Monocular Person Localization under Camera Ego-Motion

Yu Zhan[1], Hanjing Ye[1] and Hong Zhang[1*]

***Abstract*— Robust person localization from a moving camera is a fundamental skill for robots to navigate and interact with humans in the open world. However, the diversity of robot platforms and environments poses a significant generalization challenge. Learning-based methods, often trained on datasets with limited camera motion, fail in out-of-distribution (OOD) scenarios involving severe camera ego-motion. To address this, we propose an optimization-based method that models the human with a four-point skeleton to jointly estimate camera attitude and 3D person location. Our approach avoids reliance on large-scale training data and generalizes across different viewpoints and image projections. Real-robot experiments and dataset evaluations show our method outperforms existing approaches, especially in these challenging OOD scenarios. The system is deployed for person-following on an agile quadruped, demonstrating its utility for robust open-world Human-Robot Interaction (HRI).**

I. INTRODUCTION

Enabling robots to navigate the open world and collaborate with humans is a central goal in robotics, with Robot Person Following (RPF) [1] being a key interactive skill. A major challenge for open-world deployment is generalization across diverse environments and hardware embodiments. Specifically, achieving stable localization under significant viewpoint changes is critical for robust interaction. This is particularly challenging for agile robots like quadrupeds traversing rough terrains [2], [3]. As shown in Fig. 1, their dynamic movements induce severe camera ego-motion, which poses a significant out-of-distribution (OOD) problem compared to the well-studied, leveled-ground scenarios common in autonomous driving [4].

To mitigate ego-motion, using extra sensors like UWB [2], LiDAR [5]–[9], or RGB-D cameras [5] can directly provide the person's depth, thus bypassing the viewpoint dependency. However, to achieve robust monocular localization under such viewpoint shifts, it is necessary to jointly estimate both the camera's pose and the person's 3D location [10]. Some methods attempt to compensate for ego-motion using state estimation from an IMU [8], [11] or odometry [5], [6], [12], [13]. Unfortunately, state estimation for highly dynamic robots often suffers from significant drift, leading to accumulating localization errors [14], [15]. Other prior works relying on strong geometric or data-driven priors also struggle in these dynamic scenarios.

Monocular human perception is also widely studied in computer vision [16], [17]. However, learning-based models for depth estimation [18], [19], 3D detection [20], [21], or mesh recovery [22] often fail to generalize. Trained on datasets with limited camera perspectives, they perform poorly when faced with the large roll and pitch motions of a quadruped—a classic OOD failure. Furthermore, methods using complex parametric models like SMPL [23] can be sensitive to camera intrinsic parameters, limiting their hardware generalizability, while recent transformer-based architectures, despite their strong performance, are often too computationally intensive for real-time robotic applications [22].

Fig. 1. A scenario of a quadruped robot following a person through a rugged lawn. The robot view is from an onboard panoramic camera (see Sec. IV-C). The robot's dynamic motion induces severe camera ego-motion and vibration, which bring challenges for person localization.

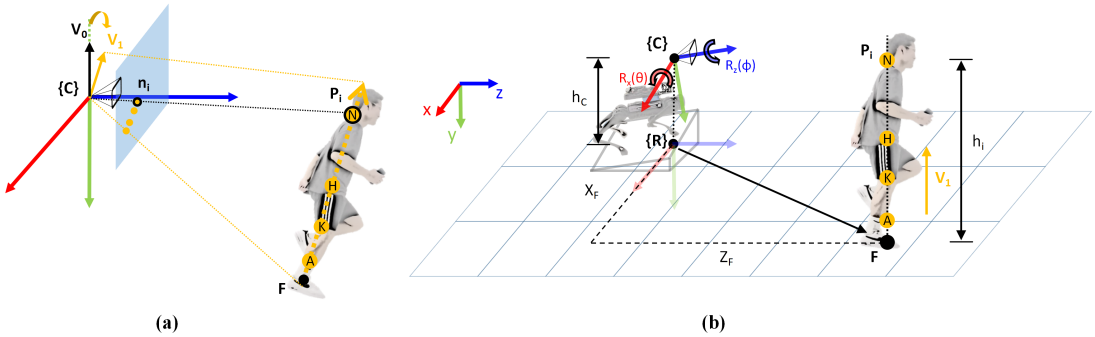
To address these generalization issues, we propose a real-time, optimization-based method that estimates 3D person location from a single image, avoiding reliance on potentially error-prone odometry. By representing the human as a four-point model, our method jointly optimizes for the camera's attitude and the person's 3D location. This model-based approach is inherently more robust to OOD scenarios than learning-based counterparts and, through a normalization step, is agnostic to specific camera intrinsics, enhancing its generalizability across different robot platforms. We demonstrate its effectiveness on a Unitree Go1 quadruped [2], achieving stable localization on rough terrain. Our code and dataset are available at https://medlartea.github.io/rpf-quadruped/.

II. RELATED WORK

A. *Geometric-Model-based Localization*

Geometric methods estimate 3D position from 2D joints by assuming an upright human on a ground plane [24]–[26]. While recent works have relaxed some constraints for robotic applications [27]–[29], their core assumption of a fixed camera pose remains. Thus, these plane geometry-based methods are suitable for fixed surveillance cameras or indoor wheeled robots on flat ground, but are ill-suited for

[1]Yu Zhan, Hanjing Ye and Hong Zhang are with Shenzhen Key Laboratory of Robotics and Computer Vision, Southern University of Science and Technology (SUSTech), and the Department of Electronic and Electrical Engineering, SUSTech. *corresponding author (hzhang@sustech.edu.cn).



(a) (b)

Fig. 2. The geometry of our observation model. (a) In the raw camera-centric view, the person appears tilted due to the robot's ego-motion. (b) Our model assumes an upright person, representing the ego-motion as a corresponding tilt of the camera.

agile robots with large-angle ego-motion, such as quadruped, humanoid, or outdoor off-road robots.

B. *Optimization-based Pose Estimation*

Optimization-based methods recover pose by fitting a model to semantic keypoints [30], [31]. For humans, state-of-the-art systems like BodySLAM++ [10] achieve high accuracy by coupling a human model with stereo VIO. However, this reliance on extra sensors makes it unsuitable for low-cost, monocular platforms, and its robustness under the severe ego-motion considered here is untested.

C. *Learning-based Pose Regression*

Learning-based methods directly regress 3D location from 2D information. Approaches that regress location from 2D joints [21], [32], [33] generalize poorly, as models trained on autonomous driving data [34] fail when presented with out-of-distribution robot viewpoints [4]. Full human mesh recovery models [16], [17], [22] often have high computational costs, suffer from depth ambiguity, and have restrictive camera assumptions. Moreover, many current 3D Human Pose Estimation (HPE) methods focus on root-relative accuracy. Consequently, few methods can simultaneously provide the absolute accuracy, real-time performance, and generalization capability required for robust person localization.

III. METHODOLOGY

We propose an optimization-based person localization method robust to camera ego-motion. By fitting a four-point human skeleton model ($\mathcal{P}_{all} = \{P_{neck}, P_{hip}, P_{knee}, P_{ankle}\}$) to 2D image observations, our method simultaneously estimates the camera's 2D attitude (roll, pitch) and the person's 3D location. We then integrate this method into an RPF system on a quadruped robot.

A. *Human Model and Observations*

Our method begins with 2D observations. We use YOLOX [35] for person detection and AlphaPose [36] for 2D joint estimation, which is robust to occlusion and distortion [28]. From the left/right joints, we take the median to define our four keypoints in $\mathcal{P}_{all}$. Crucially, all 2D points are then back-projected to a normalized image plane. This normalization makes our method independent of specific camera intrinsics, a key feature for generalizing across different robot hardware and camera embodiments. The projected lengths and ratios of the segments on this plane encode the person's distance and the camera's viewing angle (Fig. 2a).

B. *Parameterization and Constraints*

The person's pose relative to the camera is defined by 5-DoF (3D translation, 2D rotation), as we ignore the body's yaw which can be recovered separately [37]. We represent the person's position by their footprint $\mathbf{F} \in \mathbb{R}^3$ in the camera frame $\{\mathbf{C}\}$. Their structure is defined by a set of known heights $\mathcal{H}_{all} = \{h_{neck}, ..., h_{ankle}\}$ for the keypoints in $\mathcal{P}_{all}$.

Our key assumption is that the person remains upright. We therefore model the camera's ego-motion as a rotation of the camera frame itself, relative to a virtual ground plane on which the person stands (Fig. 2b). This rotation from the camera frame $\{\mathbf{C}\}$ to the robot's frame $\{\mathbf{R}\}$ is parameterized by roll and pitch Euler angles $\{\theta, \phi\}$:

$$\mathbf{R} = \mathbf{R_z}(\phi)\mathbf{R_x}(\theta), \tag{1}$$

where $\mathbf{R_z}(\phi)$ and $\mathbf{R_x}(\theta)$ are elementary rotation matrices. Our state vector $\mathbf{s}$ to be estimated thus consists of the person's planar location (X_F, Z_F), the camera height h_C, and the camera attitude:

$$\mathbf{s} = \{X_F, Z_F, h_C, \theta, \phi\} \tag{2}$$

In the robot frame $\{\mathbf{R}\}$, the vector from the camera center $\mathbf{C}$ to any point $\mathbf{P_i} \in \mathcal{P}_{all}$ is:

$$\overrightarrow{\mathbf{CP_i}} = (X_F, h_C - h_i, Z_F)^T, h_i \in \mathcal{H}_{all} \tag{3}$$

This vector is then transformed into the camera frame for reprojection:

$$\mathbf{P_i^C} = \mathbf{R^{-1}} \cdot \overrightarrow{\mathbf{CP_i}}, \tag{4}$$

where $\mathbf{P_i^C}$ are the coordinates of $\mathbf{P_i}$ in $\{\mathbf{C}\}$.

C. *Optimization Details*

To account for body articulation, we assign lower weights to mobile points (P_{knee}, P_{ankle}) and higher weights to stable points (P_{neck}, P_{hip}). We then minimize the weighted reprojection error f:

$$f(X_F, Z_F, h_C, \theta, \phi) = \sum_{i=1}^{n} w_i \left\| \mathbf{n_i} - \pi(\mathbf{P_i^C}) \right\|^2 \tag{5}$$

where $\pi(\cdot)$ is the camera projection function. We optimize the state vector $\mathbf{s}$ by partitioning it into translation $\mathbf{t} =$

$\{X_F, Z_F, h_C\}$ and rotation $\mathbf{r} = \{\theta, \phi\}$ and updating them alternately [30].

We solve this bounded nonlinear least-squares problem using the Dogbox method [38], [39] with a Cauchy cost function [40] for robustness.

D. Implementation in RPF Framework

Our method is integrated into a standard RPF pipeline (Fig. 3) [1], [27]–[29], where our normalization and optimization modules are the key components for localization. The localization process involves two phases: an offline initialization and an online tracking phase.

First, during the offline initialization with the robot's known static pose, we solve for the target's joint heights $\mathcal{H}_{all}$ and initial position. This is a linear least squares problem solved via SVD [41] by minimizing the reprojection error:

$$\{X_F^*, Z_F^*, \mathcal{H}_{all}^*\} = \arg \min_{X_F, Z_F, \mathcal{H}_{all}} \sum_{i=1}^{n} w_i \left\| \mathbf{n_i} - \pi(\mathbf{P_i^C}) \right\|^2 \tag{6}$$

Second, in the online person-following phase, the calibrated $\mathcal{H}_{all}$ is used to estimate the person's location and camera attitude in real-time as described in Sec. III-C. The resulting location (X_F, Z_F) is then used by downstream modules for data association, trajectory smoothing, person re-identification [42], and robot control.

IV. EXPERIMENTS

We conduct a series of experiments to evaluate our method's accuracy, efficiency, and generalization capability in open-world scenarios.

A. Baselines

We compare our method against geometric and deep-learning baselines from Sec. II-A and Sec. II-C:

Geo-model-based:

- **Koide's Method** [27]: Locates person via neck point, assuming a fixed camera.
- **Ye's Method** [28]: Extends [27] using four points to handle occlusion, but still assumes a fixed camera.

Deep-learning-based:

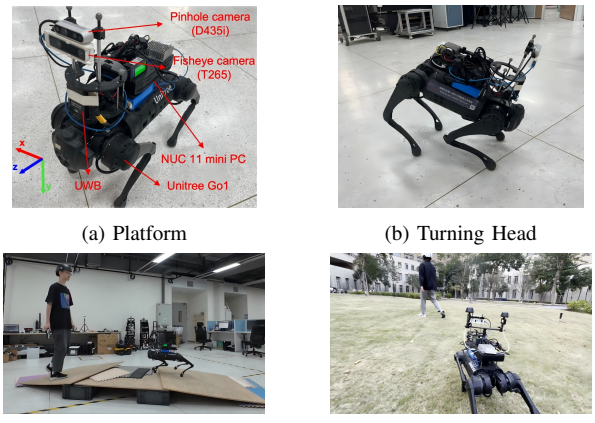
- **MonoLoco++** [21]: Regresses 3D location from 2D joints, trained on KITTI [34].
- **Depth Anything** [18]: Estimates a relative depth map. We use average joint depth for distance and the ground truth of the 1st frame for metric depth.
- **Multi-HMR** [22]: Recovers human mesh and estimates absolute distance between the pelvis and the camera.

B. Datasets

We evaluate on three datasets, positioning them as benchmarks for performance in different environments. **KITTI** [34] represents a structured, "in-distribution" scenario, typical of autonomous driving, where learning-based 3D-detection methods are trained. **FieldSAFE** [43], featuring a tractor in a field, presents a moderately challenging outdoor environment. Our custom **RPF-Quadruped** dataset, recorded on a Unitree Go1 [2] (Fig. 4a), is designed to be a benchmark for "out-of-distribution" (OOD) open-world challenges. It contains three scenarios (Fig. 4b-4d) with ground truth from a motion capture system or UWB. As shown in Table I, the significant variance in camera pitch and roll in our dataset quantitatively demonstrates a distributional shift from standard benchmarks, posing a stringent test for viewpoint generalization.

Dataset	KITTI [34]	FieldSAFE [43]	RPF-Quadruped
Distance from Camera (m)	18.44 ± 11.20	7.50 ± 1.50	3.50 ± 3.00
Camera Height (m)	2.31 ± 0.29	4.50 ± 0.09	0.50 ± 0.15
Camera Pitch (deg)	/	16.01 ± 5.27	0.5 ± 15.46
Camera Roll (deg)	/	0.32 ± 4.18	0.8 ± 10.30

TABLE I. Statistical comparison of **mean** and **standard deviation** of key parameters, highlighting the distributional shift in our RPF-Quadruped dataset.



(a) Platform (b) Turning Head

(c) Indoor Slope (d) Rugged Lawn

Fig. 4. (a) Our quadruped robot platform. (b-d) Scenarios from our RPF-Quadruped dataset.

C. Platform and Implementation Details

Our platform is a Unitree Go1 quadruped [2] (Fig. 4a) with an Intel NUC (i7/RTX 2060), using pin-hole, fisheye, and panoramic cameras. A UWB sensor provides ground-truth distance. The Go1's small size and high step frequency result in more severe ego-motion than platforms in prior RPF work [5], [7], [8]. All methods were evaluated on the robot's NUC, except for Multi-HMR [22], which ran offline on a desktop PC (RTX 3070). 2D joint detection for relevant methods was standardized as per Sec. III-A and accelerated with TensorRT.

D. Evaluation and Results

We evaluate localization accuracy and runtime. Accuracy is measured by **Average Location/Distance Error (ALE/ADE)** [21], [28]. For sequences with continuous motion, we also report the **Variance of Location/Distance Error (VLE/VDE)** to assess stability.

As shown in Table II, our method achieves the lowest error and variance on our dataset and FieldSAFE. MonoLoco++ [21] performs best on KITTI, its training domain. In the challenging *Rugged Lawn* scenario, our method is visibly more accurate and stable, as shown by the error distribution (Fig. 5) and time-series distance plot (Fig. 6). Fig. 6(a) shows

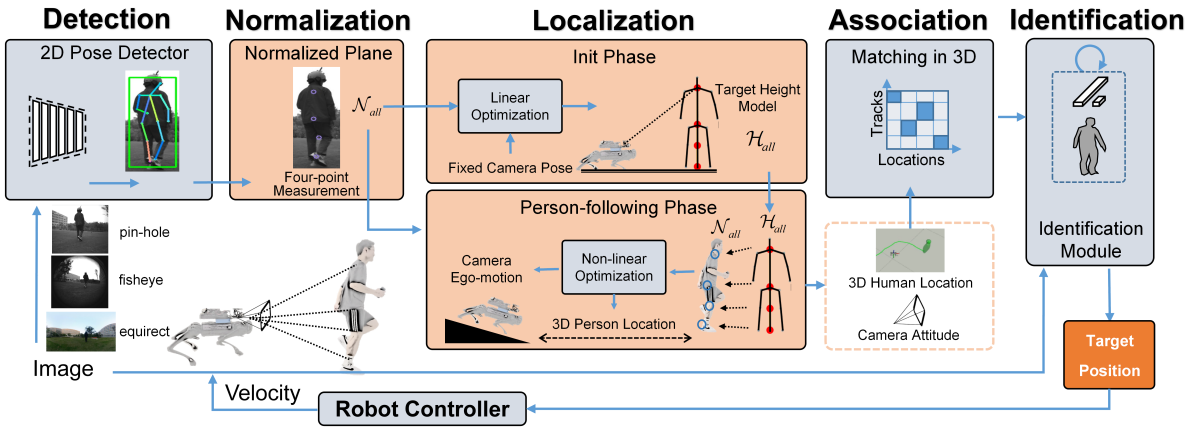


Fig. 3. Our proposed framework for monocular Robot Person Following (RPF). The modules highlighted in orange represent our key contributions: (1) a normalization step for camera-agnostic processing, and (2) a subsequent optimization-based person localization method.

TABLE II. Comparison of localization accuracy. We evaluate our method against several baselines and present an ablation study. Metrics include: Average Location/Distance Error (**ALE/ADE**) in meters (m), and their corresponding variances (**VLE/VDE**) in m^2.

Methods \ Scenarios	Turning Head *ALE* ↓	Indoor Slope *ALE* ↓	Rugged Lawn *ADE / VDE* ↓	FieldSAFE [43] *ALE / VLE* ↓	KITTI [34] *ALE* ↓
Koide's Method [27]	0.396	0.289	0.3 / 0.3	1.924 / 5.012	1.451
Ye's Method [28]	0.294	0.261	0.3 / 0.3	1.856 / 3.952	1.420
MonoLoco++ [21]	0.820	0.510	0.6 / 0.2	4.152 / 4.705	**0.940**
Depth Anything [18]	0.571	0.523	0.5 / 0.6	1.528 / 1.022	2.963
Multi-HMR [22]	0.493	0.254	0.4 / 0.3	3.066 / 0.424	1.520
Ours	**0.178**	**0.101**	**0.1 / 0.0**	**1.287 / 0.356**	1.220
Ours w/o neck	0.238	0.196	0.2 / 0.1	1.324 / 0.865	1.320
Ours w/o ankle	0.204	0.141	0.1 / 0.0	1.308 / 0.401	1.275
Ours on fisheye images	0.182	0.119	0.1 / 0.0	/	/

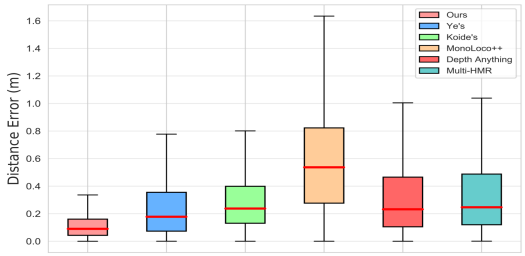


Fig. 5. A box plot illustrating the distance error of our method compared to baselines in *Rugged Lawn* scenario.

that learning-based methods generalize poorly to our scenarios, while Fig. 6(b) shows that geo-model-based methods produce large errors during ego-motion. Table III confirms our method's real-time performance.

V. CONCLUSIONS

In this paper, we presented a real-time (40 FPS), optimization-based method for monocular person localization, addressing the critical challenge of viewpoint generalization under severe camera ego-motion. Our approach, which jointly estimates camera attitude and person location using a four-point human model, was successfully demonstrated in a Robot Person Following (RPF) system on an agile quadruped. Although built upon a traditional optimization framework, our method's independence from large-scale training data and camera intrinsics makes it easy to deploy across diverse platforms. We believe that achieving such robust spatial intelligence is a prerequisite for stable decision-making and navigation in the open world.

We hope our simple yet effective method can serve as a baseline to research navigation, following, and HRI in unstructured terrains. Furthermore, we argue that scenarios with large-angle camera ego-motion are an essential testbed for a model's viewpoint generalization. Our contributed dataset, featuring low-angle perspectives and significant ego-motion, provides a resource for benchmarking these capabilities. Future work will focus on handling more postures with expressive human models and validating our approach on large-scale egocentric datasets such as TPT-bench [44].

TABLE III. Comparison of per-frame average runtime. The preprocessing time accounts for 2D human joint detection. *Runtime for Multi-HMR was measured on a different PC (see Sec. IV-C).

Method	Preprocessing (s)	Estimation (s)	Total (s)
Koide's Method [27]	0.02	**0.0006**	**0.0206**
Ye's Method [28]	0.02	0.0008	0.0208
MonoLoco++ [21]	0.02	0.09	0.11
Depth Anything [18]	/	0.23	0.23
Multi-HMR [22]*	/	1.24	1.24
Ours	0.02	0.005	0.025

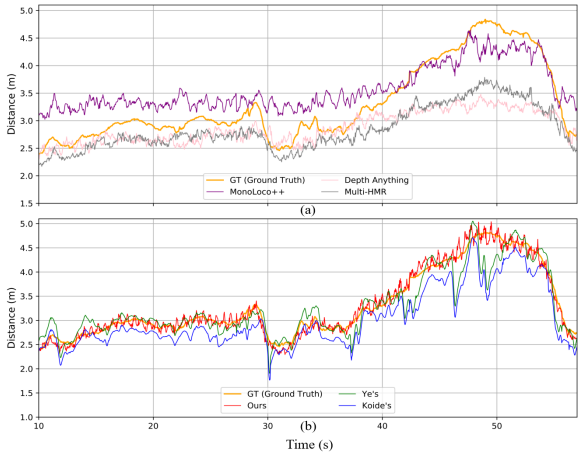


Fig. 6. Comparison of estimated distance over time on a sequence from the *Rugged Lawn* dataset (10s–57s). The plot shows the output of (a) deep-learning-based and (b) geo-model-based baselines.

REFERENCES

[1] M. J. Islam, J. Hong, and J. Sattar, "Person-following by autonomous robots: A categorical overview," *The International Journal of Robotics Research*, vol. 38, no. 14, pp. 1581–1618, 2019.

[2] "Unitree go1," https://www.unitree.com/cn/go1.

[3] "Alphard club-booster-v2," https://alphardgolf.com.

[4] X. Ma, W. Ouyang, A. Simonelli, and E. Ricci, "3d object detection from images for autonomous driving: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3537–3556, 2024.

[5] Z. Zhang, J. Yan, X. Kong, G. Zhai, and Y. Liu, "Efficient motion planning based on kinodynamic model for quadruped robots following persons in confined spaces," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 4, pp. 1997–2006, 2021.

[6] B. Mishra, D. Calvert, B. Ortolano, M. Asselmeier, L. Fina, S. McCrory, H. E. Sevil, and R. Griffin, "Perception engine using a multi-sensor head to enable high-level humanoid robot behaviors," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 9251–9257.

[7] S. Xin, Z. Zhang, M. Wang, X. Hou, Y. Guo, X. Kang, L. Liu, and Y. Liu, "Multi-modal 3d human tracking for robots in complex environment with siamese point-video transformer," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 337–344.

[8] K. Cho, S. H. Baeg, and S. Park, "3d pose and target position estimation for a quadruped walking robot," in *2013 10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, 2013, pp. 466–467.

[9] A. Leigh, J. Pineau, N. Olmedo, and H. Zhang, "Person tracking and following with 2d laser scanners," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 726–733.

[10] D. F. Henning, C. Choi, S. Schaefer, and S. Leutenegger, "Bodyslam++: Fast and tightly-coupled visual-inertial camera and human motion tracking," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 3781–3788.

[11] J. Brookshire, "Person following using histograms of oriented gradients," *I. J. Social Robotics*, vol. 2, pp. 137–146, 06 2010.

[12] A. Roychoudhury, S. Khorshidi, S. Agrawal, and M. Bennewitz, "Perception for humanoid robots," *Current Robotics Reports*, pp. 1–14, 2023.

[13] K. Aso, D.-H. Hwang, and H. Koike, "Portable 3d human pose estimation for human-human interaction using a chest-mounted fisheye camera," in *Proceedings of the Augmented Humans International Conference 2021*, 2021, pp. 116–120.

[14] F. Allione, J. D. Gamba, A. E. Gkikakis, R. Featherstone, and D. Caldwell, "Effects of repetitive low-acceleration impacts on attitude estimation with micro-electromechanical inertial measurement units," *Frontiers in Robotics and AI*, vol. 10, 2023.

[15] S. Yang, Z. Zhang, Z. Fu, and Z. Manchester, "Cerberus: Low-drift visual-inertial-leg odometry for agile locomotion," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4193–4199.

[16] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–37, 2023.

[17] W. Liu, Q. Bao, Y. Sun, and T. Mei, "Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective," *ACM Comput. Surv.*, vol. 55, no. 4, Nov. 2022.

[18] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 10 371–10 381.

[19] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "UniDepth: Universal monocular metric depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[20] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 132–10 141, 2019.

[21] L. Bertoni, S. Kreiss, and A. Alahi, "Perceiving humans: from monocular 3d localization to social distancing," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[22] F. Baradel*, M. Armando, S. Galaaoui, R. Brégier, P. Weinzaepfel, G. Rogez, and T. Lucas*, "Multi-hmr: Multi-person whole-body human mesh recovery in a single shot," in *ECCV*, 2024.

[23] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," vol. 9909, 10 2016, pp. 561–578.

[24] X. Fei, H. Wang, L. L. Cheong, X. Zeng, M. Wang, and J. Tighe, "Single view physical distance estimation using human pose," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 386–12 396.

[25] M. Aghaei, M. Bustreo, Y. Wang, G. Bailo, P. Morerio, and A. Del Bue, "Single image human proxemics estimation for visual social distancing," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2784–2794.

[26] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1577–1591, 2013.

[27] K. Koide, J. Miura, and E. Menegatti, "Monocular person tracking and identification with on-line deep feature selection for person following robots," *Robotics and Autonomous Systems*, vol. 124, p. 103348, 2020.

[28] H. Ye, J. Zhao, Y. Pan, W. Cherr, L. He, and H. Zhang, "Robot person following under partial occlusion," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 7591–7597.

[29] A. Bacchin, F. Berno, E. Menegatti, and A. Pretto, "People tracking in panoramic video for guiding robots," in *Intelligent Autonomous Systems 17*, I. Petrovic, E. Menegatti, and I. Marković, Eds. Cham: Springer Nature Switzerland, 2023, pp. 407–424.

[30] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis, "6-dof object pose from semantic keypoints," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 2011–2018.

[31] M. Pavliv, F. Schiano, C. Reardon, D. Floreano, and G. Loianno, "Tracking and relative localization of drone swarms with a vision-based headset," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1455–1462, 2021.

[32] L. Bertoni, S. Kreiss, and A. Alahi, "Monoloco: Monocular 3d pedestrian localization and uncertainty estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[33] D. T. Tran, D. D. Tran, M. A. Nguyen, Q. Van Pham, N. Shimada, J.-H. Lee, and A. Q. Nguyen, "Monois3dloc: Simulation to reality learning based monocular instance segmentation to 3d objects localization from aerial view," *IEEE Access*, vol. 11, pp. 64 170–64 184, 2023.

[34] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[35] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," 07 2021.

[36] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[37] J. Zhao, H. Ye, Y. Zhan, H. Luan, and H. Zhang, "Human orientation estimation under partial observation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 11 544–11 551.

[38] C. Voglis and I. E. Lagaris, "A rectangular trust region dogleg approach for unconstrained and bound constrained nonlinear optimization," in *Proceedings of the WSEAS International Conference on Applied Mathematics (WSEAS'04)*, 2004.

[39] J. Nocedal and S. Wright, *Numerical Optimization*, 01 2006.

[40] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment — a modern synthesis," in *Vision Algorithms: Theory and Practice*, B. Triggs, A. Zisserman, and R. Szeliski, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 298–372.

[41] G. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, vol. 2, no. 2, pp. 205–224, 1965.

[42] H. Ye, J. Zhao, Y. Zhan, W. Chen, L. He, and H. Zhang, "Person re-identification for robot person following with online continual learning," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9151–9158, 2024.

[43] M. F. Kragh, P. Christiansen, M. S. Laursen, M. Larsen, K. A. Steen, O. Green, H. Karstoft, and R. N. Jørgensen, "Fieldsafe: Dataset for obstacle detection in agriculture," *Sensors*, vol. 17, no. 11, 2017.

[44] H. Ye, Y. Zhan, W. Situ, G. Chen, J. Yu, Z. Zhao, K. Cai, A. Ajoudani, and H. Zhang, "Tpt-bench: A large-scale, long-term and robot-egocentric dataset for benchmarking target person tracking," 2025. [Online]. Available: https://arxiv.org/abs/2505.07446