Monocular Person Localization under Camera Ego-Motion

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

***Abstract*— Localizing a person from a moving monocular camera is critical for Human-Robot Interaction (HRI). To estimate the 3D human position from a 2D image, existing methods either depend on the geometric assumption of a fixed camera or use a position regression model trained on datasets containing little camera ego-motion. These methods are vulnerable to severe camera ego-motion, resulting in inaccurate person localization. We consider person localization as a part of a pose estimation problem. By representing a human with a four-point model, our method jointly estimates the 2D camera attitude and the person's 3D location through optimization. Evaluations on both public datasets and real robot experiments demonstrate our method outperforms baselines in person localization accuracy. Our method is further implemented into a person-following system and deployed on an agile quadruped robot.**

I. INTRODUCTION

Person localization is critical for robotic applications like Robot Person Following (RPF) [1] and crowd analysis [2]. While many monocular localization methods exist [3]–[5], they often rely on the restrictive assumption of a fixed camera height and attitude [1]. This assumption fails in real-world scenarios, such as for quadruped robots traversing rough terrains [6], [7], where severe ego-motion challenges stable person tracking (Fig. 1).

To mitigate ego-motion, common solutions involve extra sensors like UWB [6], LiDAR [8]–[12], or RGB-D cameras [8], but monocular cameras are desirable for their low cost. Another approach is compensating for ego-motion using state estimation from IMU [11], [13] or odometry [8], [9], [14], [15]. However, state estimation for highly dynamic robots like quadrupeds is prone to significant drift, leading to accumulating errors in the person's estimated location [16], [17].

Monocular person localization is also widely studied in computer vision [18], [19] and autonomous driving [20]. Deep learning methods for depth estimation [21], [22], 3D human pose [23], [24], or mesh recovery [25] often lack generalizability to new scenarios and camera motions. Meanwhile, methods based on complex non-rigid models like SMPL [26] are too computationally intensive for real-time applications.

To address these issues, we propose a real-time, optimization-based method that estimates 3D person location from a single-frame observation, avoiding reliance on potentially error-prone odometry. We represent the human body as a four-point model and simultaneously optimize for the camera attitude and the person's 3D location based on 2D-3D correspondences. Our method achieves state-of-the-art accuracy on public and custom datasets. We demonstrate its effectiveness in a real-world RPF system on a Unitree Go1 quadruped [6] traversing rough terrain, achieving accurate and stable localization despite severe ego-motion. Our code and dataset are available at `https://medlartea.github.io/rpf-quadruped/`.

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.

II. RELATED WORK

Estimating a person's 3D location from a monocular image is an ill-posed problem. Existing methods rely on geometric constraints (Sec. II-A), optimization (Sec. II-B), or learning-based regression (Sec. II-C). We focus on methods suitable for real-time robotic systems.

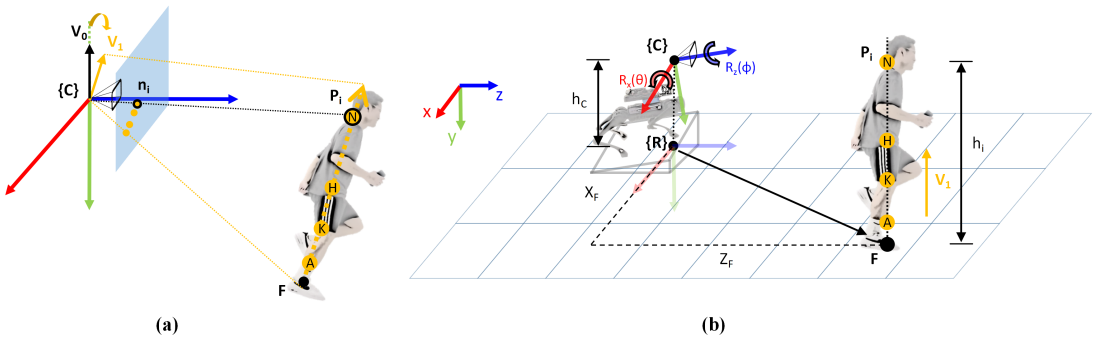
A. *Person Localization with Geometric Model*

These methods model the human as a line segment, assuming an upright posture on a ground plane to solve for 3D position from 2D joint detections. Early works for surveillance or moving cameras relied on strong assumptions like known human height [27], fixed camera pose [28], or a fixed horizon [29]. More recent methods designed for robotics [3], including extensions for occlusion [4] and omnidirectional cameras [5], still fundamentally assume a fixed camera pose relative to the ground. This assumption is violated on agile robots like quadrupeds that exhibit significant ego-motion, leading to inaccurate localization.

B. *Pose Optimization from Semantic Keypoints*

Other approaches recover 6-DoF poses using semantic keypoints. Pavlakos *et al.* [30] pioneered optimizing 6-DoF poses for rigid objects using CNN-predicted keypoints and a deformable model, a concept later applied to drone tracking

[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.

[31], but this is less effective for non-rigid humans. BodySLAM++ [32] tightly couples human pose estimation (using the SMPL model) with stereo Visual Inertial Odometry (VIO) in a factor graph. While achieving real-time, metric-scale accuracy, it depends on multi-image input and VIO, making it costly for resource-constrained robots, and its performance under severe ego-motion is untested.

C. *Learning-based Pose Regression*

Data-driven methods regress 3D location from 2D information.

Regressing 3D Location from 2D Joints. Methods like MonoLoco++ [24], [33] regress probabilistic 3D locations from 2D keypoints [24] or semantic patches [34]. However, their reliance on datasets like KITTI [35], which features limited camera perspectives, leads to poor generalization on robots with large roll/pitch ego-motion [20].

Learning-based Human Pose Estimation. End-to-end frameworks reconstruct a human's full pose, shape, and location from a single image [18], [19]. However, they often prioritize root-relative accuracy over absolute camera-relative pose, suffering from depth ambiguity in "in-the-wild" settings [18], [19]. High computational costs and restrictive camera model assumptions [25] further limit their practicality on mobile robots. Thus, few methods simultaneously achieve absolute accuracy, real-time performance, and generalization across domains.

III. METHODOLOGY

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

A. *Human Model and Observations*

We model an upright human as a rigid body of four collinear points, $\mathcal{P}_{all}$. The projected lengths and relative ratios of the segments on the image plane (Fig. 2a) encode the person's distance and the camera's viewing angle. We obtain the 2D projections of $\mathcal{P}_{all}$ using YOLOX [36] for person detection and AlphaPose [37] for 2D joint estimation, which is robust to occlusion and distortion [4]. The four keypoints are the median of corresponding left/right joints. All 2D points are back-projected to a normalized image plane, yielding points $\mathcal{N}_{all}$. This normalization makes our method independent of specific camera intrinsics, enhancing generalizability.

B. *Parameterization and Constraints*

The pose of our human model relative to the camera has 5-DoF (3D translation, 2D rotation). We ignore the body's yaw, which can be recovered separately [38]. We assume a human's footprint $\mathbf{F} \in \mathbb{R}^3$ in the camera frame $\{\mathbf{C}\}$ represents their position, and the heights of the points in $\mathcal{P}_{all}$ relative to $\mathbf{F}$ are known as $\mathcal{H}_{all} = \{h_{neck}, h_{hip}, h_{knee}, h_{ankle}\}$. In frame $\{\mathbf{C}\}$, the human's central axis is a unit vector $\mathbf{V_1}$ (Fig. 2a). Due to ego-motion, $\mathbf{V_1}$ is rotated from the camera's y-axis $\mathbf{V_0} = (0, -1, 0)^T$.

We assume the human is upright, moving on a virtual plane perpendicular to $\mathbf{V_1}$ (Fig. 2b). The robot frame $\{\mathbf{R}\}$ also lies on this plane. The rotation from $\{\mathbf{C}\}$ to $\{\mathbf{R}\}$ is given 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. The system state vector $\mathbf{s}$ to be estimated consists of the person's location (X_F, Z_F) on the virtual plane, the camera height h_C, and the camera attitude:

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

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

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

In the camera frame, this becomes:

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

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:

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

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]:

$$\text{Repeat until convergence:} \quad \begin{cases} \mathbf{t}^* \leftarrow \arg\min_{\mathbf{t}} f(\mathbf{t}, \mathbf{r}) \\ \mathbf{r}^* \leftarrow \arg\min_{\mathbf{r}} f(\mathbf{t}, \mathbf{r}) \end{cases}$$

We solve this bounded nonlinear least-squares problem using the Dogbox method [39], [40] with a Cauchy cost function [41] for robustness.

D. Implementation in RPF Framework

Our RPF framework (Fig. 3) follows a standard pipeline [1], [3]–[5], taking an image stream and outputting velocity commands. The key steps are our normalization and optimization-based localization modules.

Person localization involves two phases. First, in an offline initialization with the robot static (known camera pose), we solve for the target's joint heights $\mathcal{H}_{all}$ and initial position by minimizing the reprojection error:

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

This is a linear least squares problem solved via SVD [42]:

$$X_F^*, Z_F^*, \mathcal{H}_{all}^* = \arg\min_{X_F, Z_F, \mathcal{H}_{all}} g(X_F, Z_F, \mathcal{H}_{all}) \quad (7)$$

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. Downstream modules perform data association, trajectory smoothing, person re-identification [43], and control based on the target's estimated location (X_F, Z_F).

IV. EXPERIMENTS

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** [3]: Locates person via neck point, assuming a fixed camera.
- **Ye's Method** [4]: Extends [3] using four points to handle occlusion, but still assumes a fixed camera.

Deep-learning-based:

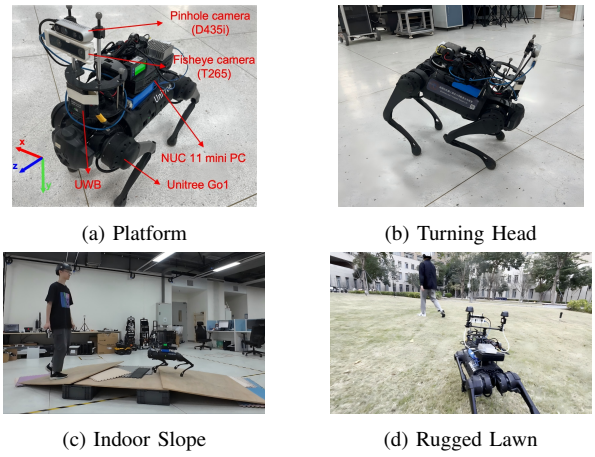
- **MonoLoco++** [24]: Regresses 3D location from 2D joints, trained on KITTI [35].
- **Depth Anything** [21]: Estimates a relative depth map. We use average joint depth for distance.
- **Multi-HMR** [25]: Recovers human mesh and estimates absolute distance between the pelvis and the camera.

B. Datasets

We evaluate on two public datasets: **FieldSAFE** [44], featuring a tractor in a field, and **KITTI** [35], an autonomous driving dataset that lacks severe ego-motion. We also introduce our **RPF-Quadruped** dataset, recorded on a Unitree Go1 [6] (Fig. 4a). It contains three scenarios (Fig. 4b-4d) with ground truth from a motion capture system or UWB. As shown in Table I, our dataset features closer-range interaction and significantly larger camera pitch/roll variations than others.

Dataset	KITTI [35]	FieldSAFE [44]	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 across different datasets.



(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 [6] (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 [8], [10], [11]. All methods were evaluated on the robot's NUC, except for Multi-HMR [25], 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)** [4], [24]. 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++ [24] performs best on KITTI, its training domain. In the challenging *Rugged Lawn* scenario, our method is visibly

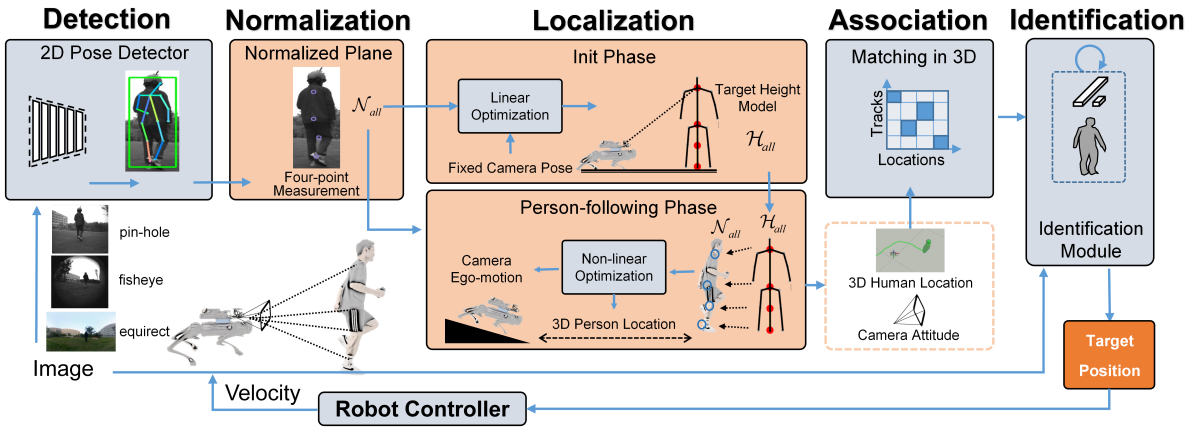


Fig. 3. Our proposed framework for monocular Robot Person Following (RPF). The modules highlighted in orange represent our key contributions: (1) a normalization step for camera-agnostic processing, and (2) a subsequent optimization-based person localization method.

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 [44] *ALE / VLE* ↓	KITTI [35] *ALE* ↓
Koide's Method [3]	0.396	0.289	0.3 / 0.3	1.924 / 5.012	1.451
Ye's Method [4]	0.294	0.261	0.3 / 0.3	1.856 / 3.952	1.420
MonoLoco++ [24]	0.820	0.510	0.6 / 0.2	4.152 / 4.705	**0.940**
Depth Anything [21]	0.571	0.523	0.5 / 0.6	1.528 / 1.022	2.963
Multi-HMR [25]	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	/	/

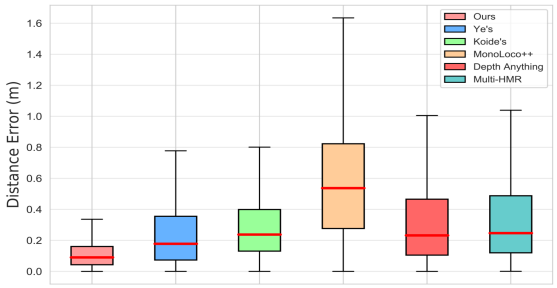


Fig. 5. A box plot illustrating the distance error of our method compared to baselines in *Rugged Lawn* scenario.

more accurate and stable, as shown by the error distribution (Fig. 5) and time-series distance plot (Fig. 6). Fig. 6(a) shows 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, outperforming deep-learning approaches in efficiency.

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 [3]	0.02	**0.0006**	**0.0206**
Ye's Method [4]	0.02	0.0008	0.0208
MonoLoco++ [24]	0.02	0.09	0.11
Depth Anything [21]	/	0.23	0.23
Multi-HMR [25]*	/	1.24	1.24
Ours	0.02	0.005	0.025

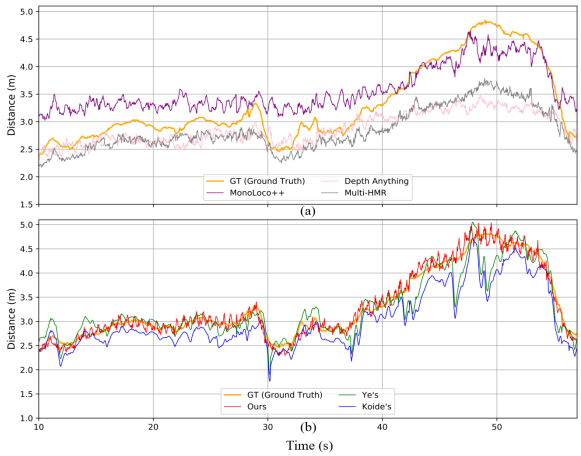


Fig. 6. Comparison of estimated distance over time on a sequence from the *Rugged Lawn* dataset (10s–57s). The plot shows the output of (a) deep-learning-based and (b) geo-model-based baselines.

V. CONCLUSIONS

In this paper, we presented a real-time (40 FPS), optimization-based method for monocular person localization under severe camera ego-motion. Our approach uses a four-point human model to jointly estimate camera attitude and person location. We demonstrated its effectiveness in a Robot Person Following (RPF) system on an agile quadruped robot and contributed a new dataset to foster research in this area, which is critical for HRI applications [1], [2]. Experiments on public and our own datasets validate our method's superior performance against geometric and deep-learning baselines. Future work will focus on handling more diverse postures with expressive human models, improving accuracy via ground plane estimation, and evaluating on large-scale egocentric datasets such as TPT-bench [45].

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] R. Jiao, Y. Wan, F. Poiesi, and Y. Wang, "Survey on video anomaly detection in dynamic scenes with moving cameras," *Artificial Intelligence Review*, vol. 56, pp. 3515 – 3570, 2023.

[3] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[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] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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