preprocessing method generally did not increase performance, and combining two methods resulted in an even greater performance drop, decreasing recall by approximately $5\%p$ (measured $@0.3d$) threshold in



(a) (b) (c)

Figure 5. **Example 6D Pose estimation results on baseline models.** The red line is prediction and green is GT. (a), (b), and (c) are the industry maintenance, sports, and emergency rescue scenarios, respectively. The top row indicates standard light condition, and the bottom row indicates extreme light condition.

Table 4. **6D object pose estimation with pre-processing for PicoPose**.

Scenario	Deblur	Dehaze	Light enhance	PicoPose		
				0.1d	0.2d	0.3d
Sports				2.81	8.80	23.02
	✔			2.61	8.52	22.24
			✔	2.87	9.05	22.05
	✔		✔	2.58	8.22	20.99
Maintenance				28.22	48.51	63.32
	✔			26.28	43.70	57.55
		✔		24.04	42.75	57.71
			✔	24.90	43.38	58.08
	✔		✔	23.39	39.73	53.28
Emergency				15.12	42.69	51.70
	✔			14.13	37.74	45.51
		✔		4.74	22.49	37.74
			✔	13.78	38.51	46.75
	✔		✔	13.72	35.58	43.65

the maintenance scenario and $8\%p$ in the emergency scenario.

Specifically, when light enhancement faces extreme conditions such as high-contrast highlights or extremely low light, these methods introduced significant noise making prediction more difficult. The current deblurring and dehazing methods also showed no meaningful positive effect. In particular, dehazing method has severely failed in the emergency scenario yielding an exceptionally low recall of only $4.74\%p$ (measured @$0.1d$). The current dehazing method introduces significant noise artifacts under non-uniform smoky conditions. This explains why our emergency scenario characterized by partial smoke was not properly dehazed, making subsequent pose estimation extremely difficult. Although human observers may perceive an improvement in the preprocessed images of some specific scenarios, the model performance decreased as illustrated in Figure 6.

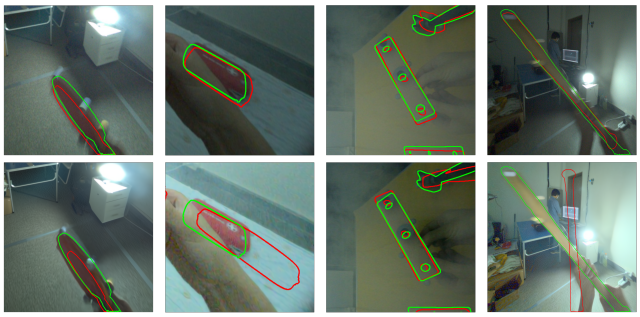


Figure 6. **Example 6D Pose estimation results with preprocessing.** The top row shows the original, non-preprocessed images. The bottom row displays the corresponding images after applying specific preprocessing: deblurring (left), light enhancement (middle), and dehazing (right).

This result demonstrates that existing preprocessing methods are insufficient for 6D object pose estimation on our dataset and its extreme conditions. This failure underscores that the challenges posed by *EgoXtreme* cannot be resolved through preprocessing alone and require deeper investigation using our dataset and benchmark.

4.3. Object pose tracking

For object pose tracking, we compare the tracking-based approach against per-frame baseline (Sec. 4.1). We focus on the highly dynamic sports scenarios. We evaluated three distinct temporal strategies: (1) **Direct temporal** (using the full pose of $t-1$ frame as initial input), (2) **Fusion temporal** (combining final rotation pose of $t-1$ frame with coarse translation pose of t frame), and (3) **Confidence-based hybrid temporal** (selectively using the temporal prior based on the current frame's prediction score).

However, the introduction of the tracking strategy did

Table 5. **6D object pose tracking for GigaPose**. Applied to sports normal scenario.

Object	Method	GigaPose		
		0.1d	**0.2d**	**0.3d**
Pingpong	Per-frame	0.53	2.97	17.63
	Direct	0.05	0.35	6.58
	Fusion	0.56	1.57	10.81
	Hybrid	0.49	1.52	16.53
Tennis	Per-frame	6.77	34.71	50.91
	Direct	3.91	14.28	22.77
	Fusion	5.47	30.64	49.56
	Hybrid	6.93	34.34	50.55
Bat	Per-frame	17.95	41.29	60.55
	Direct	4.66	9.84	14.29
	Fusion	13.68	37.10	60.59
	Hybrid	17.66	45.27	64.46
Golf	Per-frame	0.08	1.64	8.36
	Direct	0.64	1.49	3.61
	Fusion	0.45	5.76	13.98
	Hybrid	0.29	4.08	14.35
Hockey	Per-frame	0.46	5.66	18.35
	Direct	0.07	3.18	5.66
	Fusion	1.57	12.31	24.47
	Hybrid	1.08	10.72	26.13

not yield consistent performance gains across all temporal strategies (Table 5), the dynamic nature of the sports scenario clearly revealed the inherent limitations of direct temporal approaches on our dataset. Specifically, the fast object speed in the sports scenario caused large inter-frame displacement. Therefore, the simple direct temporal propagation is inadequate and sometimes adverse to performance, evidenced by a performance degradation of up to $46\%p$ (measured $@0.3d$). Furthermore, the fusion temporal approach was observed to perform worse than the per-frame baseline in some cases (e.g., pingpong and tennis). This failure is due to propagating the previous frame's rotation, which leads to inaccurate initial poses that the current frame measurement struggles to correct. Conversely, the hybrid temporal strategy successfully mitigated these failures, significantly enhancing performance in this high-motion environment. This robustness stems from its ability to selectively ignore unstable rotational priors by utilizing the current measurement's prediction score, thereby maintaining tracking robustness even under severe displacement.

These results strongly suggest that solving pose estimation in extreme dynamic environments requires not merely simple temporal propagation, but a hybrid temporal strategy that uses dynamic quality assessment to selectively apply the temporal prior. This inherent challenge, explicitly highlighted by the comparative analysis on our dynamic scenarios, confirms the utility of *EgoXtreme* as a crucial and necessary benchmark for future research. It serves not only to evaluate conventional per-frame estimation accuracy, but, more importantly, to drive the development of robust and resilient pose tracking algorithms capable of operating reliably under real-world extreme conditions.

5. Conclusion

In this paper, we filled the critical gap between existing 6D object pose benchmarks and the demanding conditions of real-world egocentric vision. We introduced *EgoXtreme*, a new large-scale 6D pose dataset featuring three challenging scenarios specifically designed to test model robustness against severe motion blur, dynamic/low lighting, and visual obstructions like smoke. Our extensive evaluation of state-of-the-art RGB-only models revealed their significant performance degradation on *EgoXtreme*, confirming that existing methods lack robustness to these extreme conditions. Our experiments showed that while image restoration partially enhances visual fidelity, this improvement fails to translate into meaningful gains for 6D object pose accuracy. Moreover, we show that incorporating temporal information through a tracking-based approach improves pose accuracy and stability in the sports scenario, where motions are extremely fast. To enable performance improvements relevant to these demanding real-world scenarios, a proper benchmark is required for an accurate evaluation, and *EgoXtreme* dataset is specifically introduced to serve this vital role.

Limitations and future works. While *EgoXtreme* provides high-fidelity 6D poses for objects, our methodology involves a few trade-offs. Firstly, the reliance on the OptiTrack system for high-accuracy GT confines data collection to specialized indoor environments, limiting the evaluation of true outdoor robustness. Secondly, we note the absence of 3D hand pose annotations. Accurately labeling intricate hand articulations, particularly due to the extreme motion blur and speed prevalent in the sports scenario, remains a significant independent challenge. We believe, however, that addressing these gaps presents valuable directions for future work, especially by generating accurate 3D hand labels through combining our existing motion capture data with advanced parametric hand models.

Our findings suggest that future research on robust egocentric object pose estimation should move beyond single-frame visual feature representation and focus on effective temporal modeling and robustness against extreme conditions to handle real-world scenarios. We believe *EgoXtreme* will serve as a critical benchmark to drive the development of the next generation of 6D pose estimation models.

Acknowledgment

This work was supported in part by an SNSF Postdoc. Mobility Fellowship (P500PT_225450), the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00216821), and the National Research Foundation (NRF) of Korea grant funded by the Korea government (MSIT) (No. RS-2023-00222663). We thank Meta for providing the Project Aria glasses used in this research. We also gratefully acknowledge the following colleagues for valuable discussions and support of our project: Keondo Park, Eunsu Baek, Joopyo Hong, Yoojin Kwon, Subeom Park, Wooseok Lee, Hun Heo, Seojun Heo, Hongjun Suh, Suahn Bae, Dayeon Woo, Yejun Ji, Wonjeong Lee, Dongik Park, and Boyeong Im.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7061–7071, 2025. 2, 3, 4

[2] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 536–551. Springer, 2014. 2, 3, 6

[3] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision – ECCV 2022*, pages 17–33, 2022. 4, 6

[4] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. 2

[5] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3583–3592, 2016. 2, 3

[6] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing mvtec itodd-a dataset for 3d object recognition in industry. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2200–2208, 2017. 2

[7] 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. 2, 4

[8] Daniel Feijoo, Juan C. Benito, Marcos V. Conde, and Alvaro Garcia. Darkir: Robust low-light image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10879–10889, 2025. 4, 6

[9] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 4, 6

[10] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1780–1789, 2020. 4

[11] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 2

[12] Yegyu Han, Taegyoon Yoon, Dayeon Woo, Sojeong Kim, and Hyung-Sin Kim. Senseshift6d: Multimodal rgb-d benchmarking for robust 6d pose estimation across environment and sensor variations. *arXiv preprint arXiv:2507.05751*, 2025. 3

[13] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1956–1963. IEEE, 2009. 4

[14] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012. 2, 3

[15] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017. 2, 3

[16] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018. 2, 3

[17] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960. 5

[18] Agastya Kalra, Guy Stoppi, Dmitrii Marin, Vage Taamazyan, Aarrushi Shandilya, Rishav Agarwal, Anton Boykov, Tze Hao Chong, and Michael Stark. Towards co-evaluation of cameras hdr and algorithms for industrial-grade 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22691–22701, 2024. 3

[19] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[20] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude)

faster and better. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 4

[21] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021. 2, 3, 4

[22] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022. 4, 6

[23] Sohyun Lee, Jaesung Rim, Boseung Jeong, Geonu Kim, Byungju Woo, Haechan Lee, Sunghyun Cho, and Suha Kwak. Human pose estimation in extremely low-light conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7761–7770, 2023. 2

[24] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81(2):155–166, 2009. 4, 6

[25] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27906–27916, 2024. 3

[26] Yongliang Lin, Yongzhi Su, Praveen Nathan, Sandeep Inuganti, Yan Di, Martin Sundermeyer, Fabian Manhardt, Didier Stricker, Jason Rambach, and Yu Zhang. Hipose: Hierarchical binary surface encoding and correspondence pruning for rgb-d 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10148–10158, 2024. 3

[27] Lihua Liu, Jiehong Lin, Zhenxin Liu, and Kui Jia. Picopose: Progressive pixel-to-pixel correspondence learning for novel object pose estimation. *Conference on Robot Learning*, 2025. 3, 4, 6

[28] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2024. 4

[29] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9890–9900, 2020. 4

[30] 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. 6

[31] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9903–9913, 2024. 3, 6

[32] Van Nguyen Nguyen, Stephen Tyree, Andrew Guo, Médéric Fourmy, Anas Gouda, Taeyeop Lee, Sungphill Moon, Hyeontae Son, Lukas Ranftl, Jonathan Tremblay, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, Stan Birchfield, Jiří Matas, Yann Labbé, Martin Sundermeyer, and Tomáš Hodaň. BOP challenge 2024 on model-based and model-free 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 6

[33] Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. In *European Conference on Computer Vision*, pages 163–182. Springer, 2024. 3, 6

[34] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11908–11915, 2020. 4

[35] Jaesung Rim, Geonung Kim, Jungeon Kim, Junyong Lee, Seungyong Lee, and Sunghyun Cho. Realistic blur synthesis for learning image deblurring. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[36] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. 4, 6

[37] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738–6748, 2022. 3

[38] Alykhan Tejani, Danhang Tang, Rigas Kouskouridas, and Tae-Kyun Kim. Latent-class hough forests for 3d object detection and pose estimation. In *European conference on computer vision*, pages 462–477. Springer, 2014. 3

[39] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13081–13088. IEEE, 2022. 2, 3

[40] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4): 376–380, 2002. 5

[41] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021. 3

[42] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 3

[43] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 2, 3, 6

[44] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 4

[45] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022. 4