sensors

MDPI

Article

Monocular Pedestrian 3D Localization for Social Distance Monitoring

Yiru Niu [1], Zhihua Xu [1,2,*], Ershuai Xu [1], Gongwei Li [1,3], Yuan Huo [1] and Wenbin Sun [1]

[1] College of Geoscience and Surveying Engineering, China University of Mining and Technology (Beijing), Beijing 100083, China; y.niu@student.cumtb.edu.cn (Y.N.); es.xu@student.cumtb.edu.cn (E.X.); ligw@aircas.ac.cn (G.L.); y.huo@student.cumtb.edu.cn (Y.H.); swb@cumtb.edu.cn (W.S.)
[2] State Key Laboratory of Coal Resources and Safe Mining, China University of Mining and Technology (Beijing), Beijing 100083, China
[3] Key Laboratory of Quantitative Remote Sensing Information Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
* Correspondence: z.xu@cumtb.edu.cn

Abstract: Social distancing protocols have been highly recommended by the World Health Organization (WHO) to curb the spread of COVID-19. However, one major challenge to enforcing social distancing in public areas is how to perceive people in three dimensions. This paper proposes an innovative pedestrian 3D localization method using monocular images combined with terrestrial point clouds. In the proposed approach, camera calibration is achieved based on the correspondences between 2D image points and 3D world points. The vertical coordinates of the ground plane where pedestrians stand are extracted from the point clouds. Then, using the assumption that the pedestrian is always perpendicular to the ground, the 3D coordinates of the pedestrian's feet and head are calculated iteratively using collinear equations. This allows the three-dimensional localization and height determination of pedestrians using monocular cameras, which are widely distributed in many major cities. The performance of the proposed method was evaluated using two different datasets. Experimental results show that the pedestrian localization error of the proposed approach was less than one meter within tens of meters and performed better than other localization techniques. The proposed approach uses simple and efficient calculations, obtains accurate location, and can be used to implement social distancing rules. Moreover, since the proposed approach also generates accurate height values, exclusionary schemes to social distancing protocols, particularly the parent-child exemption, can be introduced in the framework.

Keywords: three-dimensional localization; social distance monitoring; monocular camera; LiDAR point clouds

check for updates

Citation: Niu, Y.; Xu, Z.; Xu, E.; Li, G.; Huo, Y.; Sun, W. Monocular Pedestrian 3D Localization for Social Distance Monitoring. *Sensors* **2021**, *21*, 5908. https://doi.org/10.3390/s21175908

Academic Editor: Antonia Spano'

Received: 11 July 2021
Accepted: 28 August 2021
Published: 2 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

At the beginning of 2020, the COVID-19 pandemic swept the world, significantly impacting people's lives and global economic development. While researchers have been developing therapeutics and vaccines for coronaviruses, social distancing is still recommended, at least until 2022, to minimize the risks of another outbreak [1]. Recent studies have shown that when more than one-meter distancing is maintained, the risk of transmission declines by about 82% [2]. In many countries, various rules and strategies have been introduced to observe and monitor interpersonal distancing in public places to curb the spread of the virus [3]. South Korea and India, for example, use Global Positioning System (GPS) data to monitor the movements of pedestrians to identify potential contacts. In addition, India has been utilizing smartphones to locate and monitor COVID-19 patients in target areas with the help of GPS and Bluetooth [4]. However, inside buildings, GPS is susceptible to signal attenuation and multipath effect, resulting in incorrect positioning and limiting its use in implementing social distancing.

The fundamental challenge in implementing social distancing and suppressing the spread of the virus is how to perceive people in 3D. Over the past decades, several emerging pedestrian 3D localization technologies have been introduced, from multi-sensor approaches to vision-based solutions. Guo et al. [5] put forward the Indoor Knowledge Graph Framework, integrating spatial information from various mobile phone sensors (including inertial sensor, Wi-Fi, and magnetic field) to achieve indoor positioning. Levchev et al. [6] used cameras, LiDAR, wireless sensors, inertial gyroscopes, and other sensors as data acquisition platforms and proposed a multi-sensor hybrid database configuration. Liang et al. [7] used different sensors to acquire image data and geographic location data simultaneously to construct an indoor 3D map with geographic coordinates. Zhang et al. [8] fused 3D point clouds with scene images to obtain the 3D positioning of outdoor scenes. However, as data acquisition platforms, most of these methods rely on a variety of sensors and have great dependence on the diversity of devices. In addition, some sensors (e.g., wireless) require multiple receiving antennas. Ultra-Wide Band (UWB) needs the target to wear tags or chips, which can be difficult to deploy in actual scenarios.

Stereovision is a relatively mature 3D localization technology, but our goal is to propose a solution that primarily uses monocular closed-circuit cameras. At present, monocular vision research mainly focuses on the 3D position estimation of vehicles [9,10] since pedestrians have varying heights and body shapes and lack enough attribute information. Most traditional methods for pedestrian 3D localization using monocular cameras rely on known information about the scene or pedestrian to obtain the unknown metrics. For example, Gunel et al. [11] and Das and Meher [12] used medical statistical data to locate pedestrians based on the relationship of human height and body parts (or stride length). Using physical constraints, Bieler et al. [13] converted pixel measurements into human height through a simple motion trajectory analysis (jumping or running). In most cases, this estimation approach indirectly obtains the pedestrian's 3D position using some other metrics, which may result in error propagation.

With developments in artificial intelligence in recent years, various studies have used machine learning technology for monocular pedestrian 3D localization. For example, Deng et al. [14] proposed an end-to-end trainable neural network, which only uses synthetic data to obtain stereo images and body pose estimation. Bertoni et al. [15] used a feedforward neural network to predict confidence intervals of 2D human posture and 3D position based on the loss function of Laplace distribution. Bertoni et al. [16] also used a neural network to predict confidence intervals and estimate the 3D position and orientation of pedestrian. However, in most of these studies, pedestrians are assumed to be of equal height, resulting in a task error of about one meter for pedestrians 20 m away.

Visual-based technology has allowed the remote determination of pedestrians' posture and acquisition of texture information. However, due to the limitations of having only one projection line corresponding to each object point and the lack of depth value, additional information is needed to convert 2D images into 3D. In this context, we propose an effective alternative to achieve three-dimensional positioning. We used terrestrial LiDAR to capture 3D maps to estimate the parameters of the monocular camera. Since the pedestrian is always perpendicular to the ground, we could determine the 3D localization of pedestrians. Our method focuses on solving the limitations of plane positioning and improving the applications of traditional photogrammetry in three-dimensional space.

2. Proposed Method

A monocular pedestrian 3D localization for social distance monitoring framework is proposed in this study. Figure 1 presents the general flowchart of the proposed method. The proposed 3D localization approach includes three main parts: monocular camera calibration (preprocessing), pedestrian 3D localization (localization), and social distance calculation (analysis).

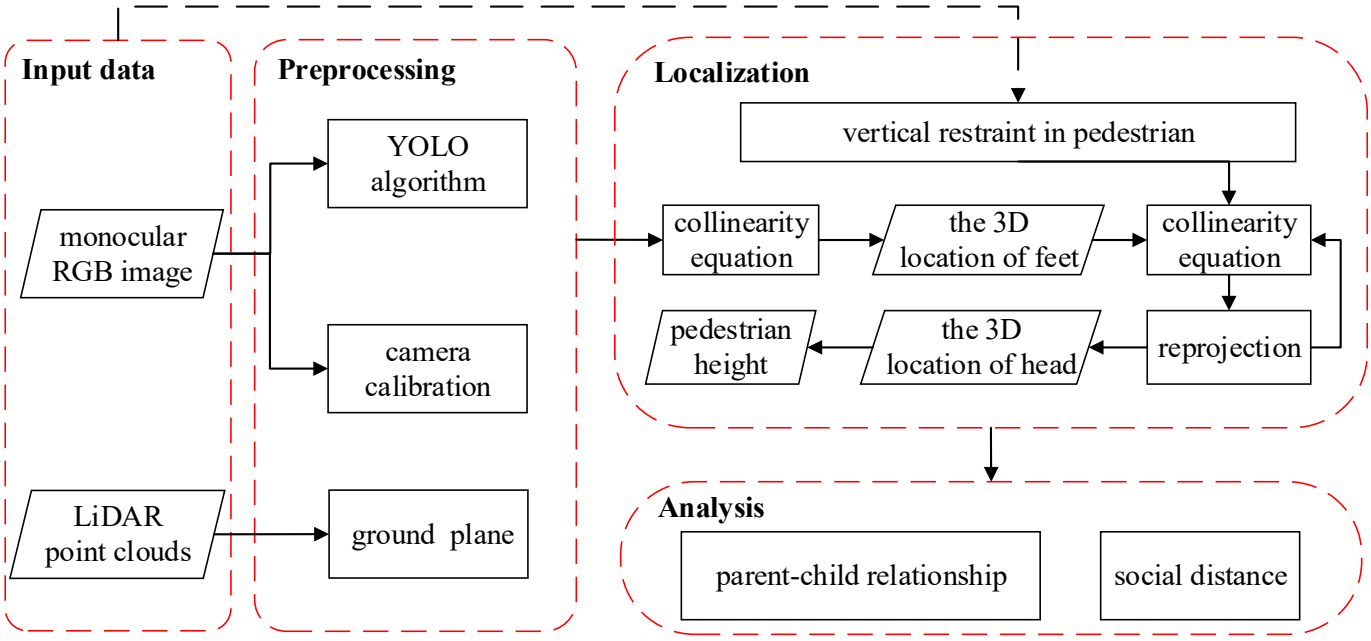


Figure 1. The flowchart of the proposed method.

2.1. Preprocessing

2.1.1. Camera Calibration

In computer vision, camera calibration is used to determine the camera position and establish the imaging model. Then the relative relationship between the space coordinate system and the image coordinate system can be obtained. To obtain 3D positioning based on monocular images, camera calibration is needed to recover internal (i.e., the focal length f and the pixel coordinates of the principal point $(u_0,\ v_0)$) and external parameters (i.e., translation vector $\boldsymbol{T}$ and rotation matrix $\boldsymbol{R}$). For every scenario, at least six pairs of feature points are manually selected from the terrestrial point clouds and monocular image to obtain the 2D-3D matching relationship. The 2D and 3D point coordinates are respectively denoted by $(u_i,\ v_i)$ and $(X_i,\ Y_i,\ Z_i)$ where i represents different matching points. We use direct linear transformation (DLT) [17] to acquire the camera orientation. The parameters can be formulated as:

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & u_0 & 0 \\ 0 & f & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ 0^{\mathrm{T}} & 1 \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} \tag{1}$$

Express it as an equation for the unknown quantity $l_j (j = 1, 2, \ldots 12)$:

$$\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \boldsymbol{L} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} = \begin{bmatrix} l_1 & l_2 & l_3 & l_4 \\ l_5 & l_6 & l_7 & l_8 \\ l_9 & l_{10} & l_{11} & l_{12} \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix} \tag{2}$$

The homogeneous coordinate l_{12} does not change the values. We set $l_{12} = 1$. Then substitute the i ($i \geq 6$) pairs of matching points into the Equation (2):

$$\begin{bmatrix} X_1 & Y_1 & Z_1 & 1 & 0 & 0 & 0 & 0 & u_1X_1 & u_1Y_1 & u_1Z_1 \\ 0 & 0 & 0 & 0 & X_1 & Y_1 & Z_1 & 1 & v_1X_1 & v_1Y_1 & v_1Z_1 \\ X_2 & Y_2 & Z_2 & 1 & 0 & 0 & 0 & 0 & u_2X_2 & u_2Y_2 & u_2Z_2 \\ 0 & 0 & 0 & 0 & X_2 & Y_2 & Z_2 & 1 & v_2X_2 & v_2Y_2 & v_2Z_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_i & Y_i & Z_i & 1 & 0 & 0 & 0 & 0 & u_iX_i & u_iY_i & u_iZ_i \\ 0 & 0 & 0 & 0 & X_i & Y_i & Z_i & 1 & v_iX_i & v_iY_i & v_iZ_i \end{bmatrix} \begin{bmatrix} l_1 \\ l_2 \\ l_3 \\ l_4 \\ l_5 \\ l_6 \\ l_7 \\ l_8 \\ l_9 \\ l_{10} \\ l_{11} \end{bmatrix} = \begin{bmatrix} -u_1 \\ -v_1 \\ -u_2 \\ -v_2 \\ \vdots \\ -u_i \\ -v_i \end{bmatrix} \quad (3)$$

Then write Equation (3) as:

$$\boldsymbol{AL} = \boldsymbol{b} \quad (4)$$

Solve for $\boldsymbol{L}$ using the following formula:

$$\boldsymbol{L} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{b} \quad (5)$$

We can then use mathematical operations to get the values of the internal and external parameters.

2.1.2. Detector

According to Section 2.1.1, it is very important to obtain accurate 2D pixel coordinates of the pedestrian. The YOLO object detector [18] is used for pedestrian detection in monocular image coordinates. Only one CNN operation is needed to realize the end-to-end prediction. The algorithm uses the whole monocular image as input to the network and simultaneously predicts the target area and its category. In this paper, only the detection results of people are selected. This operation would provide the midpoint pixel coordinates of the detection box's upper and lower edges, which would then be used to represent the head and feet of the pedestrian.

2.1.3. Extraction of the Vertical Coordinates of the Ground

The gyroscope in the LiDAR helps maintain the *z*-axis of the point cloud coordinate system in a vertical orientation, and a pedestrian remains upright while walking. The 3D coordinates of pedestrian feet are calculated with this constraint condition. We crop the terrestrial point clouds based on the view range of the monocular image. Then a grid-based point cloud segmentation algorithm is used to obtain the ground plane. The algorithm first projects the 3D point clouds onto the grid map, with a single grid regarded as the minimum computing unit. Then, the statistical characteristics of the elevation direction in each grid unit, including extremum, mean, and variance are computed. The average height difference of a grid is compared with its adjacent grids. Grids with small height differences are marked as potential ground location. Adjacent potential ground locations constitute a connected domain. The connected domain with the largest area can be regarded as the ground. Then from the point clouds, we can acquire the vertical coordinates of the ground, which provides necessary data for the subsequent pedestrian 3D localization.

2.2. 3D Localization

The midpoint t of the pedestrian bounding box's upper edge (i.e., the position of the pedestrian's head) and the midpoint b of its lower edge (i.e., the position of pedestrian's feet) are used as marking points. For pedestrians who only appear in a single frame of the monocular camera, their 3D coordinates cannot be directly obtained from a single uncalibrated image and would have to be solved using additional information. Given the presence of a gyroscope in LiDAR, the *z*-axis in its coordinate system would always be

perpendicular to the ground. Therefore, Z_g is introduced to represent the z-axis coordinate of the ground plane in the point cloud coordinate system. Based on the camera calibration and posture recovery, the 3D coordinates (X_b, Y_b, Z_b) of the pedestrian feet point b can be obtained using the space intersection:

$$\begin{cases} X_b = \frac{(Z_g - Z_S)\left(a_1 * \frac{u_0 - u_b}{f} + a_2 * \frac{v_b - v_0}{f} + a_3\right)}{c_1 * \frac{u_0 - u_b}{f} + c_2 * \frac{v_b - v_0}{f} + c_3} + X_S \\ Y_b = \frac{(Z_g - Z_S)\left(b_1 * \frac{u_0 - u_b}{f} + b_2 * \frac{v_b - v_0}{f} + b_3\right)}{c_1 * \frac{u_0 - u_b}{f} + c_2 * \frac{v_b - v_0}{f} + c_3} + Y_S \\ Z_b = Z_g \end{cases} \tag{6}$$

where (u_b, v_b) are the pixel coordinates of foot-point b, X_S, Y_S, and Z_S are the line elements in the translation vector $\boldsymbol{T}$, and a_i, b_i, and c_i (i = 1, 2, 3) are the elements in rotation matrix $\boldsymbol{R}$. The reprojection error $(\Delta u_t, \Delta v_t)$ of point t can then be calculated based on the inherent condition that the pedestrian is always perpendicular to the ground. The pixel coordinates are computed using the expressions:

$$\sqrt{(X_b - X_t)^2 + (Y_b - Y_t)^2} = 0 \tag{7}$$

$$\begin{bmatrix} \Delta u_t \\ \Delta v_t \end{bmatrix} = \begin{bmatrix} u_t \\ v_t \end{bmatrix} - \begin{bmatrix} u_t^0 \\ v_t^0 \end{bmatrix} = \begin{bmatrix} \frac{f(a_1 X_t + b_1 Y_t + c_1 Z_t + X_S)}{a_3 X_t + b_3 Y_t + c_3 Z_t + Z_S} \\ \frac{f(a_2 X_t + b_2 Y_t + c_2 Z_t + Y_S)}{a_3 X_t + b_3 Y_t + c_3 Z_t + Z_S} \end{bmatrix} + \begin{bmatrix} u_0 \\ v_0 \end{bmatrix} - \begin{bmatrix} u_t^0 \\ v_t^0 \end{bmatrix} \tag{8}$$

where (u_t, v_t) are the pixel coordinates of the head point t by central projection, and (u_t^0, v_t^0) are the pixel coordinates of point t that were directly detected. Based on Taylor's theorem, if $\boldsymbol{P} = [X,\ Y,\ Z]$, then

$$F(\boldsymbol{P}) \approx F(\boldsymbol{P}_0) + \nabla F(\boldsymbol{P}_0)^{\mathrm{T}} (\boldsymbol{P} - \boldsymbol{P}_0) \tag{9}$$

Accordingly, the reprojection error in Equation (8) is sorted into an equation of Z_t given by the formula:

$$\nabla F(\boldsymbol{P}) = \frac{\partial F}{\partial \boldsymbol{P}} = \begin{bmatrix} \frac{\partial u_t}{\partial Z_t} & \frac{\partial v_t}{\partial Z_t} \end{bmatrix} \tag{10}$$

$$\begin{bmatrix} \frac{\partial \Delta u_t}{\partial Z_t} \\ \frac{\partial \Delta v_t}{\partial Z_t} \end{bmatrix} \Delta Z_t + \begin{bmatrix} \Delta u_t \\ \Delta v_t \end{bmatrix} = 0 \tag{11}$$

In Equation (11), matrix operation is performed to calculate the correction ΔZ_t. After several iterations, the head 3D coordinates (X_t, Y_t, Z_t) satisfying the precision requirements can then be obtained. After the 3D coordinates of the pedestrian's head and feet are obtained, the height h can be calculated using the formula:

$$h = \sqrt{(X_b - X_t)^2 + (Y_b - Y_t)^2 + (Z_b - Z_t)^2} \tag{12}$$

2.3. Social Distance Monitoring

The midpoint b at the bottom edge of the pedestrian bounding box is used as the reference point (i.e., feet location). After calculating the 3D position, the interpersonal distance in a homogeneous space is determined using the person-to-person Euclidean distance [3]. If the interpersonal distance is more than the threshold value r (here, r is set to one meter based on the epidemic prevention and control requirements), the social distancing requirement is satisfied.

$$(b_i, b_j, r) = \begin{cases} 0 & \text{if} \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2 + (Z_i - Z_j)^2} \le r \\ 1 & \text{if} \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2 + (Z_i - Z_j)^2} > r \end{cases} \tag{13}$$

2.4. Performance Evaluation

For quantitative evaluation, we used two indicators: average localization error (ALE) and average localization precision (ALP). ALE is the average Euclidean distance between the predicted positions of feature points and their real positions. Here, we used the pedestrians' heads and feet as feature points for reference. ALE reflects the stability and accuracy of the proposed method and is given by the expression:

$$\mathrm{ALE} = \frac{1}{n}\sum_{k=1}^{n}\sqrt{\left(\hat{X}_k - X_k\right)^2 + \left(\hat{Y}_k - Y_k\right)^2 + \left(\hat{Z}_k - Z_k\right)^2} \tag{14}$$

where n is the number of feature points, (X_k, Y_k, Z_k) are the predicted positions of feature points, and $(\hat{X}_k, \hat{Y}_k, \hat{Z}_k)$ are their real positions.

The ALP indicator, defined by Xiang et al. [19], was originally used in evaluating car classification. In order to ensure the detection effect of YOLO, images with relatively large number of pedestrians and less occlusion between pedestrians were selected from the dataset. Then, the localization errors of all pedestrians detected in the selected images are calculated. The probabilities of errors less than the threshold r were used as the ALP_r values in our proposed method. ALP_r can be calculated using the equation:

$$\mathrm{ALP}_r = \frac{1}{n}\sum_{k=1}^{n} P(k, r) \quad r \in \{0.5,\ 1,\ 2\} \tag{15}$$

where $P(k, r)$ is the localization precision. The localization prediction is considered correct if the error between the predicted distance and the ground truth is within the given threshold.

3. Experiments and Results

For the experiments, we used two test data (i.e., CUMTB-Campus and KITTI) to evaluate the performance of our proposed method. First, the CUMTB-Campus dataset was used to show the positioning accuracy in different conditions. Second, the KITTI dataset was used to compare and contrast our proposed method against other localization methods.

3.1. Datasets

3.1.1. Self-Collected Data in CUMTB

Along the main road of CUMTB, we used a hand-held LiDAR scanner GeoSLAM-ZEB-HORIZON to collect 3D point clouds of the campus. In standard conditions, the measuring distance of this scanner is 100 m, and the relative accuracy is 1.5–3 cm. Figure 2 gives an example of our self-collected point clouds, which colored based on height. Video images were obtained simultaneously. The data includes major buildings such as teaching buildings, libraries, ethnic buildings, complex buildings, roads, and other facilities, covering about 80,000 square meters of the campus. At the same time, the video images of surveillance cameras widely distributed in the campus were obtained. In addition, non-metric camera was also used to take photos of pedestrians in some indoor and outdoor scenes on campus (such as teaching buildings, outside the library). Based on the above data, we built our self-collected dataset CUMTB-Campus. Specific parameters in standard conditions are shown in Table 1.

Table 1. Equipment parameters of two datasets.

Equipment Type	**CUMTB-Campus**		**KITTI**	
	Canon EOS 50D	**GeoSLAM-ZEB-HORIZON**	**FL2-14S3C-C**	**Velodyne 64-Wire 3D Laser Scanner**
Mode	Image	Point clouds	Image	Point clouds
Resolution	4752 × 3168	1.5–3 cm	1224 × 370	2 cm

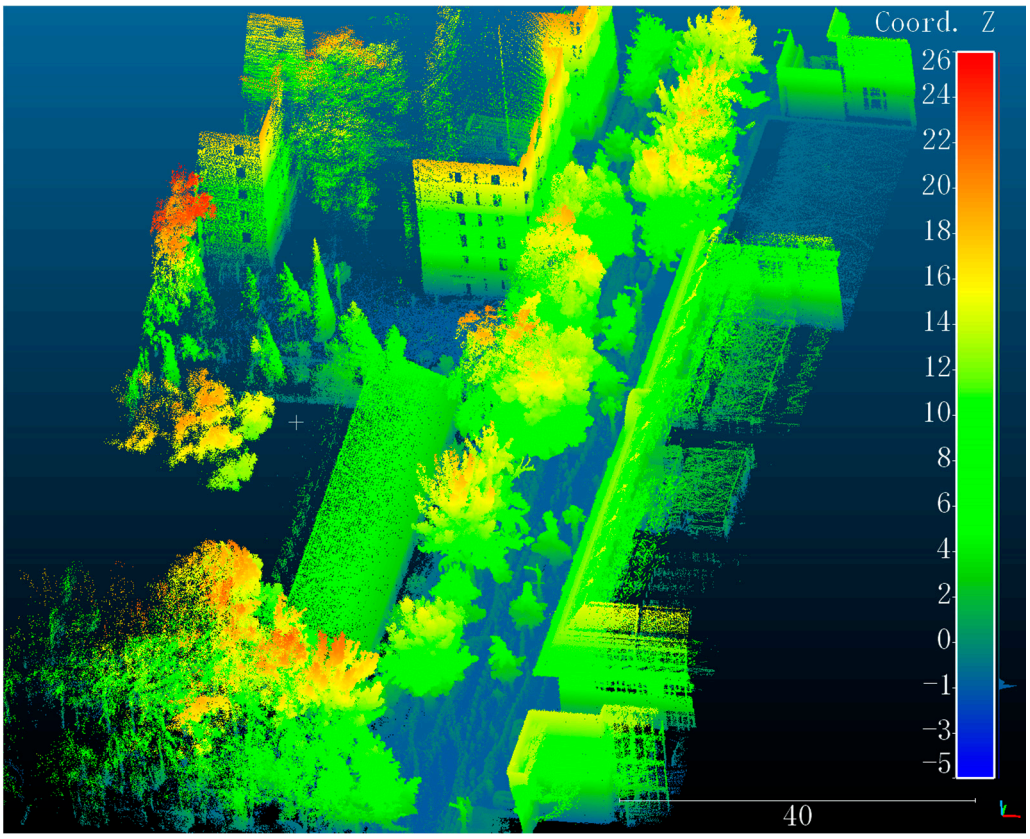


Figure 2. LiDAR point clouds colored based on height in CUMTB-Campus.

3.1.2. The KITTI Dataset

As the largest dataset for evaluating computer vision algorithms in autonomous driving scenarios, KITTI [20] contains real image data from urban, rural, and highway scenes, with up to 15 cars and 30 pedestrians per image. The dataset also has varying degrees of occlusion and truncation. In this paper, several unobstructed images of pedestrians and corresponding point clouds were selected for testing. Table 1 summarizes the device parameters of the datasets used in the study.

3.2. Implementation Details and Results

3.2.1. Implementation Details and Results for the CUMTB-Campus Dataset

The CUMTB-Campus experiment was designed based on the process described in Section 2. To calibrate the camera, non-coplanar feature points were selected from the image and point clouds, respectively. Using these feature points, the camera elements on interior and exterior orientation were obtained. Here we select the data of two different scenarios to show our positioning results.

1. Scene 1:

Scene 1 was an indoor scenario in the teaching building of CUMTB. The image with 15.1 megapixel resolution was taken by Canon EOS 50D on 24 October 2019, while the GeoSLAM-ZEB-HORIZON scanner acquired LiDAR point clouds along the corridor in the same region. We extracted the coordinates for the vertical direction of the ground plane where the pedestrian is situated from the point clouds. Then, as shown in Figure 3, the YOLO algorithm was used to obtain the pixel coordinates for each pedestrian detection box.



Figure 3. Scene 1 in CUMTB-Campus.

Camera calibration is realized based on Section 2.1.1. As shown in Figure 3, in order to facilitate the selection of feature points, we stick a number of reflectors in the scene. Here, the center points of eight uniformly distributed reflectors are manually selected from the point clouds and monocular image as matching points (represented by red five-pointed stars). The 2D pixel coordinates and 3D coordinates of matching points are read and substituted into Equation (3) to calculate the internal and external parameters of the monocular image. The results are as follows: f = 3938.104 pixels, the pixel coordinates of the principal point (u_0, v_0) are (2386.167, 1596.880), the translation vector $\boldsymbol{T} = (3.563, 5.210, 2.058)$ meters, the rotation matrix $\boldsymbol{R} = \begin{bmatrix} 0.167 & -0.985 & 0.050 \\ 0.058 & 0.041 & 0.997 \\ -0.984 & -0.170 & -0.050 \end{bmatrix}$. Based on the obtained parameters, test points are selected to verify the calibration accuracy. Taking the yellow four-pointed star as an example, its reprojected pixel coordinates are (2064.925, 2392.275) and the reprojection errors are 3.079 pixels.

Then using our proposed approach, each pedestrian's 3D position and height can be obtained using a single RGB image. Table 2 shows pedestrians' localization and height errors with different distances from the camera in Scene 1 of the CUMTB-Campus. The social distance errors between adjacent pedestrians are also listed.

Table 2. Pedestrian localization and height errors of Scene 1 in CUMTB-Campus.

Pedestrian Number	Distance (m)	Localization Error (m)	Height Error (m)	Social Distance			
				Adjacent Pedestrian	Truth Value (m)	Calculated Value (m)	Error (m)
1	10.203	0.320	0.054	—	—	—	—
2	14.271	0.242	0.007	1~2	4.832	4.770	−0.062
3	24.421	0.133	0.229	2~3	10.266	10.199	−0.067
4	36.563	0.229	0.071	3~4	13.122	13.209	0.087

In Figure 4, the left and right vertical axes represent the localization and height errors of four pedestrians in Scene 1 at different distances to the camera. The experimental results show that the localization error was less than half a meter within the 36-m vision range.

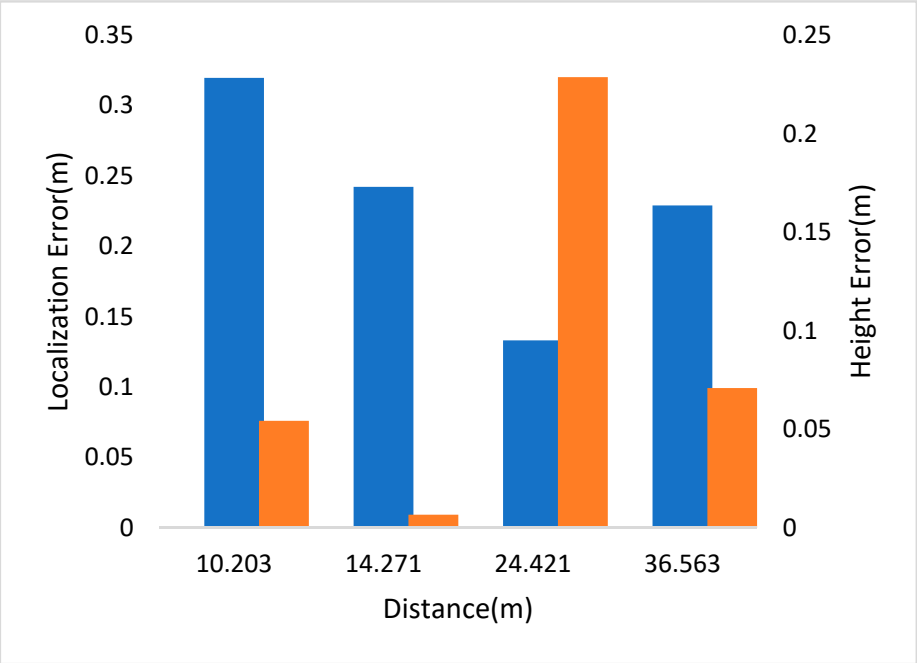


Figure 4. Localization and height error of Scene 1 in CUMTB-Campus.

2. Scene 2:

Scene 2 was an outdoor scene on the main road besides the library of the CUMTB. The image with 8.3 megapixel resolution was taken by GeoSLAM-ZEB-HORIZON scanner on 22 October 2019. The scanner also acquired LiDAR point clouds along the same road. There are more people in this scene, and they are in different positions including people overlapping. There are also several pedestrians on the steps. As shown in Figure 5, the YOLO algorithm was used to obtain the pixel coordinates for each pedestrian detection box.

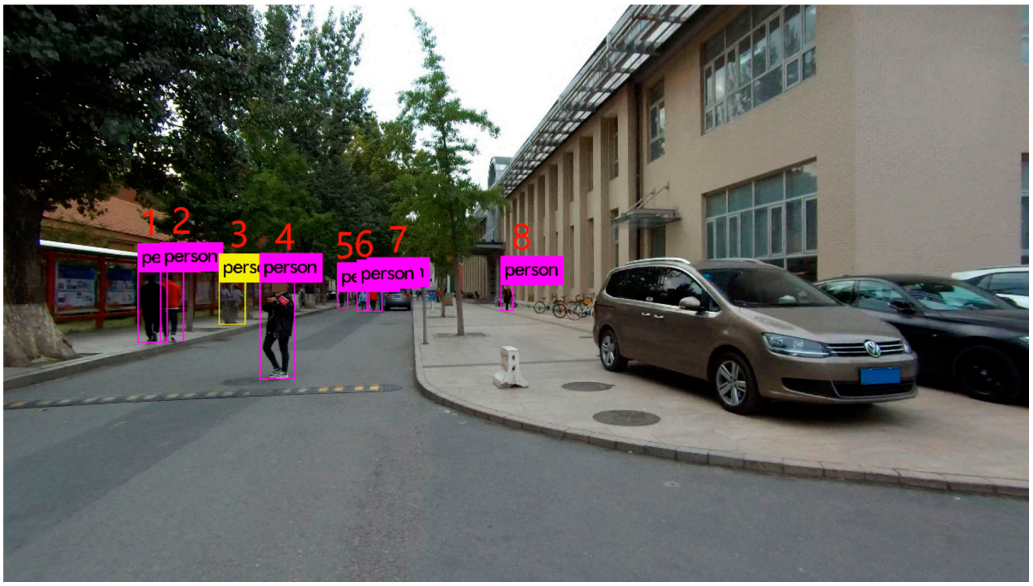


Figure 5. Scene 2 in CUMTB-Campus.

As can be seen from Figure 5, there are two pedestrians walking side by side and partially covered, and our detection algorithm only obtains one detection box (the yellow one). In order to evaluate the positioning accuracy rather than the 2D detection accuracy, the detection box is treated as a pedestrian for 3D localization. Then, using the above method, we can obtain the 3D position and height of pedestrians. Table 3 shows the localization and height errors of eight pedestrians in Scene 2. Similarly, we also list the social distance errors between adjacent pedestrians.

Table 3. Pedestrian localization and height errors of Scene 2 in CUMTB-Campus.

Pedestrian Number	Distance (m)	Localization Error (m)	Height Error (m)	Social Distance			
				Adjacent Pedestrian	Truth Value (m)	Calculated Value (m)	Error (m)
1	18.257	0.185	0.058	—	—	—	—
2	18.602	0.116	0.043	1~2	0.682	0.589	−0.093
3	25.197	0.476	0.141	2~3	8.059	8.176	0.117
4	11.238	0.093	0.017	3~4	14.709	14.548	−0.161
5	51.072	0.517	0.139	4~5	41.132	40.925	−0.207
6	42.151	0.341	0.114	5~6	9.171	9.367	0.196
7	41.984	0.392	0.128	6~7	0.676	0.558	−0.118
8	36.879	0.271	0.122	7~8	6.205	6.112	−0.093

In Figure 6, the left and right vertical axes represent the localization and height errors of eight pedestrians in Scene 2 at different positions. In this scenario, pedestrian 5 has the largest localization error because he is more than 50 m away from the camera. The second one is the yellow box due to the occlusion. The maximum relative error of social distance is 0.207 m.

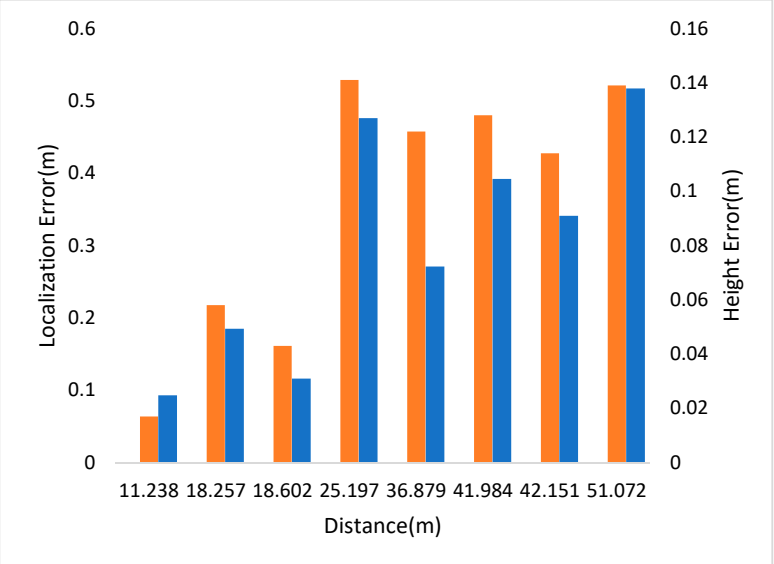


Figure 6. Localization and height error of Scene 2 in CUMTB-Campus.

3.2.2. Implementation Details and Results for the KITTI Dataset

From the KITTI dataset, we randomly selected multiple images of fully visible pedestrians. Using the camera calibration parameters provided by the dataset, the three-dimensional coordinates and height of pedestrians were obtained. Figure 7 shows Scene 3 selected from the KITTI dataset and nine pedestrian localization results. Table 4 summarizes the localization errors of the pedestrians' feet for Scene 3 and their height and social distancing errors.

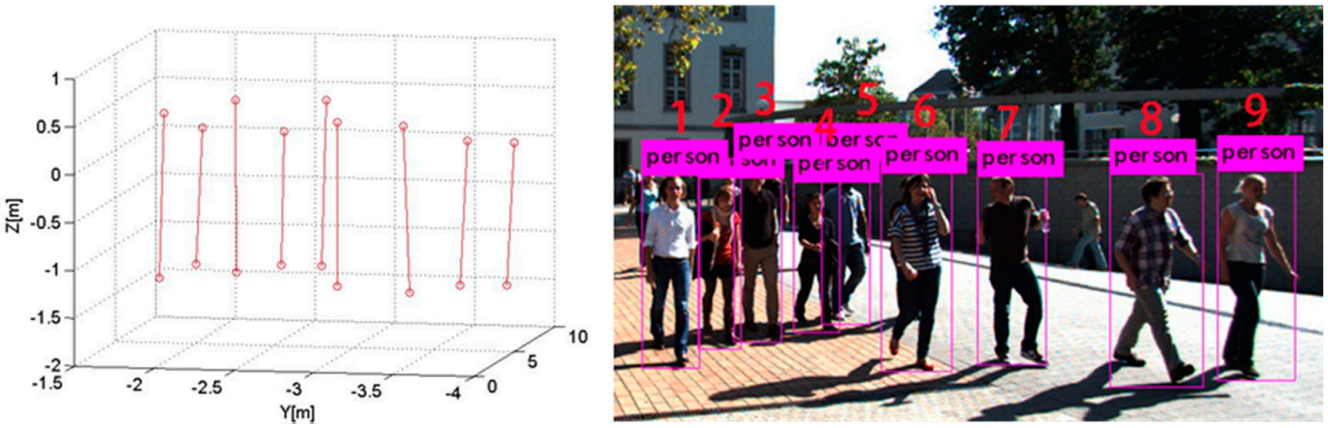


Figure 7. Scene 3 and pedestrian localization results.

Table 4. Pedestrian localization and height errors of Scene 3 in KITTI.

Pedestrian Number	Distance (m)	Localization Error (m)	Height Error (m)	Social Distance			
				Adjacent Pedestrian	Truth Value (m)	Calculated Value (m)	Error (m)
1	7.749	0.295	0.017	—	—	—	—
2	8.490	0.775	0.147	1~2	0.868	0.688	−0.180
3	8.779	0.193	0.012	2~3	0.632	0.860	0.228
4	9.048	0.688	0.082	3~4	0.683	0.585	−0.098
5	9.496	0.187	0.042	4~5	0.606	0.852	0.246
6	7.588	0.092	0.066	5~6	2.110	2.007	−0.103
7	7.687	0.066	0.012	6~7	0.929	0.930	0.001
8	7.651	0.581	0.159	7~8	1.102	1.216	0.114
9	8.006	0.731	0.150	8~9	0.677	0.643	−0.034

In Figure 8, the left and right vertical axes represent the localization and height errors of nine pedestrians in Scene 3. Figure 8 shows that the biggest localization and height errors of Scene 3 were 0.775 m and 0.159 m. Among them, pedestrian 2 and pedestrian 4 are affected by other pedestrians, leading to a large error in two-dimensional detection results, affecting the positioning accuracy. The maximum relative error of social distance is 0.246 m.

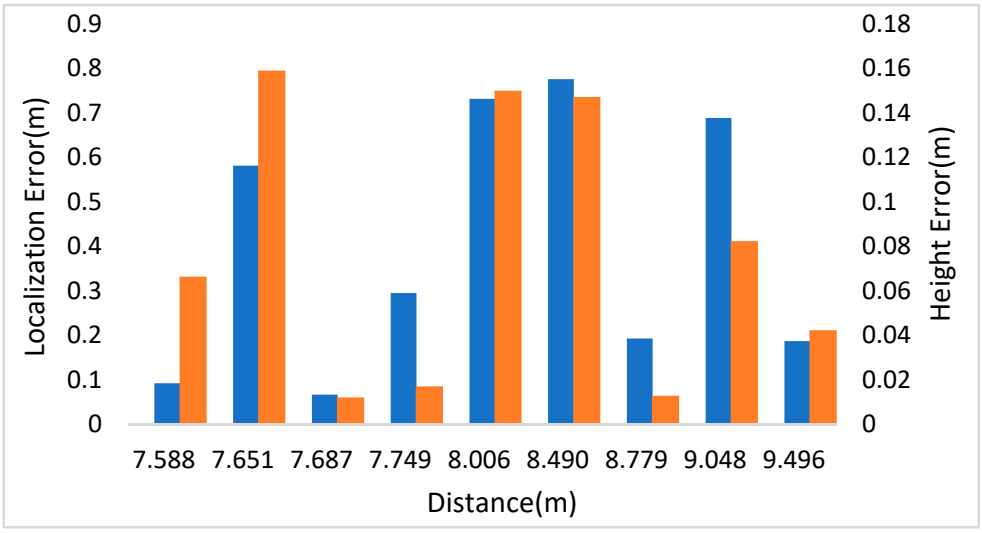


Figure 8. Localization and height error of Scene 3 in KITTI.

3.3. Methods Comparison

We then compared our positioning results with three monocular 3D localization methods based on the KITTI dataset, i.e., Mono3D [21], MonoDepth [22], and Monoloco [15]. We also evaluated our results with a binocular approach, i.e., 3DOP [23]. Its input is a stereo image pair, which uses 3D information and a context model specific to the autonomous driving field to obtain a 3D bounding box. As shown in Table 5, we evaluate our method with other localization results in [15]. Due to limitations of our implementation method, five images (including Scene 3) were randomly selected from the KITTI dataset containing 25 fully visible pedestrians at different distances. The 3D coordinates of 50 pedestrian points (each pedestrian contains two points, i.e., head and feet) were then calculated, and the accuracy was evaluated using the ALP_r(r = 0.5, 1, 2) and ALE mentioned above.

Table 5. Comparing our proposed method against baseline results on the KITTI dataset.

Method	Type	ALP_r (%)			ALE (m)
		r = 0.5 m	r = 1 m	r = 2 m	Fully Visible
Mono3D	Mono	13.2	23.2	38.9	2.32
MonoDepth + PifPaf [24]	Mono	20.5	35.3	50.6	1.69
Monoloco-trained on KITTI	Mono	29.0	49.6	71.2	0.98
3DOP	Stereo	41.4	54.9	63.2	0.71
Our proposed method	Mono	78.0	100.0	100.0	0.20

As shown in Table 5, the average localization error of our proposed method in the test data was within one meter, i.e., 0.20 m. Among them, 78% of the localization errors were within 0.5 m. With the aid of high-precision point clouds, our proposed approach outperformed other monocular localization methods, such as the Monoloco [15]. This method uses detected 2D joints of a network as input and generates the 3D location with confidence interval. However, this method assumes that all pedestrains have the same height, resulting in error-prone estimates. Then, our proposed method achieved comparable performance with the 3DOP, which uses stereo images for training and testing. However, our method does not require stereo images to obtain better accuracy and does not require datasets for training. In addition, our proposed method was able to calculate pedestrians' height. In the selected images, the average height error of twenty-five pedestrians was 0.05 m.

3.4. Social Distance Monitoring

Based on COVID-19 prevention and control protocols, one-meter social distancing should be observed between pedestrians. However, this rule on social distancing is not considered absolute. For instance, if adults and children walk side by side, the rule on social distancing is bypassed even if the distance between them is less than one meter. To account for this special circumstance, an exclusion scheme for parent and child pedestrians, determined based on height, is adopted. We adopt the standard of free admission for children in most public places (such as tourist attractions, amusement parks, cinemas, etc.) in China and choose 1.2 m as the reference height for children. At the same time, according to the relevant statistical data [25], the average height of 1.715 m was selected as the adult height reference. So in the proposed scheme, pedestrians walking side by side with a height difference of more than 51.5 cm would be considered parent and child.

Since our proposed localization method includes 3D location determination and height estimation, the parent-and-child exemption could be executed. In Figures 9 and 10, the 3D position and height of pedestrians are marked in different colors. Red is used to mark violations to social distancing rules and green pertains to adherence to social distancing protocols.

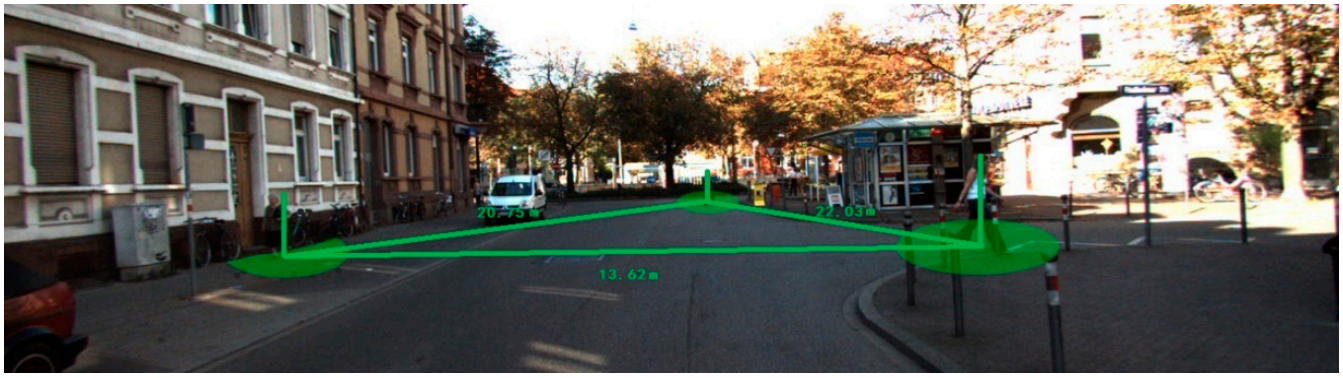


Figure 9. An example scene showing acceptable social distancing.



Figure 10. An example scene with various social distancing types.

4. Discussion

This study investigated the monocular pedestrian 3D localization for social distance monitoring. According to the experimental results in Section 3, the pedestrian localization and height errors obtained by the proposed method are mainly affected by the following factors:

(1). The quality of point clouds:

Mobile mapping systems (MMS) devices are used to acquire the point clouds in both tests (CUMTB-Campus and KITTI). When there are pedestrians or other moving objects in the scenes, the integrity of point clouds may be affected. In addition, the observation accuracy of LiDAR scanner is also related to the scanning distance and the fine degree of scanning. In the process of camera calibration, 3D coordinates of feature points are obtained from point clouds. In addition, point cloud segmentation is needed to obtain the vertical coordinate of the ground plane. Therefore, point cloud accuracy will have a certain impact on camera calibration and localization accuracy. However, in the proposed method, this factor has little influence on the final result.

(2). Select the corresponding point pairs:

As described in Section 2.1.1, at least six pairs of matching points should be selected from the point clouds and monocular image for DLT to obtain the camera parameters. As shown in Figure 3, eight evenly distributed feature points are selected around pedestrians. Then, the number and location distribution of the selected points will have a certain impact on the calibration results.

(3). Distance from pedestrian to the projection center of monocular image:

In the indoor scene shown in Figure 3, the distance differences between the pedestrians and the camera are not too large. The final result is mainly affected by other factors such as

calibration accuracy and research environment. Thus, the localization and height errors are not linearly connected to the distances amplitude. As shown in Figure 5, some pedestrians are far from the camera. The YOLO algorithm may fail to obtain the pedestrian detection box or the coordinates of the detection box are inaccurate. In addition, the actual distance corresponding to a pixel is larger. In this case, the accuracy is mainly affected by distance, and the localization and height errors are approximately linearly related to distance.

(4). Environmental situation:

Our method may be influenced by the environmental situation, such as whether there are steps or slopes in the scene and whether pedestrians are blocked or truncated. As shown in Figure 5, it can be seen that for pedestrians in different height ground planes, the proposed method can still achieve pedestrian 3D localization. However, since our approach requires the assistance of terrestrial point clouds, it is difficult to extract the ground plane of pedestrians if there are slopes or steps with large inclination in the scene. In addition, by analyzing the experimental results, it can be found that when pedestrians are blocked or truncated, the YOLO algorithm may fail to obtain accurate detection box coordinates. At this time, the localization error mainly comes from 2D detection. The use of additional auxiliary information may help solve this problem. For example, Gunel et al. [11] confirmed that human height could be inferred from images by obtaining joint posture, bone length ratio, and facial features. Kundegorski and Breckon [26] used infrared images for pedestrian tracking and obtained 3D locations based on the medical statistics assuming adult height.

To sum up, the accuracy of our proposed method relies on some external factors. During data collection, the quality of point clouds should be ensured as far as possible, and a certain number of uniformly distributed matching points can improve the localization accuracy. In general, our method is a simple and accurate monocular pedestrian 3D localization and social distance monitoring solution. The proposed approach can be extended and applied to other fields, such as automatic piloting, crowd analysis, fire safety, and merchant passenger flow statistics.

5. Conclusions

An innovative and accurate approach using monocular RGB images to obtain pedestrian localization and assess social distancing is proposed in this paper. The proposed method takes advantage of the inherent condition that the pedestrian is always perpendicular to the ground plane. Our method improves on the current localization in various aspects. First, our method only needs to obtain the pedestrian detection box without extracting the skeleton. Second, it does not rely on the base station, network or tag constraints, and can obtain the pedestrian's 3D position based on monocular image and terrestrial point clouds, and the precision meets the social distancing protocol of one meter. Finally, since the proposed localization approach also generates accurate height values, exclusionary schemes to social distancing protocols, particularly the parent-child exemption, can be introduced in the framework. Using height difference between pedestrians, the proposed approach can identify violations of social distancing rules and determine cases most likely covered by the parent-child exemption. During the epidemic prevention and control period, for some places requiring key prevention and control, such as schools, hospitals, shopping malls, LiDAR scanners can be used to obtain their 3D point clouds. During data collection, only the point clouds on the ground containing the pedestrian accessible area need to be scanned. Then, using widely distributed surveillance cameras, 3D positioning of pedestrians in the range of view can be obtained. In future works, we will continue to study the pedestrian 3D localization when they are blocked or truncated and try to solve this problem by obtaining the pedestrian body proportions through skeleton detection. While social distancing is important to minimize community transmission of the virus, a moderate amount of social interaction is also necessary. Using statistical analysis of the spatial-temporal data of pedestrians, high-risk areas for virus transmission and community

spread can be identified. This would help authorities reevaluate pedestrian traffic design and implement more proactive measures to mitigate risks.

Author Contributions: Z.X. and Y.N. contributed to the study's conception and design. Y.N. designed the experiments, analyzed the data, and wrote the first draft. Z.X. proposed key ideas and gave suggestions for the manuscript. E.X., G.L. and Y.H. participated in the data curation. W.S. gave suggestions for modifications to the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the National Natural Science Foundation of China, grant number 41701534, in part by the Open Fund of State Key Laboratory of Coal Resources and Safe Mining under Grant SKLCRSM19KFA01, in part by the State Key Laboratory of Geohazard Prevention and Geoenvironment Protection under Grant SKLGP2019K015, and in part by the China Fundamental Research Funds for Central University under Grant 2020YQDC02 and 2020YJSDC27.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Open-source dataset KITTI is available from Karlsruhe Institute of Technology at http://www.cvlibs.net/datasets/kitti/raw_data.php accessed on 25 May 2021. Self-collected dataset CUMTB-Campus that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: We would like to thank Xuesheng Zhao, and Jun Li of College of Geoscience and Surveying Engineering, China University of Mining and Technology (Beijing) for giving some comments and writing suggestions. We would also like to thank the anonymous reviewers and the editors for their comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kissler, S.M.; Tedijanto, C.; Goldstein, E.; Grad, Y.H.; Lipsitch, M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **2020**, *368*, 860–868. [CrossRef] [PubMed]
2. Chu, D.K.; Akl, E.A.; Duda, S.; Solo, K.; Yaacoub, S.; Schünemann, H.J.; Chu, D.K.; Akl, E.A.; El-harakeh, A.; Bognanni, A.; et al. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: A systematic review and meta-analysis. *Lancet* **2020**, *395*, 1973–1987. [CrossRef]
3. Rezaei, M.; Azarmi, M. DeepSOCIAL: Social Distancing Monitoring and Infection Risk Assessment in COVID-19 Pandemic. *Appl. Sci.* **2020**, *10*, 7514. [CrossRef]
4. Jhunjhunwala, A. Role of Telecom Network to Manage COVID-19 in India: Aarogya Setu. *Trans. Indian Natl. Acad. Eng.* **2020**, *5*, 157–161. [CrossRef]
5. Guo, S.; Niu, G.; Wang, Z.; Pun, M.O.; Yang, K. An Indoor Knowledge Graph Framework for Efficient Pedestrian Localization. *IEEE Sens. J.* **2021**, *21*, 5151–5163. [CrossRef]
6. Levchev, P.; Krishnan, M.N.; Yu, C.R.; Menke, J.; Zakhor, A. Simultaneous Fingerprinting and Mapping for Multimodal Image and WiFi Indoor Positioning. In Proceedings of the 2014 International Conference on Indoor Positioning and Indoor Navigation, Busan, Korea, 27–30 October 2014; pp. 442–450.
7. Liang, J.Z.; Corso, N.; Turner, E.; Zakhor, A. Image Based Localization in Indoor Environments. In Proceedings of the 2013 Fourth International Conference on Computing for Geospatial Research and Application, San Jose, CA, USA, 22–24 July 2013; pp. 70–75.
8. Zhang, R.; Candra, S.A.; Vetter, K.; Zakhor, A. Sensor Fusion for Semantic Segmentation of Urban Scenes. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation, Seattle, WA, USA, 26–30 May 2015; pp. 1850–1857.
9. Mousavian, A.; Anguelov, D.; Flynn, J.; Kosecka, J. 3D Bounding Box Estimation Using Deep Learning and Geometry. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 5632–5640. [CrossRef]
10. Zhang, S.; Wang, C.; He, Z.; Li, Q.; Lin, X.; Li, X.; Zhang, J.; Yang, C.; Li, J. Vehicle global 6-DoF pose estimation under traffic surveillance camera. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 114–128. [CrossRef]
11. Gunel, S.; Rhodin, H.; Fua, P. What Face and Body Shapes Can Tell Us About Height. In Proceedings of the 2019 IEEE/Cvf International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019; pp. 1819–1827. [CrossRef]
12. Das, S.; Meher, S. Automatic Extraction of Height and Stride parameters for Human Recognition. In Proceedings of the 2013 Students Conference on Engineering and Systems, Allahabad, India, 12–14 April 2013.
13. Bieler, D.; Günel, S.G.; Fua, P.; Rhodin, H. Gravity as a Reference for Estimating a Person's Height From Video. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.

14. Deng, W.; Bertoni, L.; Kreiss, S.; Alahi, A. Joint Human Pose Estimation and Stereo 3D Localization. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 2324–2330.
15. Bertoni, L.; Kreiss, S.; Alahi, A. Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6861–6871.
16. Bertoni, L.; Kreiss, S.; Alahi, A. Perceiving Humans: From Monocular 3D Localization to Social Distancing. *IEEE Trans. Intell. Transp. Syst.* **2021**. [CrossRef]
17. Abdel-Aziz, Y.I.; Karara, H.; Hauck, M. Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. *Photogramm. Eng. Remote Sens.* **2015**, *81*, 103–107. [CrossRef]
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788. [CrossRef]
19. Xiang, Y.; Choi, W.G.; Lin, Y.Q.; Savarese, S. Data-Driven 3D Voxel Patterns for Object Category Recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1903–1911.
20. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
21. Chen, X.Z.; Kundu, K.; Zhang, Z.Y.; Ma, H.M.; Fidler, S.; Urtasun, R. Monocular 3D Object Detection for Autonomous Driving. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2147–2156. [CrossRef]
22. Godard, C.; Mac Aodha, O.; Brostow, G.J. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 6602–6611. [CrossRef]
23. Chen, X.Z.; Kundu, K.; Zhu, Y.K.; Berneshawi, A.; Ma, H.M.; Fidler, S.; Urtasun, R. 3D Object Proposals for Accurate Object Class Detection. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Neural Information Processing Systems (Nips): La Jolla, CA, USA, 2015; Volume 28.
24. Kreiss, S.; Bertoni, L.; Alahi, A. PifPaf: Composite Fields for Human Pose Estimation. In Proceedings of the 2019 IEEE/Cvf Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11969–11978. [CrossRef]
25. Visscher, P.M. Sizing up human height variation. *Nat. Genet.* **2008**, *40*, 489–490. [CrossRef] [PubMed]
26. Kundegorski, M.E.; Breckon, T.P. A Photogrammetric Approach for Real-time 3D Localization and Tracking of Pedestrians in Monocular Infrared Imagery. In *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence X; and Optical Materials and Biomaterials in Security and Defence Systems Technology Xi*; Burgess, D., Owen, G., Rana, H., Zamboni, R., Kajzar, F., Szep, A.A., Eds.; International Society for Optics and Photonicsing: Bellingham, WA, USA, 2014; Volume 9253, p. 92530I.