Systematic Evaluation of Reference Point Selection for Pedestrian Localization in Aerial-Ground Multi-View Monitoring Systems

Sunwoong Paik [1], Joonhee Youn [2], Phillip Kim [3], Jiheon Jung [4]

[1] Department of Future & Smart Construction Research, Korea Institute of Civil Engineering and Building Technology (KICT), Gyeonggi-do, Republic of Korea, sw.paik@kict.re.kr
[2] Corresponding Author, Department of Future & Smart Construction Research, Korea Institute of Civil Engineering and Building Technology (KICT), Gyeonggi-do, Republic of Korea, younj@kict.re.kr
[3] Department of Future & Smart Construction Research, Korea Institute of Civil Engineering and Building Technology (KICT), Gyeonggi-do, Republic of Korea, kpl@kict.re.kr
[4] Department of Future & Smart Construction Research, Korea Institute of Civil Engineering and Building Technology (KICT), Gyeonggi-do, Republic of Korea, jungjh97@kict.re.kr

Keywords: Aerial-ground camera system, Pedestrian localization, Pedestrian monitoring, Reference point, Urban surveillance

Abstract

Accurate pedestrian localization is fundamental to urban monitoring systems such as crowd management and anomalous behavior detection. Multi-view systems combining CCTV and UAV imagery extend monitoring coverage, but pedestrian locations estimated from each domain must be consistent in a common world coordinate system to enable cross-domain integration and reliable pedestrian tracking. Despite this requirement, no systematic comparison of how to define a reference point for pedestrian localization across aerial and ground-level camera domains has been conducted. This study proposes ten reference point selection rules combining bounding box- and pose estimation-based methods across CCTV and UAV domains, and evaluates their localization errors under varying camera heights, UAV altitudes, distances from the camera, and pedestrian poses in a simulation environment. Results show that using the bilateral toe tip midpoint as the reference point achieved the lowest overall localization error of 0.169 m, as toe tips are the body part closest to the ground. However, foot center- and ankle-based rules outperformed it under poses where the toe tips are lifted and at far distances, suggesting that condition-aware reference point selection may yield better overall performance. To the best of our knowledge, this is the first systematic comparison of reference point selection rules for pedestrian localization in aerial-ground multi-view systems, providing a quantitative foundation for reference point selection in practical monitoring system design.

1. Introduction

Monitoring and analyzing pedestrian movement in real time in urban environments plays a central role in smart city applications such as crowd density management and anomalous behavior detection (Liu et al., 2025). One of the core technologies underlying these systems is accurate pedestrian localization. Accordingly, extensive research has been conducted using single or multiple ground-level cameras (Hou et al., 2020; Hou and Zheng, 2021), and more recently, studies leveraging UAV imagery from aerial perspectives to monitor pedestrians over wider areas have become increasingly active (Shao et al., 2022; Dakic et al., 2025). These two domains have distinct imaging characteristics and can complement each other's limitations, which has led to growing adoption of heterogeneous multi-view configurations combining CCTV and UAV in real-world monitoring environments (Nguyen et al., 2024; Ito et al., 2025).

Since the pedestrian location estimated from each domain must correspond to the same position in a common world coordinate system, a consistent rule for selecting the reference point — the point derived from the pedestrian used to estimate their location — must be applied across both domains. An inconsistent or inaccurate reference point directly propagates into downstream tasks such as cross-domain trajectory association and crowd density estimation, where positional errors of even tens of centimeters can lead to identity mismatches or inaccurate density maps. As localization in both domains relies on image-based inverse projection, the localization error is directly affected by the actual height (Z value) of the reference point. A larger Z value amplifies this error, particularly as camera height decreases and as the pedestrian moves farther from the principal point (Lu and Dai, 2024; Kruber et al., 2020). Prior studies have estimated pedestrian locations using either bounding boxes (Kruber et al., 2020) or pose estimation algorithms (Rahimi et al., 2021; Abdulrahman and Zhu, 2023) in UAV and CCTV imagery. However, no study has systematically compared reference point selection rules across both aerial and ground-level camera domains simultaneously. To address this gap, the present study defines ten reference point selection rules applicable to both CCTV and UAV domains, and evaluates their localization errors under varying conditions including CCTV height, UAV altitude, horizontal distance from the camera, and pedestrian pose. Based on this analysis, the optimal reference point selection rule for each condition is derived, providing a quantitative basis for reference point selection in the design of aerial-ground multi-view pedestrian monitoring systems.

2. Experimental Setup

2.1 Simulation Environment

To evaluate the localization error of each reference point selection rule, a virtual simulation environment based on Blender was constructed. This setting was adopted to isolate the effect of reference point selection from confounding factors such as detection or pose-estimation errors, by providing exact ground-truth keypoint coordinates and camera parameters. The experimental scene models a simple urban intersection layout, as shown in Figure 1, with a road width of approximately 9 m. The scene includes nine pedestrians represented by rigged human models based on Mixamo, each consisting of 25 full-body joints as shown in Figure 2. To ensure diversity in stride length, the nine pedestrians were divided into three groups: Large (0.71 m), Medium (0.55 m), and Small (0.48 m), with three individuals per group. Stride length was defined as the horizontal distance between the left and right ankles at maximum stride. Pedestrians were placed randomly on the road within the scene, and a total of three scenes with different pedestrian placements were constructed. Two types of cameras were used: CCTV and UAV. The CCTV was configured as a fixed camera with a 2.8 mm wide-angle lens, positioned at a downward angle to face the center of the intersection. The UAV was configured with camera parameters based on a 35 mm full-frame equivalent and was set to capture the scene from a top-down view.

Figure 1. Oblique view of the simulated environment.

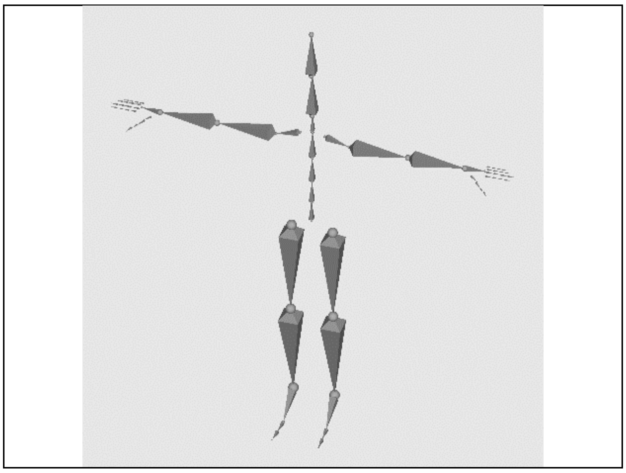

Figure 2. Full-body skeleton model used in the simulation.

2.2 Experimental Conditions

The experimental variables consist of camera height, pedestrian stride length, and pedestrian pose, as summarized in Table 1.

Variable	Conditions				
CCTV height	3m		5m		7m
UAV altitude	30m		50m		70m
Pose	Initial Contact	Loading Response	Mid Stance	Terminal Stance	Pre-swing

Table 1. Experimental conditions and variable settings.

2.2.1 Camera Height: The CCTV was set at three heights above the ground: 3 m, 5 m, and 7 m, with the depression angle adjusted at each height to face the center of the intersection. The UAV was set at three altitudes: 30 m, 50 m, and 70 m, maintaining a nadir orientation at each altitude. Example images captured under each condition are shown in Figures 3 and 4. Pedestrians that fell outside the captured frame under some low-altitude UAV conditions were excluded from the analysis.

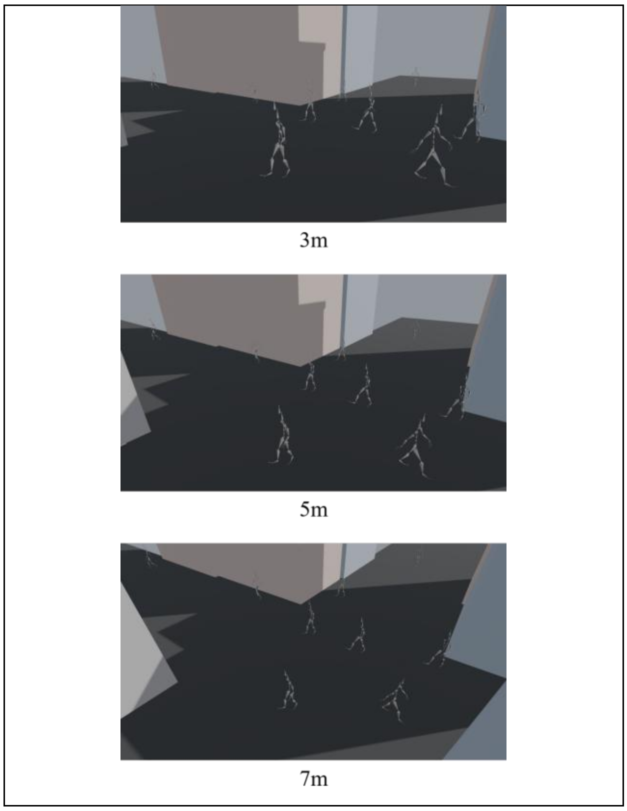

Figure 3. Example images captured at each CCTV height (3m, 5m, and 7m).

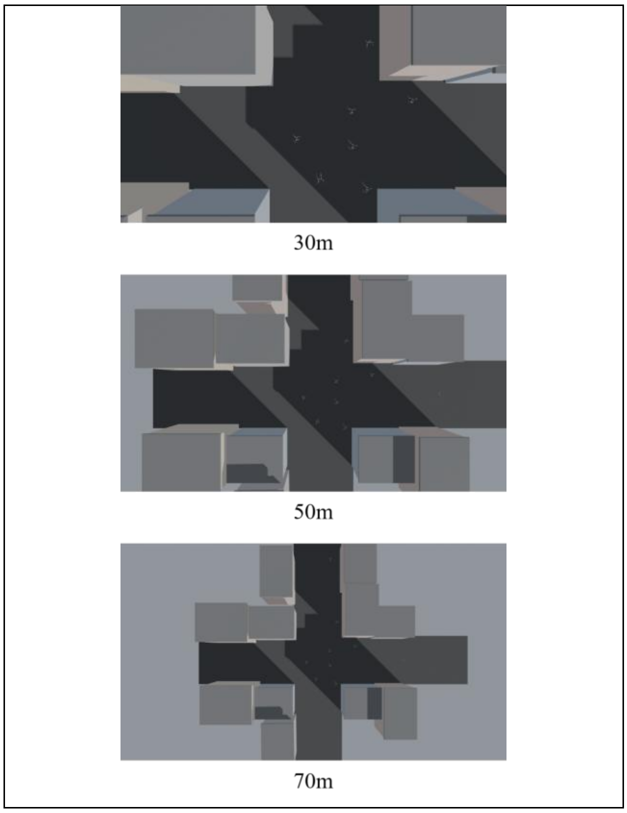

Figure 4. Example images captured at each UAV altitude (30m, 50m, and 70m).

https://doi.org/10.5194/isprs-archives-L-4-W3-2026-131-2026

2.2.2 Pedestrian Pose: Pedestrian poses were defined according to the classification by Perry and Burnfield (2010), consisting of five representative poses from the stance phase of the gait cycle: Initial Contact, Loading Response, Mid Stance, Terminal Stance, and Pre-swing, as illustrated in Figure 5. These five poses were selected after reviewing the full gait cycle, excluding redundant configurations.

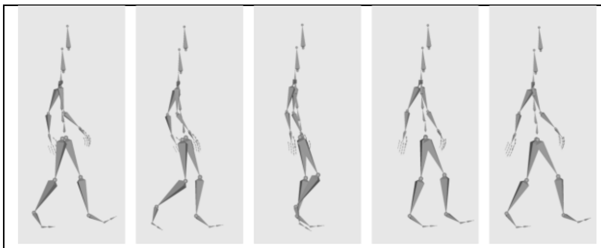

Figure 5. Five representative poses selected from the gait cycle.

2.3 Reference Point Selection Rules

This study defines four candidate reference points for each domain: (1) bounding box point (bottom center for CCTV, center for UAV), (2) bilateral ankle midpoint, (3) bilateral toe tip midpoint, and (4) bilateral foot center midpoint. Bounding box-based points represent the simplest approach without pose estimation and have been widely adopted (Kruber et al., 2020; Lu and Dai, 2024). Ankle keypoints have been used in pedestrian localization studies (Huang et al., 2023; Kim et al., 2017). Toe tip keypoints have become feasible with recent whole-body pose estimation models based on COCO-WholeBody (Jin et al., 2020; Xu et al., 2023), and the foot center is defined as the midpoint between the ankle and toe tip. For keypoint-based combinations, only those referencing the same body part in both domains are considered valid, while bounding box-based points are exempt from this constraint as they are not tied to any specific body part. This yields ten valid combinations in total, as defined in Table 2, along with their ground truth (GT) references — the exact 3D positions of the corresponding keypoints of the pedestrian skeleton model in the simulation environment.

2.4 Metric

Localization error quantifies the discrepancy between the estimated position of each rule's reference point projected onto the Z = 0 ground plane and its GT. The estimated position is computed by back-projecting the pixel coordinates of the reference point — obtained directly from the virtual environment — using camera parameters to find the intersection with the Z = 0 ground plane. A ray direction is defined from the camera position $C = (C_x, C_y, C_z)$and the 3D coordinates $P = (P_x, P_y, P_z)$ of the reference point, where t is the ray parameter at which the ray intersects the Z = 0 plane, computed as in Equations (1) and (2).

$$t = \frac{-C_z}{P_z - C_z}, \tag{1}$$

$$\hat{x} = C_x + t(P_x - C_x), \quad \hat{y} = C_y + t(P_y - C_y), \tag{2}$$

Here, $(\hat{x}, \hat{y})$ is the estimated location. The GT is defined as the XY coordinates (x_{GT}, y_{GT}) obtained by vertically projecting the 3D coordinates of each rule's GT point onto the Z = 0 plane. Localization error E is computed as the Euclidean distance (m) between the estimated location and the GT, as given in Equation (3). The final evaluation metric for each rule is defined as the sum of the localization errors computed separately in the CCTV and UAV domains, referred to as the localization error.

$$E = \sqrt{(\hat{x} - x_{GT})^2 + (\hat{y} - y_{GT})^2}, \tag{3}$$

Rule	CCTV	UAV	GT
Rule 1	Bounding box bottom center	Bounding box center	Ankle midpoint
Rule 2	Bounding box bottom center	Ankle midpoint	Ankle midpoint
Rule 3	Bounding box bottom center	Toe tip midpoint	Toe tip midpoint
Rule 4	Bounding box bottom center	Foot center midpoint	Foot center midpoint
Rule 5	Ankle midpoint	Bounding box center	Ankle midpoint
Rule 6	Ankle midpoint	Ankle midpoint	Ankle midpoint
Rule 7	Toe tip midpoint	Bounding box center	Toe tip midpoint
Rule 8	Toe tip midpoint	Toe tip midpoint	Toe tip midpoint
Rule 9	Foot center midpoint	Bounding box center	Foot center midpoint
Rule 10	Foot center midpoint	Foot center midpoint	Foot center midpoint

Table 2. Reference point combinations and GT definitions

3. Experimental Results

3.1 Effect of Camera Height

Analysis of localization error by CCTV height (Table 3) reveals distinct trends depending on the reference point type. Rules using CCTV keypoint-based reference points (Rules 5–10) showed increasing errors at lower camera heights, while rules using bounding box bottom center (Rules 1–4) showed the opposite trend, with lower errors at lower heights. Rule 8 recorded the lowest errors at 5 m (0.151 m) and 7 m (0.110 m), while at 3 m, Rule 4 ranked first (0.167 m), followed by Rule 3 (0.199 m) and Rule 2 (0.202 m), with Rule 8 ranking fourth at 0.248 m. Analysis by UAV altitude (Table 4) showed that error differences across altitude conditions were generally small. Rule 8 recorded the lowest errors across all altitudes, followed consistently by Rule 4, with rule rankings remaining largely stable regardless of UAV altitude.

Rule	CCTV Height		
	3 m	5 m	7 m
1	0.308	0.330	0.360
2	<u>0.202</u>	0.221	0.255
3	**0.199**	<u>0.216</u>	0.250
4	**<u>0.167</u>**	**0.189**	<u>0.227</u>
5	0.750	0.495	0.389
6	0.678	0.407	0.295
7	0.389	0.297	0.259
8	0.248	**<u>0.151</u>**	**<u>0.110</u>**
9	0.516	0.345	0.274
10	0.459	0.277	**0.202**

Table 3. Localization error by CCTV height for each rule (Bold underlined: 1st, bold: 2nd, underlined: 3rd.)

Rule	UAV Altitude		
	30 m	50 m	70 m
1	0.338	0.330	0.331
2	0.237	0.224	<u>0.218</u>
3	<u>0.226</u>	<u>0.221</u>	<u>0.218</u>
4	**0.202**	**0.192**	**0.188**
5	0.530	0.538	0.565
6	0.471	0.457	0.451
7	0.322	0.312	0.312
8	**<u>0.174</u>**	**<u>0.169</u>**	**<u>0.166</u>**
9	0.381	0.373	0.381
10	0.320	0.311	0.307

Table 4. Localization error by UAV altitude for each rule (Bold underlined: 1st, bold: 2nd, underlined: 3rd.)

3.2 Effect of Distance from the Camera

Analysis of localization error by distance from the camera (Table 5) divided the recorded distance range into three intervals: Near (< 8 m), Mid (8–14 m), and Far (> 14 m). Rule 8 achieved the lowest error in the Near range (0.0718 m) but showed the highest sensitivity to distance, increasing 4.4 times to 0.315 m in the Far range. Rule 4 demonstrated the most robust performance across all distance ranges, maintaining relatively low errors from 0.178 m (Near) to 0.220 m (Far), with only a 1.2- times increase. These results suggest that combining Rule 8 for near ranges with Rule 4 for far ranges may yield better overall performance than applying a single fixed rule.

3.3 Effect of Pedestrian Pose

Analysis of localization error by pedestrian pose (Table 6) showed that Rule 8 recorded notably low errors in Loading Response (0.057 m) and Mid Stance (0.082 m), where the feet are closest to the ground. In contrast, Rule 8 showed substantially higher errors in Terminal Stance (0.281 m), Pre-swing (0.219 m), and Initial Contact (0.210 m), where the toe tips are lifted from the ground. Figure 6 presents a comparison of Rule 8 alongside Rules 2, 3, and 4, which showed comparable or lower errors in these poses. In Terminal Stance, Rule 4 (0.160 m), Rule 2 (0.185 m), and Rule 3 (0.226 m) all recorded lower errors than Rule 8 (0.281 m), and in Pre-swing, Rule 4 (0.180 m) and Rule 2 (0.213 m) also outperformed Rule 8 (0.219 m). This suggests that in poses where the toe tips are lifted from the ground, using a bounding box-based reference point in the CCTV domain may contribute to more stable localization than relying solely on toe tip keypoints.

Rule	Distance		
	Near	Mid	Far
1	0.338	0.328	0.342
2	0.227	<u>0.218</u>	<u>0.246</u>
3	0.183	0.230	**0.239**
4	<u>0.178</u>	**0.190**	**<u>0.220</u>**
5	0.343	0.523	0.886
6	0.232	0.413	0.804
7	0.222	0.313	0.445
8	**<u>0.0718</u>**	**<u>0.150</u>**	0.315
9	0.227	0.366	0.620
10	**0.151**	0.280	0.556

Table 5. Localization error by distance from the camera for each rule (Bold underlined: 1st, bold: 2nd, underlined: 3rd.)

Rule	Stance Phase				
	Initial Contact	Loading Response	Mid Stance	Terminal Stance	Pre Swing
1	0.377	0.359	0.286	0.305	0.338
2	0.260	0.262	0.211	**0.185**	**0.213**
3	<u>0.248</u>	<u>0.219</u>	<u>0.192</u>	<u>0.226</u>	0.223
4	**0.233**	0.232	**0.166**	**<u>0.160</u>**	**<u>0.180</u>**
5	0.535	0.587	0.534	0.522	0.546
6	0.441	0.519	0.478	0.421	0.441
7	0.327	**0.146**	0.270	0.473	0.360
8	**<u>0.209</u>**	**<u>0.0565</u>**	**<u>0.0818</u>**	0.281	<u>0.219</u>
9	0.365	0.325	0.360	0.446	0.395
10	0.324	0.283	0.277	0.350	0.329

Table 6. Localization error by stance phase for each rule (Bold underlined: 1st, bold: 2nd, underlined: 3rd.)

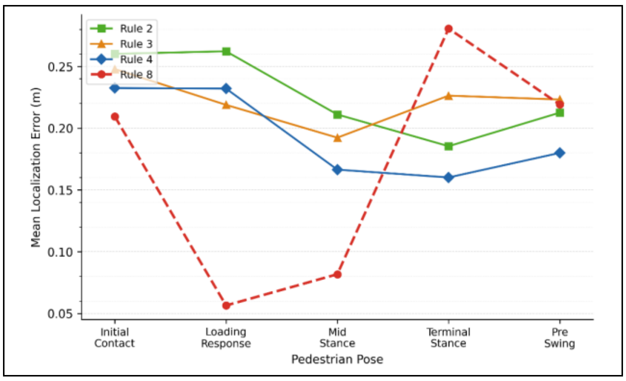

Figure 6. Localization error by pedestrian pose for Rules 2, 3, 4, and 8 (Rule 2: green, Rule 3: orange, Rule 4: blue, Rule 8: red).

3.4 Overall Localization Error

Analysis of overall mean localization error (Figure 7) shows that Rule 8 achieved the lowest error of 0.169 m, followed by Rule 4 (0.194 m), Rule 3 (0.222 m), and Rule 2 (0.226 m). In the CCTV domain, rules using bounding box bottom center or toe tip keypoints (Rules 1, 2, 3, 4, 7, 8) showed consistently lower errors, while ankle-based rules (Rules 5, 6) resulted in the highest overall errors. In the UAV domain, rules using ankle, foot center, or toe tip keypoints (Rules 2, 3, 4, 6, 8, 10) showed lower errors, whereas rules relying on bounding box center (Rules 1, 5, 7, 9) showed relatively higher UAV error contributions. Across both domains, toe tip-based reference points consistently demonstrated the lowest errors, as reflected in the strong overall performance of Rule 8.

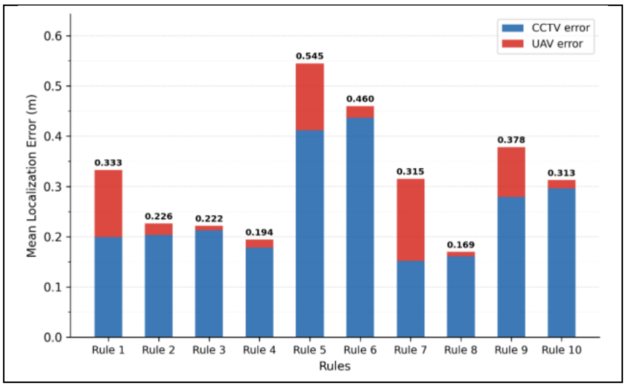

Figure 7. Mean localization error by rule (CCTV error: blue, UAV error: red).

4. Discussion

The overall results reveal distinct patterns in reference point performance across the two domains. In the CCTV domain, oblique viewing geometry causes reference points with larger Z values to incur greater horizontal localization error during inverse projection. Ankle-based rules (Rules 5, 6) use keypoints located higher on the body, resulting in larger Z values and thus greater susceptibility to this effect. Consequently, they showed substantially higher errors compared to rules using bounding box bottom center or toe tip keypoints. In the UAV domain, keypoint-based rules using ankle, foot center, or toe tip showed lower errors. In contrast, bounding box center-based rules (Rules 1, 5, 7, 9) showed higher UAV error contributions, likely due to relief displacement causing the projected position to deviate from the actual pedestrian location.

Rule 8 was not optimal under all conditions. Among pedestrian poses, in Terminal Stance, Rule 4 (0.160 m), Rule 2 (0.185 m), and Rule 3 (0.226 m) all outperformed Rule 8 (0.281 m), and in Pre-swing, Rule 4 (0.180 m) and Rule 2 (0.213 m) also showed lower errors than Rule 8 (0.219 m). In terms of distance from the camera, Rule 4 (0.220 m), Rule 3 (0.239 m), and Rule 2 (0.246 m) all outperformed Rule 8 (0.315 m) in the Far range. Rules 2, 3, and 4, which use a bounding box-based reference point in the CCTV domain, showed comparable or lower errors than Rule 8 under these conditions. As shown in Figure 8, which presents the standard deviation of localization error across CCTV height, distance, and pose conditions, these rules exhibited substantially lower variability than Rule 8. Rule 8 achieves the lowest errors when conditions are favorable, but its error increases sharply when the toe tips are lifted or when the pedestrian is at far distances — making it a high-reward but condition-sensitive choice.

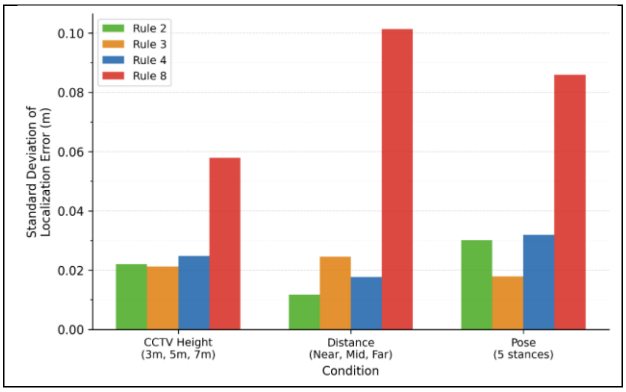

Figure 8. Standard deviation of localization error across conditions for Rules 2, 3, 4, and 8 (Rule 2: green, Rule 3: orange, Rule 4: blue, Rule 8: red).

One potential approach for improving localization accuracy is to dynamically select the reference point rule based on the current condition. By analyzing the gait cycle across consecutive frames, it is possible to identify whether the current pose corresponds to a toe-grounded phase (Loading Response, Mid Stance) or a toe-lifted phase (Terminal Stance, Pre-swing), enabling dynamic rule selection — Rule 8 for the former and Rule 4 or Rule 2 for the latter. A distance-based switching mechanism, applying Rule 4 or Rule 2 when the pedestrian is estimated to be in the Far range, can be incorporated alongside pose-based selection. Such a condition-aware hybrid strategy may yield lower localization error across a wider range of conditions compared to applying a single fixed rule.

The feasibility of such a strategy depends in part on the implementation complexity of the rules. Rule 8 requires accurate whole-body pose estimation, particularly reliable toe tip keypoint detection. Unlike object detection, pose estimation must predict the precise locations of individual body keypoints, making it an inherently more demanding task. In particular, whole-body models that estimate 133 keypoints present greater challenges than models that estimate 17 keypoints, owing to extreme scale variation across body parts (Jin et al., 2020). In UAV imagery, pedestrians are frequently captured with occluded body parts, further reducing the reliability of toe tip keypoint detection. Performance in real-world environments requires further validation (Farooq et al., 2026). In conditions where toe tip detection is unreliable, selectively incorporating bounding box-based reference points as a fallback may offer a more robust localization strategy. Alternatively, advances in pose estimation models with improved keypoint detection under such conditions could enhance the reliability of toe tip-based rules.

5. Conclusion

This study defined ten reference point selection rules (Rule 1–10) for pedestrian localization in a multi-view environment consisting of CCTV and UAV domains, and evaluated the localization error of each rule under varying conditions of camera height, UAV altitude, horizontal distance from the camera, and pedestrian pose. Experiments were conducted in a Blender-based simulation environment.

The results show that Rule 8, which uses the bilateral toe tip midpoint as the reference point in both domains, achieved the lowest overall mean localization error of 0.169 m. However, rules using a bounding box-based reference point in the CCTV domain (Rules 2, 3, and 4) showed comparable or lower errors under certain conditions, with substantially lower variability across conditions. These findings suggest that a condition-aware hybrid strategy — dynamically selecting the reference point rule based on the detected condition — may yield better overall performance than applying a single fixed rule.

By using a simulation environment where camera parameters and keypoint coordinates are exactly known, this study was able to isolate and evaluate the effect of reference point selection on localization error, independently of confounding factors such as detection model errors. To the best of our knowledge, this represents the first systematic comparison of reference point selection rules for pedestrian localization in heterogeneous aerial-ground multi-view systems, providing a quantitative foundation for reference point selection in practical monitoring system design. Future work should include end-to-end evaluation in real-world environments, where detection and pose estimation errors inevitably affect localization accuracy. Developing strategies that account for such errors, including model fine-tuning and integration with condition-aware rule selection, will be essential for practical deployment in aerial-ground multi-view monitoring systems.

References

Abdulrahman, B., and Zhu, Z., 2023. Real-time pedestrian pose estimation, tracking and localization for social distancing. Machine Vision and Applications, 34(1), 8. doi.org/10.1007/s00138-022-01356-0

Dakic, K., Thilakarathna, K., Calheiros, R.N., and Lim, T.J., 2025. A multi-drone multi-view dataset and deep learning framework for pedestrian detection and tracking. arXiv preprint arXiv:2511.08615. doi.org/10.48550/arXiv.2511.08615

Farooq, H., et al., 2026. FlyPose: Towards robust human pose estimation from aerial views. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2026). arXiv:2601.05747. doi.org/10.48550/arXiv.2601.05747

Hou, Y., Zheng, L., and Gould, S., 2020. Multiview detection with feature perspective transformation. Proceedings of the 16th European Conference on Computer Vision (ECCV 2020), Lecture Notes in Computer Science, vol. 12352, Springer, Cham, pp. 1–18. doi.org/10.1007/978-3-030-58571-6_1.

Hou, Y., and Zheng, L., 2021. Multiview detection with shadow transformer (and view-coherent data augmentation). Proceedings of the 29th ACM International Conference on Multimedia (MM '21), ACM, New York, pp. 1–10. doi.org/10.1145/3474085.3475310.

Huang, H.-W., Yang, C.-Y., Jiang, Z., Kim, P.-K., Lee, K., Kim, K., Ramkumar, S., Mullapudi, C., Jang, I.-S., Huang, C.-I., and Hwang, J.-N., 2023. Enhancing multi-camera people tracking with anchor-guided clustering and spatio-temporal consistency ID reassignment. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 5239–5249. doi.org/10.48550/arXiv.2304.09471

Ito, Y., Sasaki, T., and Kondo, S., 2025. Crowd anomaly detection from drone and ground. IEEE Access, 13, 28824–28835. doi.org/10.1109/ACCESS.2025.3537855

Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., and Luo, P., 2020. Whole-body human pose estimation in the wild. Proceedings of the 16th European Conference on Computer Vision (ECCV 2020), Lecture Notes in Computer Science, vol. 12352, Springer, Cham, pp. 196–214. doi.org/10.1007/978-3-030-58545-7_12

Kim, K., Heo, B., Byeon, M., and Choi, J.Y., 2017. Deep learning architecture for pedestrian 3-D localization and tracking using multiple cameras. Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), pp. 1147–1151. doi.org/10.1109/ICIP.2017.8296461.

Kruber, F., Sánchez Morales, E., Chakraborty, S., and Botsch, M., 2020. Vehicle position estimation with aerial imagery from unmanned aerial vehicles. Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV). doi.org/10.48550/arXiv.2004.08206.

Liu, S., Zhang, L., Ge, J., Li, W., and Long, Y., 2025. Framework for timely perception of spatiotemporal crowd congestion risk in public spaces based on video pedestrian tracking and geographic mapping. GIScience & Remote Sensing, 62(1), 2480416. doi.org/10.1080/15481603.2025.2480416

Lu, L., and Dai, F., 2024. Accurate road user localization in aerial images captured by unmanned aerial vehicles. Automation in Construction, 158, 105257. doi.org/10.1016/j.autcon.2023.105257

Nguyen, H., Nguyen, K., Sridharan, S., and Fookes, C., 2024. AG-ReID.v2: Bridging aerial and ground views for person re-identification. IEEE Transactions on Information Forensics and Security, 19, 2896–2908. doi.org/10.1109/TIFS.2024.3353078

Perry, J., and Burnfield, J.M., 2010. Gait Analysis: Normal and Pathological Function, 2nd ed. SLACK Incorporated, Thorofare, NJ. ISBN: 978-1-61711-430-4.

Rahimi, M.M., Khoshelham, K., Stevenson, M., and Winter, S., 2021. Pose-aware monocular localization of occluded pedestrians in 3D scene space. ISPRS Open Journal of Photogrammetry and Remote Sensing, 2, 100006. doi.org/10.1016/j.ophoto.2021.100006

Shao, Z., Cheng, G., Ma, J., Wang, Z., Wang, J., and Li, D., 2022. Real-time and accurate UAV pedestrian detection for social distancing monitoring in COVID-19 pandemic. IEEE Transactions on Multimedia, 24, 2069–2083. doi.org/10.1109/TMM.2021.3075566.

Xu, Y., Zhang, J., Zhang, Q., and Tao, D., 2023. ViTPose++: Vision transformer for generic body pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(2), 1212–1230. doi.org/10.1109/TPAMI.2023.3330016

Acknowledgements

This research was funded by the Korea Agency for Infrastructure Technology Advancement grant funded by the Ministry of Land, Infrastructure, and Transport (Grant No. RS-2022-00143782, Development of Fixed/Moving Platform Based Dynamic Thematic Map Generation Technology for Next Generation Digital Land Information Construction).