Location Determination of Dynamic Objects Using a Single CCTV with Monocular Depth Estimation

Jiheon Jung [1], Junhee Youn [2,] Jieun Kim[1], Junho Gong[1], Phillip Kim[1], Sunwoong Paik[1]

[1] Dept. of Future & Smart Construction Research, Korea Institute of Civil Engineering and Building Technology, 10223 Goyang-Si, Gyeonggi-Do, Republic of Korea; [2] Corresponding Author : Dept. of Future & Smart Construction Research, Korea Institute of Civil Engineering and Building Technology, 10223 Goyang-Si, Gyeonggi-Do, Republic of Korea

Keywords: Single CCTV, Location Determination, Monocular depth estimation

Abstract

This contribution presents a method to determine ground coordinates of pedestrians from a single CCTV frame using monocular depth estimation and orthophoto-based ground control points. Urban crowd monitoring requires pedestrian location information, but many CCTV-based approaches rely on accurate camera calibration or multi-view configurations, which are often unavailable in real deployments. In this study, we exploit relative depth values from a monocular depth estimation model (Depth Anything V2) and ground control points jointly identifiable in both the CCTV frame and an orthophoto in EPSG:5186. For each frame, depth-based distance ratios between the pedestrian and ground control point pairs are used to construct Apollonius circles in the orthophoto plane, and the pedestrian position is estimated by a weighted least-squares adjustment of their intersections. The method is evaluated on 180 frames across two scenes from an urban testbed with camera–target distances within approximately 50 m, across three GCP placement scenarios. For the optimal configuration (Scenario A), a mean RMSE of 1.989 m was achieved, excluding frames in which GCPs were temporarily occluded by moving objects, demonstrating that single-frame CCTV imagery combined with an orthophoto can achieve an accuracy of approximately 2 m without any EOP/IOP information, which is practically useful for urban crowd monitoring and dynamic thematic mapping. The influence of GCP placement geometry and occlusion conditions on estimation accuracy is also analyzed.

1. Introduction

1.1 Background & Motivation

Urban CCTV systems are widely used not only for simple detection and surveillance, but also as key data sources for a variety of public decision-making tasks, including traffic flow analysis, fire monitoring, and disaster response. In particular, quantitative pedestrian location information is essential for applications directly related to public safety, such as urban crowd management, accident prevention, and monitoring of densely populated areas (Dardour et al., 2025; Li et al., 2025).

Although CCTV is highly useful as an urban observation infrastructure due to its dense deployment and relatively low operating cost, most cameras used in real-world environments do not provide geometric sensor parameters such as exterior and interior orientation parameters (EOP/IOP). In addition, practical CCTV imagery frequently suffers from narrow fields of view, distortion, and occlusion, which further complicate geometric interpretation (Criminisi et al., 2000; Abdulrahman and Zhu, 2023).

As a result, estimating pedestrian ground coordinates quantitatively from a single CCTV frame remains a highly practical yet technically challenging task, especially when no prior camera geometry is available and only limited scene constraints can be assumed (Zeng et al., 2024).

To address these limitations, this study proposes a lightweight georeferencing pipeline that combines orthophoto-based ground control points (GCPs) with monocular depth estimation (MDE), without requiring detailed field calibration or precision surveying.

1.2 Related Work

Traditionally, image-based ground coordinate estimation has relied on camera calibration, multi-view triangulation, and plane-induced mappings such as homography estimation (Zhang 2000; Hartley and Sturm, 1997; Criminisi et al., 2000). In the context of CCTV-based pedestrian localization, prior studies have mainly adopted multi-camera configurations (Kim and Davis, 2006; Song et al., 2021), pre-calibrated camera parameters (Cosma et al., 2019), or ground-plane homography assumptions (Cosma et al., 2019; Abdulrahman and Zhu, 2023). However, when only a single CCTV frame is available and no prior EOP/IOP information is given, reliable coordinate estimation remains challenging because the observation geometry is weakly constrained and the absolute scale is inherently ambiguous (Criminisi et al., 2000; Eigen et al., 2014; Zeng et al., 2024).

Recently, monocular depth estimation (MDE) has emerged as a promising alternative because it can infer relative scene depth from a single image (Eigen et al., 2014; Ranftl et al., 2021). Nevertheless, the output of MDE is generally relative rather than metric depth, making direct conversion to absolute distance difficult, and its reliability may vary depending on scene structure, illumination, and material properties (Eigen et al., 2014; Ranftl et al., 2020; Zeng et al., 2024). To alleviate these issues, prior studies have explored additional sensor fusion, camera-geometry-based scale recovery, and landmark-based scale correction (Nützi et al., 2011; Lee et al., 2022; Cecille e al., 2024). However, such approaches typically require auxiliary devices or prior calibration information, which still limits their direct applicability to a fixed single-CCTV setting without

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-49-2026 |

EOP/IOP (Nützi et al., 2011; Lee et al., 2022; Cecille et al., 2024).

1.3 Contribution

This study proposes a method for estimating pedestrian ground coordinates by combining a single CCTV frame with orthophoto-based pseudo ground control points (Pseudo GCPs), even when EOP/IOP information is unavailable. The main contributions are as follows.

1. A pipeline is presented for estimating pedestrian absolute ground coordinates using only an uncalibrated monocular CCTV frame and planar GCPs obtained from an orthophoto. This addresses a practical limitation of real-world CCTV systems, most of which do not provide calibration information.

2. A dynamic object localization framework is introduced in which relative depth values from MDE are converted into distance-ratio constraints between GCP pairs, formulated as Apollonius circles and solved using weighted least squares.

3. Experiments on a real urban testbed show that the proposed single-frame CCTV-based coordinate estimation achieves practical positioning accuracy, and the effects of GCP geometric configuration and occlusion on estimation performance are analyzed.

2. Methodology

2.1 Overview

The overall workflow proposed in this study is shown in Figure 1. To estimate pedestrian absolute ground coordinates without camera calibration, a single CCTV frame and an orthophoto are used as inputs. In the CCTV frame, ground control points that are also identifiable in the orthophoto are selected, and the Depth Anything V2 model is applied to generate a pixel-wise relative depth map. In the orthophoto, the map coordinates of the selected control points are extracted in the KGD2002/Central Belt 2010(EPSG:5186) coordinate system. Distance ratios are then derived from the relative depth values between the pedestrian and the control points, formulated as Apollonius circles, and solved by weighted least squares to estimate the pedestrian’s ground coordinates. Finally, the estimated coordinates are compared with manually measured reference coordinates, and the positioning accuracy is evaluated using RMSE.

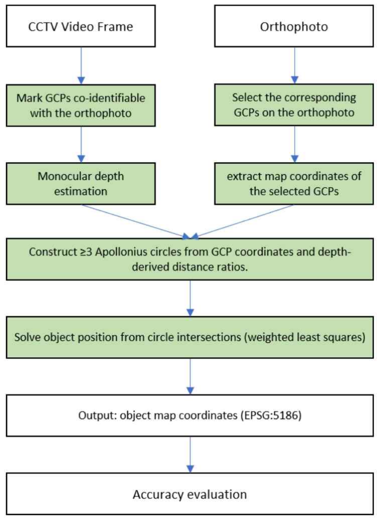
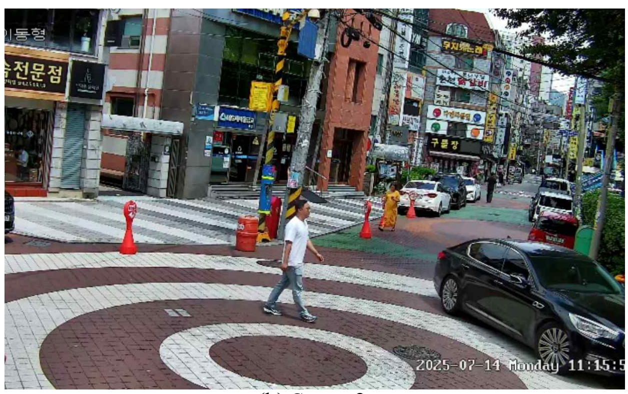


Figure 1. The flowchart of proposed methodology.

2.2 Generation and Preprocessing of Monocular Depth Estimation (MDE) Images

Frame-wise relative depth images were generated using the Depth Anything V2 model (Yang et al., 2024). The model has been reported to provide reliable depth estimation performance in outdoor short-range environments (within 50 m). To maintain scale consistency and numerical precision of the estimated depth values, the outputs were stored in 32-bit floating-point (float32) TIFF format. Figure 2 shows a portion of a CCTV frame used in the experiment, and Figure 3 shows the corresponding MDE output.

(a) Scene 1.



(b) Scene 2.

Figure 2. Sample CCTV video frames used in the experiment: (a) Scene 1; (b) Scene 2.

(a) Scene 1.

(b) Scene 2

Figure 3. Monocular depth estimation (MDE) results corresponding to the video frames in Figure 2: (a) Scene 1; (b) Scene 2.

2.3 Control Point Selection

Ground control points are selected as four or more ground features that can be commonly identified in both the orthophoto and the CCTV frame. Since at least three control-point pairs are required to construct Apollonius circles, a minimum of four control points is needed. To improve the geometric conditioning of the intersection, the control points should be distributed as evenly as possible over the entire image. A biased configuration concentrated on one side may weaken the geometric constraints of the distance-ratio formulation and increase the estimation error. As a general rule, points affected by occlusion or severe distortion are excluded from the control-point set. If the number of valid control points decreases due to occlusion, the stability of the estimation may also deteriorate.

2.4 Distance-Ratio Estimation and Apollonius Circle Construction

For each control-point pair (A, B), a distance ratio between the pedestrian P and the two control points is derived from the relative depth values of P, A, and B. Once k_{AB} is obtained, the pedestrian location P is constrained to the locus of points satisfying

$$\|P - A\| = k\|P - B\|$$

For each control-point pair (A, B), this locus forms an Apollonius circle. When $k \neq 1$, the center and radius of the circle are given by

$$C = \frac{A - k^2 B}{1 - k^2}, \quad r = \frac{k}{|1 - k^2|}\|A - B\|$$

An Apollonius circle is constructed on the orthophoto plane for each control-point pair. Ideally, all circles intersect at a single point. In practice, however, the pedestrian coordinate P is estimated using weighted least squares based on the set of constructed circles.

Instead of applying ordinary least squares to estimate the circle intersection, weighted least squares (WLS) is adopted. As the distance ratio k approaches 1, the radius of the Apollonius circle becomes larger, which expands the feasible region in which the pedestrian may exist and makes the estimation of P more difficult. This behavior is analogous to the geometric dilution of precision (GDOP) effect in GPS, where poor satellite geometry degrades positioning accuracy (Spilker, 1996). Therefore, the weight of each circle is defined as w = |1 - k^2|, so that the influence of circles with k close to 1 is automatically reduced. Multiple location candidates are generated from different combinations of control-point pairs, and outliers are removed using a MAD-based robust Z-score (Rousseeuw and Croux, 1993). The final coordinate is then determined as the median of the valid candidates.

3. Experiments

The input data consist of frame-wise extracts from a single CCTV video acquired in an urban street environment with mixed pedestrian and vehicle traffic in Indeogwon-dong, Dongan-gu, Anyang-si, Gyeonggi-do, Republic of Korea, together with an orthophoto of the same area. The dataset was constructed by the Korea Institute of Civil Engineering and Building Technology (KICT). The CCTV video was captured using an INTER COAX QUATTRO-LITE device modified for mobile deployment. The video was recorded at a resolution of 1920 × 1080 pixels and a frame rate of 30 fps. The observation distance to the target pedestrians was within approximately 50 m. The orthophoto was generated for the same area based on drone imagery and ground control point surveying data, and was provided in the Korean Central Belt coordinate system (EPSG:5186) with a ground sampling distance (GSD) of approximately 25 cm.

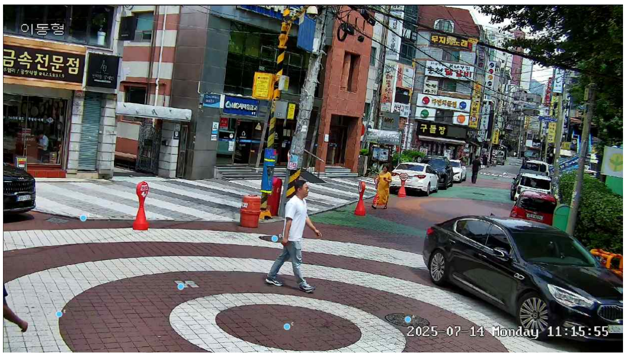
The experiments were conducted on a system equipped with an AMD Ryzen 9 9950X3D CPU and an NVIDIA GeForce RTX 5090 GPU. The computational bottleneck was MDE inference, which required approximately 5 min per frame, whereas the proposed coordinate estimation stage required only approximately 0.1 s per frame. Since the MDE output can be generated offline in advance, the coordinate estimation stage itself remains computationally lightweight.

The experiments were performed on a total of 180 frames, with 90 frames selected from each of two acquisition locations (Scene 1 and Scene 2) within the same testbed. Each frame was processed independently to evaluate the robustness of the proposed single-frame localization method under varying scene conditions. The two scenes represent slightly different observation conditions within the same urban testbed. The pedestrian ground-truth (GT) coordinates were manually identified on the orthophoto for each frame and used as reference coordinates in the EPSG:5186 system.

To analyze the effect of control-point configuration on localization performance, three scenarios were designed. These scenarios were intended to examine the effects of control-point geometry and point availability on localization accuracy. Scenario A is the baseline condition, in which six control points are evenly distributed across the image. Scenario B uses the same six control points, but arranged in a spatially biased configuration on one side of the image. Scenario C evaluates the minimum-configuration case by using only four control points, which is the minimum number required to construct the Apollonius-circle formulation. The control-point pair configurations for each scenario are listed in Table 1, and example layouts are shown in Figure 4. The average RMSE was computed for each scenario and compared accordingly. Frames affected by visible occlusion at selected control-point locations were excluded from the valid-frame analysis because such occlusions could distort the depth values used for distance-ratio estimation. Accordingly, RMSE was reported for both the full set of frames and the valid-frame subset.

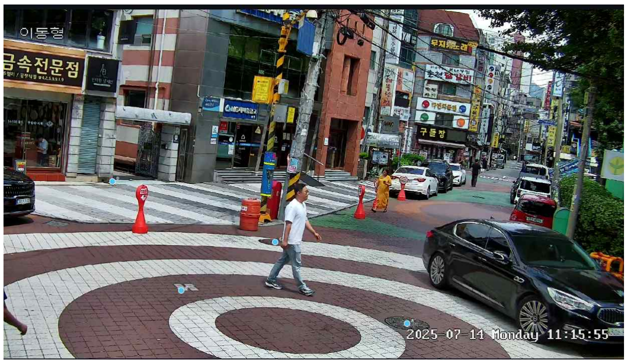
Scenario	No. of GCPs	Placement	GCP Pairs
A	6	Uniform distribution	A-B, C-D, E-F
B	6	Biased(one-sided)	A-B, C-D, E-F
C	4	Uniform distribution	A-B, B-C, C-D

Table 1. GCP configuration for each experimental scenario



(a) Scenario A : uniform distribution

(b) Scenario B : biased placement



(c) Scenario C : minimum GCPs

Figure 4. GCP placement examples for each scenario (Scene 2). (a) Scenario A: uniform distribution, (b) Scenario B: biased placement, (c) Scenario C: minimum GCPs.

4. Results

The pedestrian ground-coordinate estimation results for each scenario and acquisition location are summarized in Table 2. Scenario A achieved the lowest error among the three scenarios, with a mean RMSE of 1.989 m. Despite the absence of prior EOP/IOP information, meter-level accuracy was achieved under favorable control-point configurations, while degraded configurations led to substantially larger errors.

Scenario	Scene 1 RMSE (m)	Scene 2 RMSE (m)	Mean RMSE (m)
A	2.211	1.767	1.989
B	7.259	5.479	6.369
C	4.139	2.007	3.073

Table 2. RMSE values are computed on valid frames excluding GCP-occluded frames.

In Scenario A, where the control points were evenly distributed, the uncertainty region induced by the intersections of the Apollonius circles was effectively reduced even under realistic error conditions in which an ideal single intersection point was not formed. This contributed to stable coordinate estimation using weighted least squares (WLS). These results suggest that a well-distributed control-point configuration provides stronger geometric constraints for robust localization.

In contrast, Scenario B, in which the same six control points were arranged in a spatially biased configuration, produced a mean RMSE of 6.369 m, which is approximately 3.2 times larger than that of Scenario A. This quantitatively demonstrates the influence of geometric imbalance in control-point placement on estimation accuracy. Scenario C, which used only the minimum number of four control points, yielded a mean RMSE of 3.073 m, demonstrating that effective localization is still achievable even under the minimum control-point condition. This indicates that the proposed framework can remain operational even when only a limited number of valid control points are available.

Figure 5 shows an overlay of the estimated pedestrian coordinates on the orthophoto for a representative frame in Scenario A. The results indicate that, except for extreme occlusion cases caused by vehicles or personal mobility devices, pedestrian positions were estimated consistently across consecutive frames.

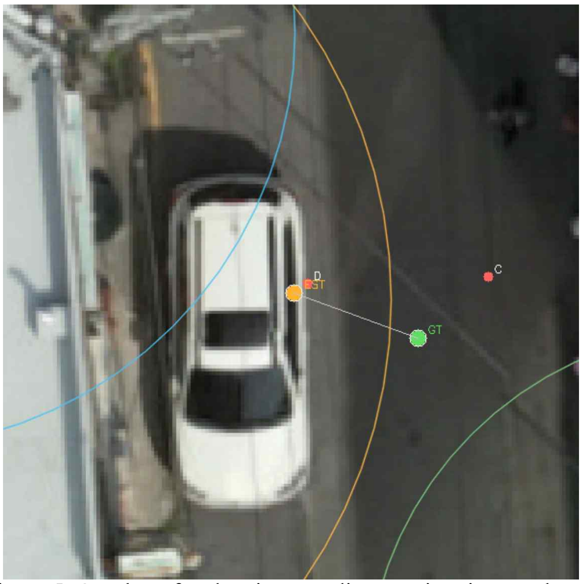


Figure 5. Overlay of pedestrian coordinate estimation results on the orthophoto (Scenario A, Scene 1).

5. Conclusions

The mean RMSE of 1.989 m obtained in the experiments indicates that meter-level localization accuracy can be achieved using only a single CCTV frame without geometric sensor parameters (EOP/IOP) and an orthophoto of the same area. This level of accuracy is practically meaningful for urban geospatial applications such as crowd and pedestrian monitoring and the generation of dynamic thematic maps.

The scenario-based analysis shows that the localization performance of the proposed method depends strongly on the control-point configuration strategy. When the control points were distributed evenly across the field of view, as in Scenario A, the uncertainty region associated with the circle intersections was reduced, resulting in more stable coordinate estimation. In contrast, when the control points were spatially biased to one side, as in Scenario B, the distance-ratio constraints became weaker and the localization accuracy deteriorated. When the number of valid control points decreased due to occlusion by moving objects or surrounding structures, the stability of the estimation also tended to decrease, which is one of the main limitations of the proposed method. In addition, the reference coordinates used in this study were manually obtained from an orthophoto with a GSD of approximately 25 cm, and therefore may include interpretation error. Although control points are currently selected manually, the practical applicability of the proposed pipeline could be significantly improved by extending it to an automatic matching framework that identifies corresponding feature points between the orthophoto and the CCTV frame.

The favorable performance observed within an observation distance of approximately 50 m suggests that the relative depth information obtained from MDE can be effective within a limited distance range. This also implies that additional correction or modeling may be required for longer observation distances. In addition, pedestrian coordinates were found to be produced continuously over consecutive frames, suggesting the potential for future extension to pedestrian trajectory generation and tracking. However, verification of absolute depth accuracy and trajectory-based analysis using consecutive frames are beyond the scope of the present study and are therefore left for future work.

Overall, the proposed approach, which formulates the relationship between MDE-derived relative depth and orthophoto-based GCP distance ratios as Apollonius circles, was experimentally shown to enable meter-level pedestrian localization in real urban CCTV environments without camera calibration. This suggests that the method can be practically integrated with existing geospatial infrastructures for urban pedestrian monitoring.

Future work will include validation of the absolute depth accuracy of the MDE output, performance analysis across different observation distances, extension to pedestrian trajectory generation and tracking based on consecutive frames, and investigation of automatic GCP selection between orthophotos and CCTV frames. In addition, further experiments under more diverse imaging conditions and environments are needed to evaluate the reliability and generalizability of the proposed method.

Acknowledgements

This research was funded by a Korea Agency for Infrastructure Technology Advancement (KAIA) grant from the Ministry of Land, Infrastructure, and Transport, grant number RS2022-00143782 (Development of Fixed/Moving Platform Based Dynamic Thematic Map Generation Technology for Next Generation Digital Land Information Construction).

References

Abdulrahman, B., Zhu, Z., 2023. Real-time pedestrian pose estimation, tracking and localization for social distancing. Machine Vision and Applications, 34(1), 6. https://doi.org/10.1007/s00138-022-01356-0

Cecille, A., Duffner, S., Davoine, F., Neveu, T., Agier, R., 2024. Ground Constraint for Metric Self-Supervised Monocular Depth Estimation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds), Computer Vision – ECCV 2024. Lecture Notes in Computer Science. Springer, Cham.

Cosma, A., Radoi, I.E., Radu, V., 2019. CamLoc: Pedestrian Location Estimation through Body Pose Estimation on Smart Cameras. In: 2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Pisa, Italy, 30 September–3 October 2019.

Criminisi, A., Reid, I., Zisserman, A., 2000. Single View Metrology. International Journal of Computer Vision, 40(2), 123–148. https://doi.org/10.1023/A:1026598000963

Dardour, A., El Haji, E., Achkari Begdouri, M., 2025. Video Surveillance and Artificial Intelligence for Urban Security in Smart Cities: A Review of a Selection of Empirical Studies from 2018 to 2024. Smart Cities, 10(1), 15. https://doi.org/10.3390/smartcities10010015

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In: Advances in Neural Information Processing Systems 27 (NeurIPS 2014), 2366–2374.

Hartley, R.I., Sturm, P., 1997. Triangulation. Computer Vision and Image Understanding, 68(2), 146–157. https://doi.org/10.1006/cviu.1997.0547

Kim, K., Davis, L.S., 2006. Multi-camera Tracking and Segmentation of Occluded People on Ground Plane Using Search-Guided Particle Filtering. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds), Computer Vision – ACCV 2006. Lecture Notes in Computer Science, 3852, Springer, Berlin, Heidelberg, 98–109.

Lee, J.C., Chen, C.C., Shen, C.T., Lai, Y.C., 2022. Landmark-Based Scale Estimation and Correction of Visual-Inertial Odometry for VTOL UAVs in a GPS-Denied Environment. Sensors, 22(24), 9654. https://doi.org/10.3390/s22249654

Li, H., Wang, Z., Wu, Y., 2025. A review on the use of top-view surveillance videos for crowd management and safety in public buildings. Accident Analysis & Prevention, 212, 107903. https://doi.org/10.1016/j.aap.2025.107903

Nützi, G., Weiss, S., Scaramuzza, D., Siegwart, R., 2011. Fusion of IMU and Vision for Absolute Scale Estimation in Monocular SLAM. Journal of Intelligent and Robotic Systems, 61, 287–299. https://doi.org/10.1007/s10846-010-9490-z

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V., 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3), 1623–1637.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision Transformers for Dense Prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 12179–12188.

Rousseeuw, P.J., Croux, C., 1993. Alternatives to the Median Absolute Deviation. Journal of the American Statistical Association, 88(424), 1273–1283. https://doi.org/10.2307/2291267

Song, L., Wu, J., Yang, M., Zhang, Q., Li, Y., Yuan, J., 2021. Stacked Homography Transformations for Multi-View Pedestrian Detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 6049–6057.

Spilker, J.J., 1996. Satellite Constellation and Geometric Dilution of Precision. In: Parkinson, B.W., Spilker, J.J., Jr (eds), Global Positioning System: Theory and Applications, Volume I. American Institute of Aeronautics and Astronautics, Washington, DC.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth Anything V2. In: Advances in Neural Information Processing Systems 38 (NeurIPS 2024).

Zeng, Z., Wu, Y., Park, H., Wang, D., Yang, F., Soatto, S., Lao, D., Hong, B.-W., Wong, A., 2024. RSA: Resolving Scale Ambiguities in Monocular Depth Estimators through Language Descriptions. In: Advances in Neural Information Processing Systems 38 (NeurIPS 2024).

Zhang, Z., 2000. A Flexible New Technique for Camera Calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(11), 1330–1334.