applied sciences

MDPI

Article

Method for 2D-3D Registration under Inverse Depth and Structural Semantic Constraints for Digital Twin City

Xiaofei Hu, Yang Zhou * and Qunshan Shi

PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China
* Correspondence: zhouyang3d@163.com

Abstract: A digital twin city maps a virtual three-dimensional (3D) city model to the geographic information system, constructs a virtual world, and integrates real sensor data to achieve the purpose of virtual–real fusion. Focusing on the accuracy problem of vision sensor registration in the virtual digital twin city scene, this study proposes a 2D-3D registration method under inverse depth and structural semantic constraints. First, perspective and inverse depth images of the virtual scene were obtained by using perspective view and inverse-depth nascent technology, and then the structural semantic features were extracted by the two-line minimal solution set method. A simultaneous matching and pose estimation method under inverse depth and structural semantic constraints was proposed to achieve the 2D-3D registration of real images and virtual scenes. The experimental results show that the proposed method can effectively optimize the initial vision sensor pose and achieve high-precision registration in the digital twin scene, and the Z-coordinate error is reduced by 45%. An application experiment of monocular image multi-object spatial positioning was designed, which proved the practicability of this method, and the influence of model data error on registration accuracy was analyzed.

Keywords: digital twin city; 2D-3D registration; virtual–real fusion; inverse depth; structural semantics

Citation: Hu, X.; Zhou, Y.; Shi, Q. Method for 2D-3D Registration under Inverse Depth and Structural Semantic Constraints for Digital Twin City. *Appl. Sci.* **2022**, *12*, 8543. https://doi.org/10.3390/app12178543

Academic Editor: Rubén Usamentiaga

Received: 27 July 2022
Accepted: 24 August 2022
Published: 26 August 2022

1. Introduction

A digital twin city rebuilds a corresponding virtual city in cyberspace by digitizing all elements of a physical city, forming a pattern of coexistence and blending of a physical city and digital city in the information dimension [1]. Digital twin cities are widely used in urban management, planning, and navigation. In recent years, with the maturity of tilt photography and unmanned aerial vehicle (UAV) technology, the local orthophoto, tilt, or LIDAR point cloud data of the city can be accurately obtained, and the urban entity can be comprehensively measured from the land and sky so as to obtain a virtual 3D city model that is very similar to the real world [2]. The digital twin city maps the model to the geographic information system (GIS) system, builds a virtual world, and then integrates real sensor data, achieving the goal of virtual–real fusion. Among numerous sensor types, the vision sensor is one of the most common sensor devices in the digital twin city, such as surveillance cameras, mobile portable devices, and vehicles. Vision sensor integration occurs with the high-accuracy registration of a 2D image taken by a camera and the digital twin virtual 3D scene, which is one of the key technologies of the digital twin city.

The purpose of 2D-3D vision sensor registration is to estimate or optimize the 6-DOF pose of the vision sensor in a virtual digital twin scene and obtain the spatial position of any object in the image. Hence, 2D-3D registration is one of the key technologies for applications such as augmented reality (AR) [3] and Video GIS [4]. It has attracted extensive attention in recent years. Generally, 2D-3D registration methods are divided into hardware-based and vision-based registration methods [5]. The registration method based on hardware generally obtains the position and orientation of the equipment in space through GPS positioning, acceleration, and geomagnetic sensors, identifying the positioning of ground

objects. It is significantly affected by the sensor accuracy and outdoor environment, and the accuracy has difficulty meeting the requirements of 2D-3D registration in the outdoor environment. Vision-based registration technology uses the image obtained by the vision sensor to optimize the camera's 6-DOF pose and then achieves the fusion of the image and virtual scene. Because of seasonal and weather changes, in digital twin applications, sensor image and virtual scene appearances often vary significantly, which makes matching the virtual and real images difficult. Additionally, because a city scene image often contains a sky background and has a large depth range, its feature points are unevenly distributed in the vertical direction, resulting in inaccurate positioning. In addition, urban outdoor scenes have typical Manhattan-world [6] characteristics, containing rich structural semantic features. For more accurate and robust camera pose estimation, semantic features can be used as constraints. However, only a few works have applied structural features to the outdoor environment.

This study focused on the accuracy problem of vision sensor registration in the virtual digital twin city scene. Regarding the characteristics of the city scene with structural semantic information, we propose a 2D-3D registration method for the real image and virtual scene under inverse depth and structural semantic constraints. First, perspective view and inverse-depth nascent technology (PDNT) is used to obtain the perspective and inverse depth images of digital twin scenes, and the plumb line, which contains structural semantic information, is extracted from the vision sensor image. Based on the general vision-based framework, we also propose a simultaneous feature matching and pose estimation method utilizing inverse depth coordinate and structural semantic constraints (MP-IDSSC) to optimize the position and orientation of the vision sensor, which effectively solves the pose error caused by the uneven distribution of feature points. Finally, we use the ray intersection method to obtain the location of any object in the real image and achieve the registration of the vision sensor.

The main contributions of this study are as follows:

- We developed a 2D-3D registration method with constraints of structural semantics and inverse depth coordinates, which effectively solves the problem of the large error caused by the uneven distribution of feature points in the vertical direction and achieves high-accuracy registration of monocular images in the digital twin scene.
- The proposed method seamlessly integrates with existing digital twin platforms. This method utilizes PDNT technology, which can be directly implemented using the basic functions of digital twin applications.

The remainder of this paper is arranged as follows: Section 2 outlines 2D-3D registration methods. Section 3 introduces the proposed method and framework, and Section 4 describes the experiments that we conducted. Finally, Section 5 summarizes the study and discusses potential future research directions.

2. Related Works

The 2D-3D registration process is different from traditional image registration because it consists of two parts: camera pose estimation and object positioning. The related work is introduced from these two aspects.

2.1. Pose Estimation

Pose estimation is used to estimate the camera's own position and orientation by registering the image acquired by the camera device with the virtual model data. Application scenarios can be divided into indoor and outdoor scenarios. Ma et al. [7] presented an indoor 2D-3D registration method based on scene recognition, which detects and recognizes the target scene in a video frame image to track targets and estimate the camera pose. Li et al. [8] developed a novel camera localization workflow based on a highly accurate 3D prior map optimized by the RGB-D SLAM method. Outdoor pose estimation is more difficult than indoor pose estimation. The indoor environment is small, so binocular cameras and RGB-D can be used because the distance is relatively close. However, the

outdoor area is larger, and the image appearance is highly affected by the environment. Wu et al. [9] presented a 3D registration method for mobile augmented reality of a building environment based on SURFREAK and KLT. Yue et al. [10] presented an image matching method for distorted buildings with automatic viewpoint correction and fusion. Huang et al. [11] proposed an outdoor AR system registration mechanism based on 3D GIS. Most of the existing methods make use of natural outdoor features, but they do not adequately mine the structural features of cities. Additionally, in practice, we find that because street view images often contain a sky background and have a large depth range, they have an uneven distribution of features in the vertical direction, resulting in errors in camera pose estimation in the vertical direction, which affects the camera pose accuracy.

From the view of technology, the traditional framework of the pose estimation system has three steps: feature extraction, feature matching, and pose calculation. The focus of research is feature extraction and matching. Presently, difficulties are encountered in practical application. First, the appearance of images often differs significantly with a change in season and weather. The key to solving this problem is having a local feature extraction method with high discrimination, such as SIFT [12]. To address the aforementioned shortcomings of traditional feature extraction methods, scholars have applied the convolution neural network (CNN) to image feature extraction, gradually evolved from using handcrafted and deep convolutional features, and successively proposed several methods, including SuperPoint [13] and D2-Net [14]. Deep convolutional features use high-level semantic information for feature point extraction, have strong generalizability, and ultimately show great potential in solving image matching in changing scenes. Second, repeated texture leads to difficulty in matching. There are many repeated texture features, especially in building targets, which place feature point matching requirements on the algorithm. The random sample consensus (RANSAC) algorithm is a widely used matching method, and, while its basic idea is simple and effective, there are two prerequisites for using it [15]: (1) The proportion of interior points must be at a high level because, if less than 50%, the RANSAC method requires a large number of iterations [16] and may fail or become very time-consuming [17]. (2) An assumed model must be given and then satisfied by the interior points [18]. In [19], the spatial constraints of the Delaunay triangulation were used to improve the robustness of feature point matching, but the method was time-consuming. Bian et al. [20] proposed a feature point extraction algorithm that could run in real time; however, its accuracy failed to meet localization requirements. Achieving good performance in challenging environments is difficult when using the existing mainstream feature point matching methods.

2.2. Object Positioning

Regarding the outdoor object positioning problem, the multi-view location method is mainstream and has been applied to a variety of scenes, such as with UAVs and handheld devices. With the popularization of UAV applications, using UAVs to achieve target localization has attracted some attention [21,22]. These types of methods usually require the vision sensor to move along a desired track and take multiple photographs or videos of the target; thereafter, it uses a multi-view method to achieve target localization. Zhang et al. [23] studied the yaw angular errors and relative height estimation problems of two flight scenarios (flyover and wandering) and identified 3D target geolocation. Using vehicle platforms to achieve short-range target localization in an outdoor environment [24–27] is a popular topic. Additionally, Tekaya et al. [28] introduced an algorithm to estimate the distance from a target in super stereo images using mobile devices, but it was only used for relative target localization. These methods essentially use other means to obtain high-precision camera poses and multi-view methods for target localization, requiring the camera to move in a fixed pattern. Moreover, in urban areas with many buildings, multi-view localization methods are used for target positioning but are limited in practical application because of the occlusion of buildings.

3. Materials and Methods

The framework for the 2D-3D registration of virtual and real images in the digital twin city is shown in Figure 1. The first step was to employ PDNT according to the initial positions and orientations to obtain the virtual perspective and inverse depth images of the digital twin scene. Then, the CNN feature point extraction method was adopted to improve the discrimination and consistency of feature extraction. The 3D point coordinate information in the inverse depth map was used to establish the 2D-2D-3D triplet correspondence, and semantic features, such as the plumb line, were extracted from the real image by the 2-line minimal solution set (MSS) method. Under the constraints of inverse depth coordinates and structural semantics, the MP-IDSSC method was proposed. This method improves the matching accuracy and corrects the pose of the vision sensor. Finally, the ray intersection method was used to achieve 3D registration between a real image and a virtual scene.

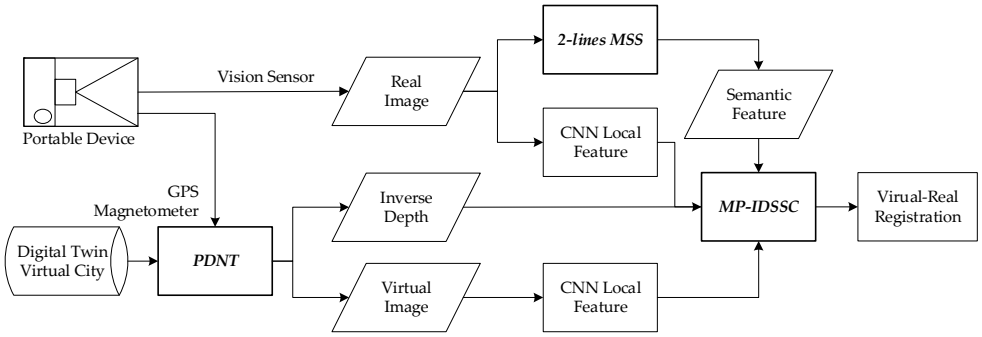


Figure 1. Flowchart of the algorithm used in the proposed method. PDNT is used to obtain virtual perspective and inverse depth images. The 2-line MSS method is used to extract structural semantic features. MP-IDSSC achieves 2D-3D registration between the real image and virtual scene.

3.1. PDNT: Perspective View and Inverse-Depth Image Nascent Technology

The position and orientation information obtained from portable devices was used as the initial values. The perspective images and inverse depth images of the virtual digital twin scene in the current view were obtained based on the spatial relationship between the scene and camera pose. The virtual perspective image was used to match the real image, and the inverse depth image provided 3D coordinates for each image point. The formula used to calculate these coordinates is:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = [Mat_m][Mat_c]\left[Mat_{proj}\right]\begin{bmatrix} u_0 \\ v_0 \\ w_0 \end{bmatrix} \quad (1)$$

where $\begin{bmatrix} u_0 \\ v_0 \\ w_0 \end{bmatrix}$ is the coordinate vector for the normalized image point, and $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ is the position vector of the calculated target. $[Mat_m]$ is the model transformation matrix, and $[Mat_c]$ is the transformation matrix from the world coordinate system to the camera coordinate system. $\left[Mat_{proj}\right]$ is the projection transformation matrix.

The significance of generating a virtual perspective image is to establish a triplet (2D-2D-3D) correspondence between the real, perspective, and inverse depth images. In addition, under the constraint of the initial pose, perspective images have a high spatial similarity to the real image, which reduces the difficulty in image matching caused by geometric differences. An inverse depth image stores the inverse depth coordinates of the current scene and can easily convert each feature point into world coordinates. Because the simulated orientation was accurate, the real-world coordinates of each pixel could

be calculated, and the sensor position and orientation could be corrected using the pose estimation method.

3.2. Structural Semantic Feature Extraction

Urban scenes often have typical Manhattan-world characteristics. The Manhattan world is the abbreviation for a structured scene that conforms to the Manhattan hypothesis. It has strong structural regularity; thus, it contains rich structural semantic features. The plumb line is a typical structural semantic feature contained in city scene images. As a type of global information, it reflects the spatial relationship between the camera coordinate system and the world coordinate system. An accurate and robust estimation of the camera pose is likely with the use of the semantic feature of the plumb line. This study first used the EDLines [29] algorithm to extract line features and then used the 2-line MSS [30] method constrained by prior information to extract plumb lines and, finally, the spatial characteristics of plumb lines as constraints in 2D-3D registration.

3.2.1. EDLines Line Feature Detection

Presently, the EDLines algorithm is one of the most popular line detection algorithms. Compared with the traditional line feature extraction methods, such as Hough [31] and LSD [32], EDLines can detect the line features in an image with high accuracy. The principle of the algorithm is to use the edge drawing algorithm to generate a clean and continuous edge pixel chain and then generate a straight-line segment based on the edge pixels. This algorithm is efficient and fast. In comparison, the detection speed of EDLines is 10 times higher than that of LSD.

3.2.2. Two-Line MSS Plumb LINE Extraction with Prior Constraint

The direction of gravitational acceleration is obtained with the help of inertial measurement elements or when the initial values of the position and attitude are known. The gravitational acceleration g_c in the camera coordinate system is obtained by using inertial measurement elements. Among the three orthogonal vanishing points in the Manhattan world, only the vertical vanishing point is parallel to the direction of gravitational acceleration. The vertical vanishing point provides an a priori constraint for the extraction of the plumb line. Using this prior constraint, we can first conduct a preliminary screening before comparing the evaluation value of the candidate hypothesis with the current optimal result. Because the angle between the vertical vanishing vector and vector of gravity is close to zero, the vector cross product is calculated, and the fault-tolerant threshold τ_1 is set as follows:

$$\left| min\left\{ arccos\left(\frac{l_i \times g_c}{\|g_c\|} \right) \right\} \right| < \tau_1,\ i = 1, 2, 3 \tag{2}$$

The candidate set of the plumb line can be screened out from the extracted line segments, and the vertical vanishing point can be calculated by using the 2-line MSS method. Assuming that the proportion of segments belonging to the vertical vanishing point in the candidate set of segments is 0.5, the probability of randomly selecting two segments belonging to the vertical vanishing point is $p = \mu 0.5^2$. μ is the regulating factor. At a confidence coefficient of 0.9999, the number of iterations required to obtain at least one inner 2-line MSS is:

$$it_c = log(1 - 0.9999)/log(1 - p) \tag{3}$$

Any it_c line segment pairs are selected, and the plane coordinates of the intersection image of the line segments are calculated and then converted to the equivalent spherical coordinates:

$$\begin{cases} \phi = acos\left(Z/\sqrt{X^2 + Y^2 + Z^2} \right) \\ \lambda = atan2(X, Y) + \pi \end{cases} \tag{4}$$

where (x, y) are the image plane coordinates of the endpoint of the line segment. f is the principal length of the camera. $\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$ are spherical coordinates:

$$\begin{cases} X = x - x_0 \\ Y = y - y_0 \\ Z = f \end{cases} \tag{5}$$

With candidate points, the optimal estimation of the position of the vertical vanishing point is calculated by using the threshold value τ_2, and the set of vertical line segments is extracted.

$$argmax \sum_{i=1}^{n} VP * l_i < \tau_2 \tag{6}$$

3.3. MP-IDSSC: Matching and Pose Estimation under Inverse Depth and Structural Semantic Constraints

The traditional vision system framework has three steps: feature extraction, feature matching, and pose estimation. Each step depends on the result of the previous step. In challenging conditions, there may not be enough matching point pairs for the pose estimation step. To solve this problem, we propose a simultaneous feature matching and pose estimation method under the constraints of inverse depth coordinates and structural semantics. Because all pixels in perspective images correspond to 3D coordinates in the map data, the correspondence between 3D point coordinates and 2D pixel coordinates in the target image can be established. The pose estimation algorithm can now be used directly to calculate the real image's pose, and the 3D point coordinate image plane projection error and plumb line error can be used to constrain feature point matching. Thus, the feature point matching and pose estimation algorithm are completed in one step, and the pose estimation no longer depends on the result of the feature matching, which improves the robustness of the algorithm.

3.3.1. Cross-Validation of Dynamic Adaptive Threshold

To improve the efficiency and accuracy of the algorithm, this study adopted the cross-validation method of a dynamic adaptive threshold based on the selected initial value. The fast library for approximate nearest neighbors (FLANN) algorithm was used to seek the matching point pairs for screening, which generally included a first matching point with the closest Euclidean distance and a second matching point with the second closest Euclidean distance. It is generally believed that, for a certain matching point pair to be screened, the smaller the distance dis of the closest matching point compared with the distance $dis^{'}$ of the second closest matching point, the better the matching quality. Traditional algorithms generally use a fixed scale factor t as the threshold. In particular, when $dis < t{\cdot}dis^{'}$ is satisfied, the matching point pair is used as a candidate. However, because of the radiation difference, the distribution range of the Euclidean distance difference between point features is unpredictable, and it is generally necessary to manually adjust the threshold t to select a better matching point pair. To solve this problem, we designed a cross-validation method for a dynamic adaptive threshold. First, the FLANN cross-nearest neighbor search was used to calculate the average distance difference between the closest and second closest matching point in the cross search.

$$dis_{avg} = \frac{\left(\sum_{i=1}^{N}\left(dis^{'} - dis\right)\right)}{N} \tag{7}$$

When the distance difference between the closest and second closest points is less than dis_{avg}, the matching point pair in the cross search is used as the initial candidate. Using the average distance difference as a comparison criterion can adapt to the distance

between feature points, retain high-quality matching points, and improve the stability of the RANSAC algorithm.

3.3.2. Fast P5P Method

The pose of the image was used to constrain image matching and incorporate the pose estimation algorithm into RANSAC. Therefore, it was necessary to improve the computing speed as much as possible while still ensuring estimation algorithm accuracy. However, the number of points involved in the computation must not be excessive. Therefore, this study adopted the P5P fast pose estimation method to solve the problem.

The mathematical relationship for the standard central projection is expressed as follows:

$$s\boldsymbol{u} = \boldsymbol{KPX} \tag{8}$$

where $\boldsymbol{K}$ is the intrinsic camera parameter matrix, and the intrinsic camera parameters can be obtained by calibration; s is an unknown scale parameter; $\boldsymbol{u} = \begin{bmatrix} u & v & 1 \end{bmatrix}$; and $\boldsymbol{P} = [\boldsymbol{R}|\boldsymbol{t}]$ is a 3×4 matrix that contains rotation and translation information, expressed as follows:

$$P = [\boldsymbol{R}|\boldsymbol{t}] = \begin{bmatrix} r_1 & r_2 & r_3 & t_1 \\ r_4 & r_5 & r_6 & t_2 \\ r_7 & r_8 & r_9 & t_3 \end{bmatrix} \tag{9}$$

Using ${u_i}^{\wedge} u_i = 0$ and eliminating the scale variable s, we obtain:

$$\begin{bmatrix} 0 & -1 & v \\ 1 & 0 & -u \\ -v & u & 0 \end{bmatrix} \begin{bmatrix} r_1 & r_2 & r_3 & t_1 \\ r_4 & r_5 & r_6 & t_2 \\ r_7 & r_8 & r_9 & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = 0 \tag{10}$$

Expanding the third row yields:

$$-v(r_1X + r_2Y + r_3Z + t_1) + u(r_4X + r_5Y + r_6Z + t_2) = 0 \tag{11}$$

where r_i *and* t_i are unknown variables to be determined. Each feature point provides a linear constraint equation; five points construct a 5×8 coefficient matrix. Let $w = [r_1, r_2, r_3, t_1, r_4, r_5, r_6, t_2,]^T$. Through the above system of linear equations, we can obtain a solution consisting of three non-zero basis vectors:

$$w = \lambda_1 n_1 + \lambda_2 n_2 + \lambda_3 n_3 \tag{12}$$

Let $\lambda_3 = 1$; based on the constraint that R is an orthogonal matrix, we can quickly obtain four groups of candidate solutions [33]. For each solution group, we can further solve $[r_7, r_8, r_9, t_3]^T$. From

$$\begin{cases} r_1r_7 + r_2r_8 + r_3r_9 = 0 \\ r_4r_7 + r_5r_8 + r_6r_9 = 0 \end{cases} \tag{13}$$

r_7 and r_8 can be linearly represented by r_9, indicating that only r_9 and t_3 are real unknowns. Using the first or second row of (10),

$$(r_1X + r_2Y + r_3Z + t_1) - u(r_7X + r_8Y + r_9Z + t_3) = 0 \tag{14}$$

We can construct a system of linear equations with a coefficient matrix of size 5×2 and finally obtain the optimal solution by minimizing the projection error.

$$\begin{bmatrix} \alpha_1 u_1 & u_1 \\ \vdots & \vdots \\ \alpha_i u_i & u_i \end{bmatrix} \begin{bmatrix} r_9 \\ t_3 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_i \end{bmatrix} \tag{15}$$

where $\alpha_i = \frac{r_2r_6-r_3r_5}{r_1r_5-r_2r_4}X_i + \frac{r_2r_6-r_3r_5}{r_1r_5-r_2r_4}Y_i + Z_i$, and $\beta_i = r_1X_i + r_2Y_i + r_3Z_i + t_1$.

Random point selection is a key part in this process. We adopted a uniform random sampling method for the image plane. First, the image plane was evenly divided into 2×2 grids, and each rectangle covered exactly one-quarter of the image. Thereafter, a region of equal size was placed at the center of the image, and the image plane was divided into regions named from A to E, as shown in Figure 2. During sampling, five points were randomly selected from these five regions.

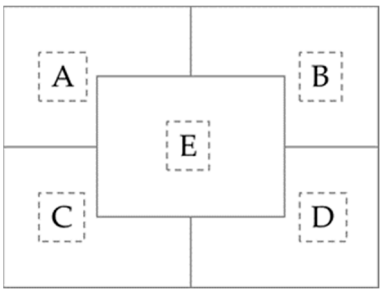


Figure 2. Random sampling distribution.

3.3.3. Error Function Combined with Structural Semantic Constraint

Because the plumb line contains special semantic information, it can be used as a constraint in 3D registration. For integrating point and plumb line features, the error function model of the image pose solution is:

$$\delta = \sum_{i=1}^{n} argmin\delta_{pi} + \delta_l \tag{16}$$

where δ_{pi} and δ_l are the point characteristic error and plumb line constraint error, respectively. The point characteristic error, δ_{pi}, consists of two parts: camera pose error and projection error. First, the camera pose error was calculated, and the projection error was only calculated when the error was less than the threshold. The optimal solution often has the largest number of interior points, and the core method for determining interior points was to calculate the projection error. For each feature point, when the calculated projection error was less than the threshold, it was recorded as an interior point. When the number of interior points was the largest, the pose parameter was optimal, and the corresponding point pair was the matching result. The calculation formula for the projection error is:

$$\sigma_p = \sum_{i=1}^{n} argmin \sum_{k=1}^{m} \|u_{ik} - KPX_i\|_2^2 < \varepsilon \tag{17}$$

where σ_p is the projection error, u_{ik} is the image point, and X_i represents the corresponding 3D point coordinates.

For the calculation of δ_l, the vertical line segment is mapped to the virtual perspective image by the feature mapping results, and then the vertical line segment in the image is mapped to the 3D model space according to the spatial relationship between the virtual and real images, as shown in Figure 3, to obtain the 3D vertical line segment. The error of the vertical line segment in 3D model space is used as a constraint, and its error is calculated by:

$$\sigma_l = \sum_{i=1}^{m} atan2\left(\sqrt{dX_i^2 + dY_i^2}, dZ_i\right) \tag{18}$$

where dX_i, dY_i, dZ_i are the coordinate difference between the two endpoints of the i-th 3D space segment in the direction of the three coordinate axes.

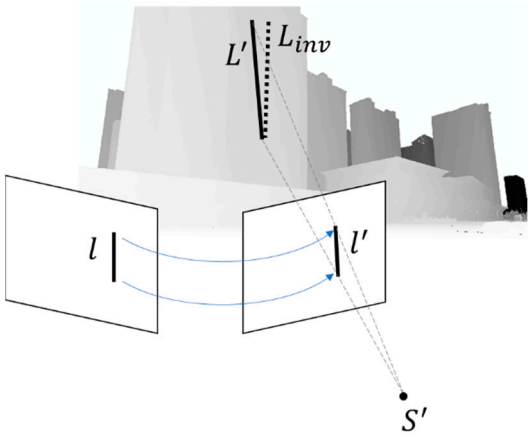


Figure 3. Mapping the vertical line segment to 3D model space. l is the vertical line segment extracted from the real image. l' is the corresponding segment in the virtual perspective image, which is mapped to L'. L_{inv} is the true plumb line of l; due to the existence of error, L' will not be a plumb line.

In the process of mapping the line segment on the image to the 3D model space, the mapping may fail, for example, by mapping the edge line of a building to the sky background without depth coordinates. To avoid this problem, this study adopted the random drift algorithm. After mapping the image line segment to the inverse depth space, it determines whether the inverse depth value of the two endpoints of the line segment is valid and whether the difference between the two is less than the threshold. If it is greater than the threshold, it indicates that there is a mapping failure, and the line segment must be randomly shifted by k pixels until a valid mapped line segment is obtained. Since there is often a plane near the edge line, the random offset does not affect the calculation of the final error.

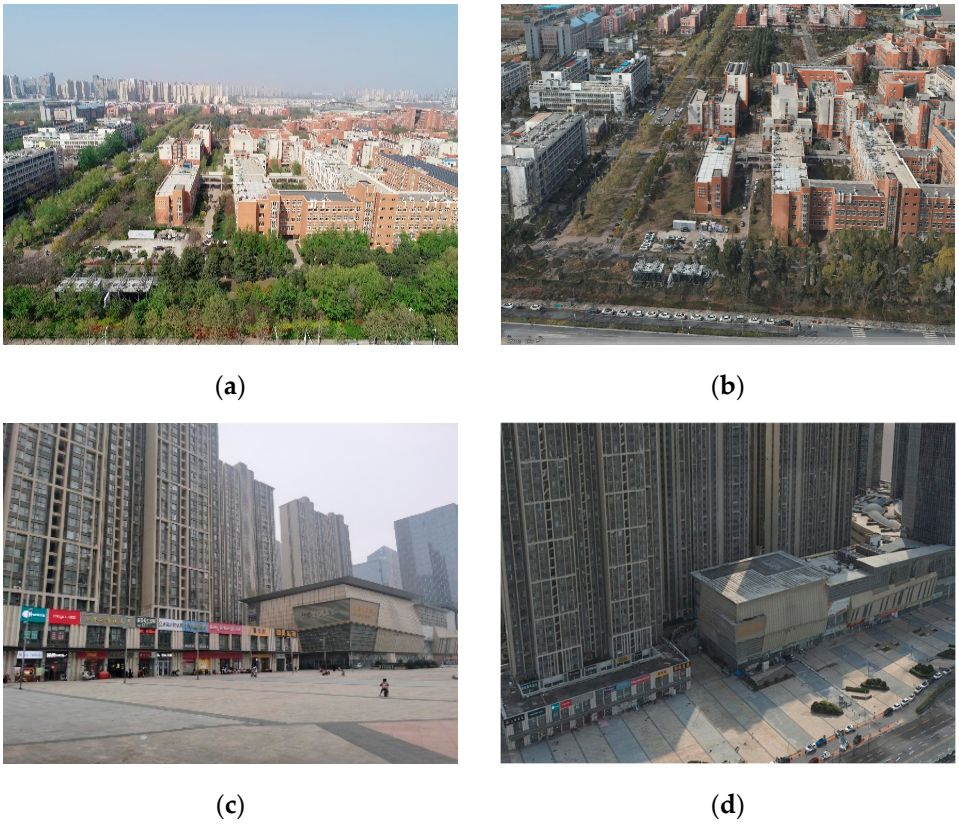
4. Results and Discussion

In order to verify this method, a 2D-3D registration experiment was designed, and the performance of this method in multi-target localization of a monocular image was analyzed with target localization as an application case. To evaluate the influence of the accuracy of the digital twin city model on registration, this study also designed a simulation experiment.

4.1. Experiment of 2D-3D Registration

4.1.1. Experimental Environment

We used images captured using a smartphone in the 2D-3D registration experiment. A Huawei Meta 30 smartphone was used because it has built-in general-purpose GPS and geomagnetic sensors, which were necessary for this experiment. Two photos were used as real images for the experiment, named P1 (Figure 4a) and P2 (Figure 4c), which were taken in September 2021. The estimated and actual poses are presented in Table 1, and the coordinates are described by the ECEF coordinate system. The estimated pose was obtained using the CamPOS application. CamPOS is an image acquisition program developed in this study. While shooting photos, it can acquire GPS and geomagnetic sensor information without redundant optimization processing. The virtual digital twin scenes (Figure 4b,d) are oblique photography 3D models generated from UAV aerial images. The images were captured in December 2020. Because of the difference in seasons and time, the texture of the virtual scene 3D model was different from the appearance of the test image, which increased the difficulty of image matching.



(**a**) (**b**)

(**c**) (**d**)

Figure 4. Real images and virtual scenes. (**a**,**c**) Real images taken with a vision sensor; (**b**,**d**) digital twin virtual scene.

Table 1. Poses of the Real Images.

Image	CamPOS		Actual		Error	
	Position (m) (X/Y/Z)	**Orientation (°) (Y/P/R)**	**Position (m) (X/Y/Z)**	**Orientation (°) (Y/P/R)**	**Position(m) (X/Y/Z)**	**Orientation (°) (Y/P/R)**
P1	−2093390.42 4806197.15 3621100.70	284.9 −13.5 360.0	−2093394.17 4806194.13 3621107.46	269.67 −5.78 353.68	3.75 3.02 −6.76	15.23 −7.72 6.32
P2	−2093850.41 4806374.77 3620496.15	217.2 3.1 359.0	−2093851.67 4806372.58 3620501.33	222.46 10.63 1.39	1.26 2.18 −5.18	−5.26 −7.53 −2.39

4.1.2. PDNT Experiment

According to the proposed method, the Cesium digital earth platform was used as the 3D GIS system in the application of the digital twin city, and the virtual perspective and inverse depth maps of the digital twin city model were generated according to the estimated pose, as shown in Figure 5. The perspective image was matched with the real image, and 3D coordinates for the feature points were obtained from the inverse depth image.

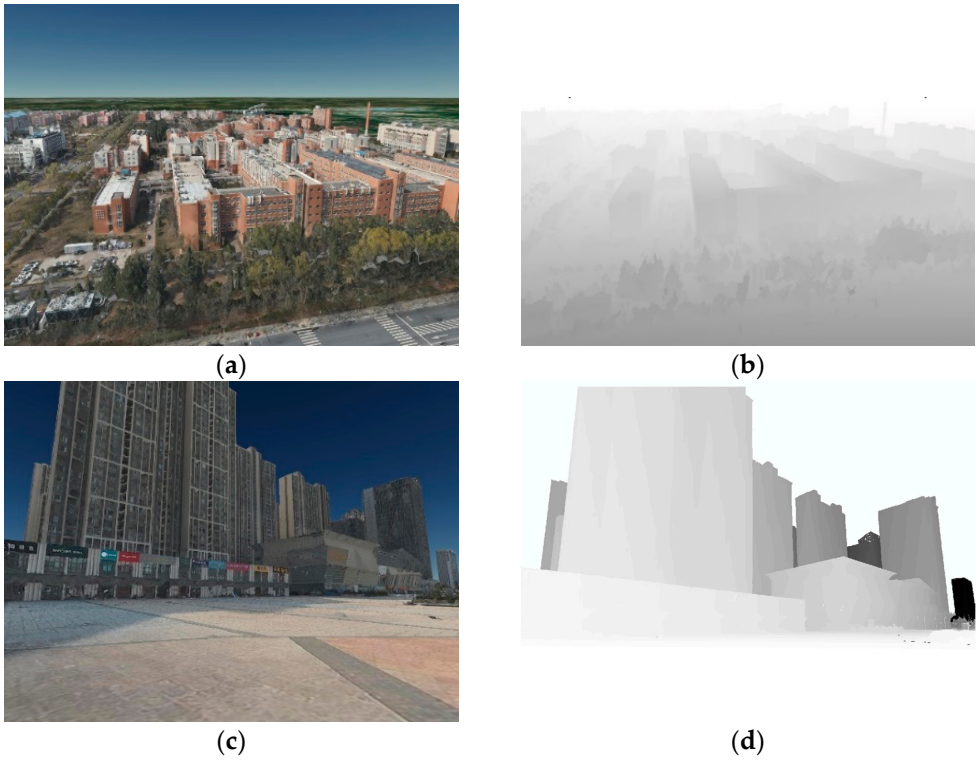


Figure 5. Perspective images and inverse depth images: (**a**) perspective image of P1; (**b**) inverse depth image of P1; (**c**) perspective image of P2; and (**d**) inverse depth image of P2.

4.1.3. Structural Semantic Feature Extraction Experiment

EDLines was used to extract the line feature of the image. According to the direction of gravitational acceleration, the candidate plumb line set was obtained: set $\mu = 0.5$, $it_c = 105$. The vanishing point was calculated by the two-line MSS method, and the plumb line was extracted. The experimental results are shown in Figure 6.

4.1.4. MP-IDSSC Experiment

To achieve accurate camera positioning and orientation, one must establish as many correct matching relationships as possible. To verify the effectiveness of the proposed method, we set up several experimental groups (Table 2). Each experimental group used a different combination of feature extraction and pose estimation methods. This study used the D2-Net, SuperPoint, and SIFT feature extraction methods for comparative experiments. SIFT is an artificially designed feature description and extraction method that is widely used. D2-Net can extract similar scale-invariant features from different images and has strong adaptability and robustness, and SuperPoint is a feature point detection and descriptor extraction method based on self-supervised training. To improve the feature extraction time performance and ensure the same conditions in the experiment, the number of feature points extracted was set to 1000. Then, to validate the performance of our IDSSC-MP algorithm, we selected three advanced algorithms for comparison in terms of feature point matching: SuperGlue [34], GMS, and RANSAC. Their source codes can be downloaded from the internet, and the default parameters were used in the settings.

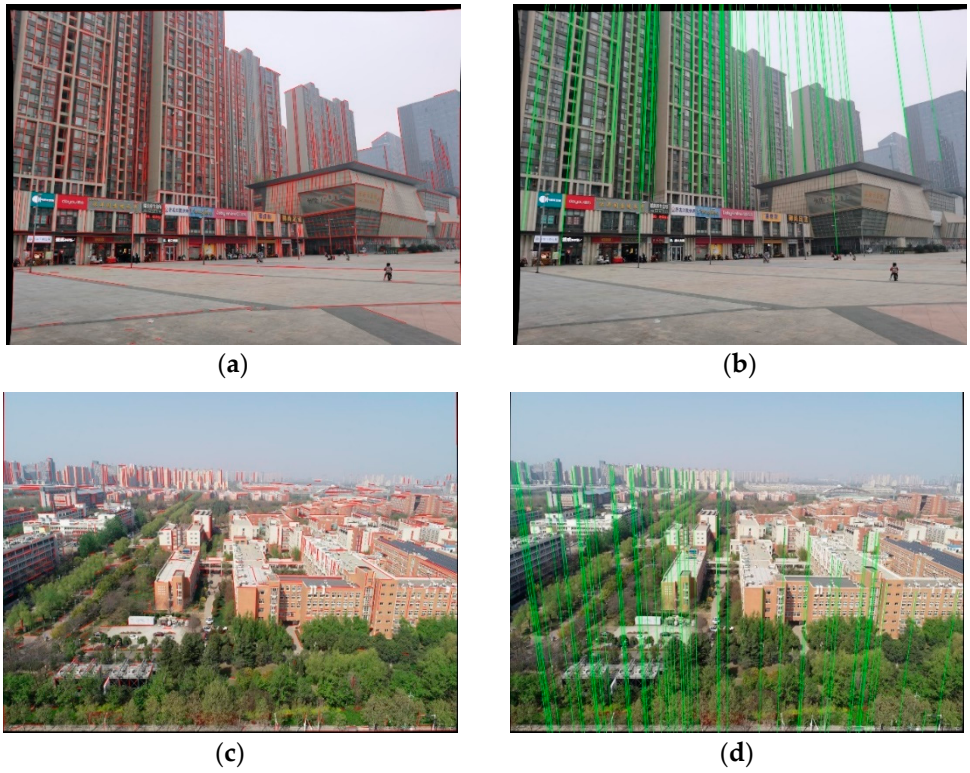


Figure 6. EDLines features and plumb lines: (**a**) red line segments are the EDLines features of P1; (**b**) green lines are the plumb lines of P1; (**c**) red line segments are the EDLines features of P2; (**d**) green lines are the plumb lines of P2.

Table 2. Experimental Groups.

Trial	Feature Extraction	Pose Estimation Method
Trial 1	SuperPoint	IDSSC-MP
Trial 2	D2-NET	IDSSC-MP
Trial 3	SuperPoint	IDC
Trial 4	SuperPoint	SuperGlue
Trial 5	SuperPoint	GMS
Trial 6	SIFT	Ransac

1. Evaluation Metrics

The experimental results are presented in Table 3. In this experiment, six quantitative evaluation metrics were adopted: number of correct matches (NCM), root mean square error (RMSE), camera position (CP), camera position error (CPE), camera orientation (CO), and camera orientation error (COE). The NCM and RMSE are the two main metrics used to evaluate registration performance. The NCM represents the number of matching point pairs used for the pose calculation; the higher the number, the better the registration performance. RMSE is a metric for evaluating the accuracy of the matching point position; the smaller the RMSE, the higher the matching accuracy. CP denotes the camera coordinates calculated from experiments in the ECEF coordinate system, and CPE is the error value of CP compared with the actual camera coordinates. CO denotes the camera orientation calculated from experiments in the yaw/pitch/roll system, and the unit is degree. COE is the orientation error value of the CO compared with the actual camera orientation. The unit is mrad, and 1 mrad is about 0.057 degrees. This means an error of about 1 meter beyond 1000 m.

Table 3. Experimental Results of MP-IDSSC.

Trials	Image	NCM	RMSE	CP X/Y/Z(m)	CPE(m)	CO Y/P/R(°)	COE (*mrad*)
Trial 1	P1	332	0.60	−2093392.02	2.15	270.81	19.89
				4806196.97	2.83	−7.39	28.18
				3621103.69	3.77	354.87	20.76
	P2	373	0.59	−2093851.7	0.44	221.33	19.63
				4806372.58	0.89	9.48	20.07
				3620501.74	1.18	0.70	12.13
Trial 2	P1	223	1.83	−2093390.58	3.59	271.37	29.67
				4806196.99	2.85	−7.78	34.99
				3621102.49	4.97	355.83	37.61
	P2	273	1.66	−2093851.79	0.56	221.29	20.42
				4806372.7	1.01	9.16	25.56
				3620501.78	1.60	0.73	11.51
Trial 3	P1	173	0.86	−2093392.05	2.18	271.31	28.53
				4806197.18	3.05	−10.02	74.00
				3621100.77	6.69	356.40	47.56
	P2	219	0.67	−2093851.84	0.61	221.28	20.68
				4806372.32	0.63	8.60	35.34
				3620501.88	2.23	0.81	10.03
Trial 4	P1	182	0.82	−2093392.95	1.22	271.33	28.97
				4806197.72	3.58	−10.03	74.17
				3621100.53	6.93	356.38	47.21
	P2	146	0.96	−2093851.6	0.40	221.33	19.72
				4806372.82	1.13	8.455	37.96
				3620500.33	1.79	0.685	12.30
Trial 5	P1	30	1.89	−2093391.17	3.00	271.55	32.81
				4806196.71	2.57	−9.75	69.37
				3621098.40	9.06	356.13	42.84
	P2	33	1.33	−2093851.7	0.43	221.25	21.11
				4806373.89	2.20	8.34	39.88
				3620498.87	3.24	0.7	12.04
Trial 6	P1	13	1.05	−2093246.81	147.36	183.38	1506.04
				4806182.10	12.03	−34.71	504.92
				3621131.57	24.11	76.59	1447.05
	P2	7	1.69	−2093842.5	8.77	218.86	23.38
				4806368.51	3.17	17.66	162.83
				3620501.05	1.07	7.16	124.96

2. Results Analysis

The results show that the camera pose estimated by using the traditional point feature matching method (SuperGlue and GMS) has a larger vertical error because the Z-coordinate error is larger than the X, Y coordinate error, and the pitch angle error is larger than the yaw

and rolling angle error. This is because real images contain sky, trees, and barren ground, resulting in an uneven vertical distribution of feature points in the image. It can also be seen in Figures 7 and 8 that in the image matching results, there are more feature points in the middle of the image, while there are fewer feature points in the upper and lower parts. This leads to insufficient constraints in the vertical direction, resulting in a large vertical error. In Trials 1-2, because this method uses the structural semantic information in the image and the direction of the plumb line to constrain, the error in the vertical direction is significantly reduced. Comparing Trials 4-5, our IDSSC-MP method obtains better results when the same feature point extraction algorithm is used. In Trials 4-5, the average error of the Z-coordinate is 5.26 m, but 2.88 m in Trials 1-2, which is 45% less than that in Trials 4-5. In Trials 1-2, the accuracy of the NCM and RMS is also higher, and the calculated camera position and attitude errors are smaller, which shows the effectiveness of the MP-IDSSC method.

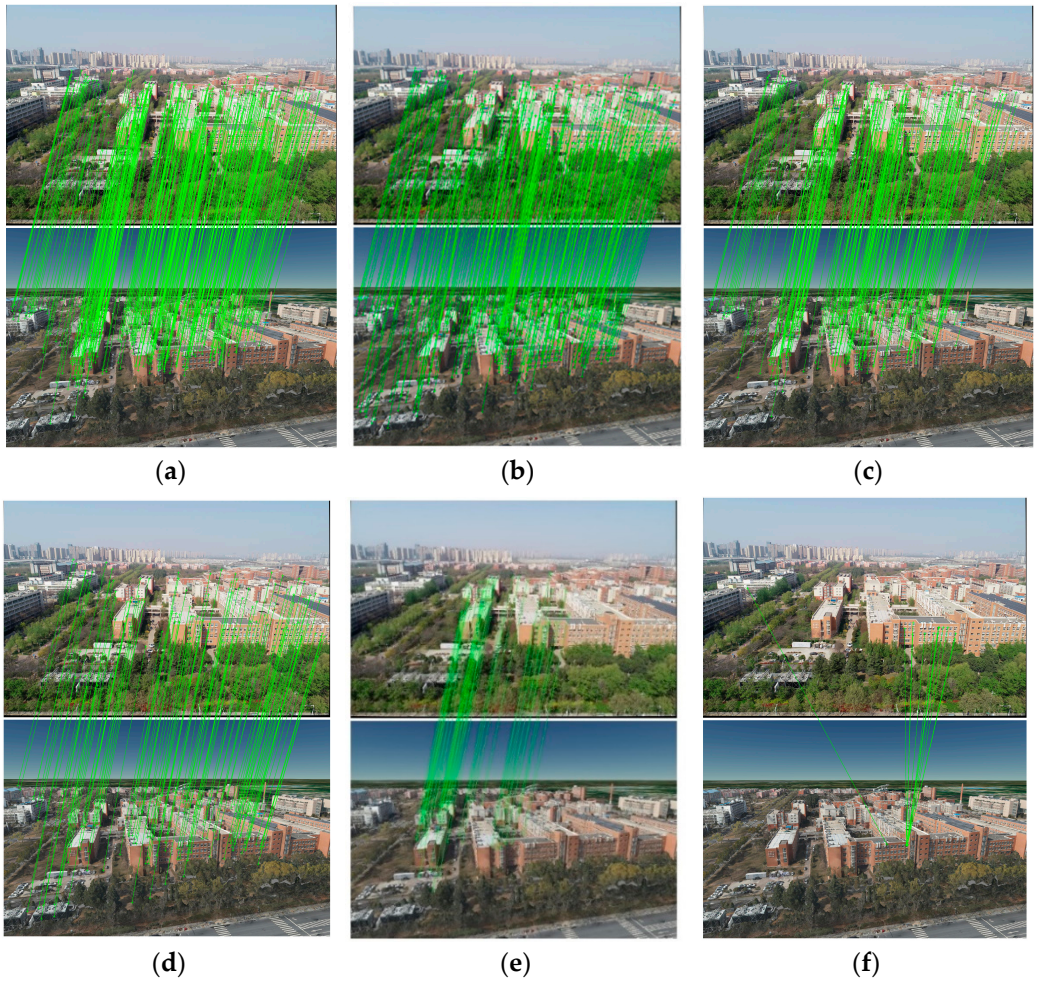


Figure 7. Image matching experiment results of image P1: (**a**–**f**) results of Trails 1-6. Green lines indicate correct matches when considering a threshold value of 2 pixels. The feature points are unevenly distributed in the vertical direction.

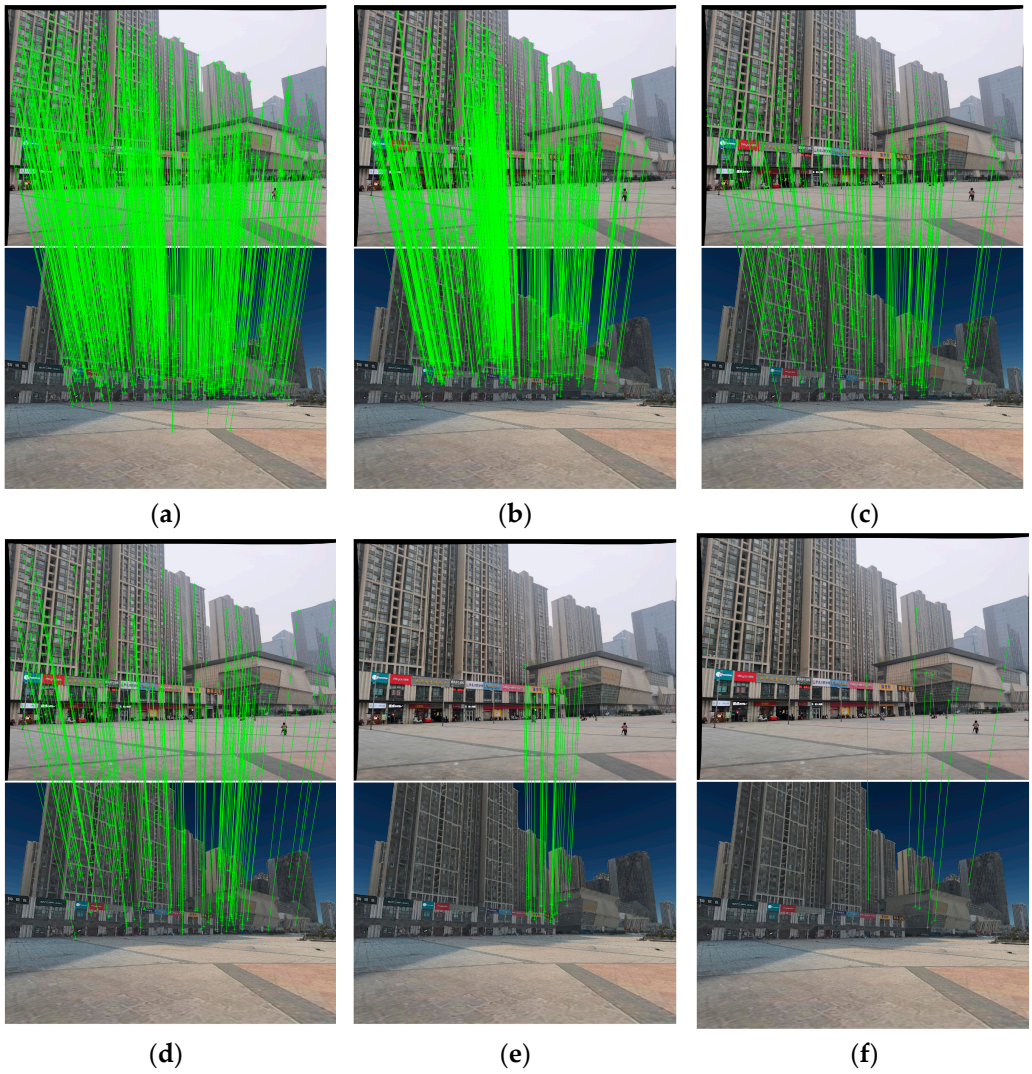


Figure 8. Results of image matching experiment result of image P2: (**a**–**f**) results of Trails 1-6. Green lines indicate correct matches when considering a threshold value of 2 pixels. The feature points are unevenly distributed in the vertical direction.

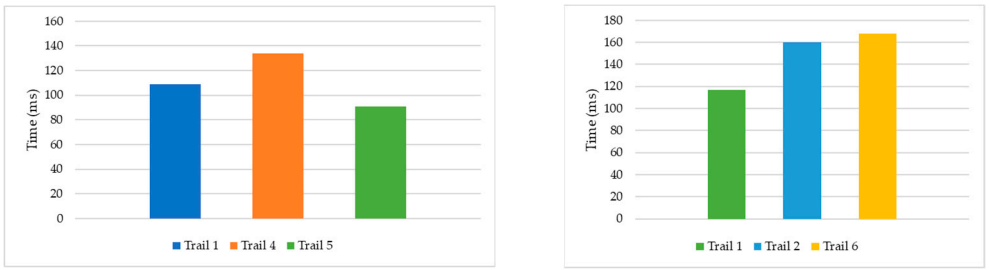
In order to further illustrate the effectiveness of structural semantic constraints, we designed Trial 3. In Trial 3, there is only the constraint of inverse depth coordinates, and there is no constraint of structural semantics. The results show that the vertical error is also high, but compared with Trials 4-5, the experimental results still have advantages. This is because the SuperGlue and GMS methods only restrict the spatial position relationship of feature points in the image plane using a relative graph structure and may not be able to find an accurate matching relationship. For the most accurate correspondence to reduce false matching caused by similar multiple feature points, Trial 3 used the inverse depth coordinates of feature points and the 3D–2D projection relationship as constraints.

Comparing Trials 1 and 4, the SuperPoint method is better than the D2-Net method when the MP-IDSSC is used as the feature matching method. Although D2-Net performs better than SuperPoint in the cross-modality image environment, the positioning accuracy of the D2-Net feature point is lower. After eliminating the feature points with large projection errors using IDSSC constraints, the remaining feature points participating in the camera pose estimation are fewer, and the RMS accuracy of the feature points is still low. In Trial 6, the results of matching and pose estimation are obviously wrong, which indicates that the traditional artificial feature extraction and RANSAC matching methods are not effective for images with large differences in appearance.

Considering the above experimental results, this method can optimize the initial vision sensor pose effectively, utilizes the constraints of inverse depth and structure semantic information to estimate the camera pose more accurately, and reduces the Z-coordinate error caused by the uneven distribution of feature points in the vertical direction. In addition, the proposed method is more accurate, has stronger adaptability than existing algorithms, and performs effectively in challenging conditions.

4.1.5. Time Efficiency Analysis Experiment

Figure 9 shows the time efficiency of the proposed method. The runtime of the algorithm is largely affected by the computer configuration. In this experiment, the computer was configured with a Windows 10 64-bit operating system on an Intel Core i9-9980XE 3.0 GHz CPU, with 64 GB of RAM and an NVIDIA GeForce RTX™ 3090 graphics card.



(**a**) (**b**)

Figure 9. Experimental results of time performance test: (**a**) time performance comparison of image matching algorithm; (**b**) time performance comparison of feature extraction.

Figure 9a shows the time performance of the feature matching and pose estimation methods under the constraint of inverse depth coordinates. The results show that the proposed method has advantages in Trial 1 in terms of time performance compared with the results of Trial 4. In Trial 5, the GMS method was used, which reduces the matching step time, but its feature point location accuracy is low. Figure 8b compares the time performance of different feature extraction methods and shows that the SuperPoint method took the shortest time.

4.2. Application Experiment of Multi-Object Positioning for Monocular Image

Object positioning is commonly used in digital twin applications. This study designed a multi-object spatial positioning application experiment based on a monocular image to verify the effectiveness of this method in object positioning. Once the pose of the camera is computed, the object target can be located. This experiment was conducted using the digital twin city model data. The error between the result of object geolocation and that obtained from the model data was analyzed. Three different points in each image were selected as target points in the image; the position of each point in the image is shown in Figure 10. The coordinates of the points were calculated based on Trials 1, 4, and 5 and the coordinates of the target points corresponding to the model data as true values.

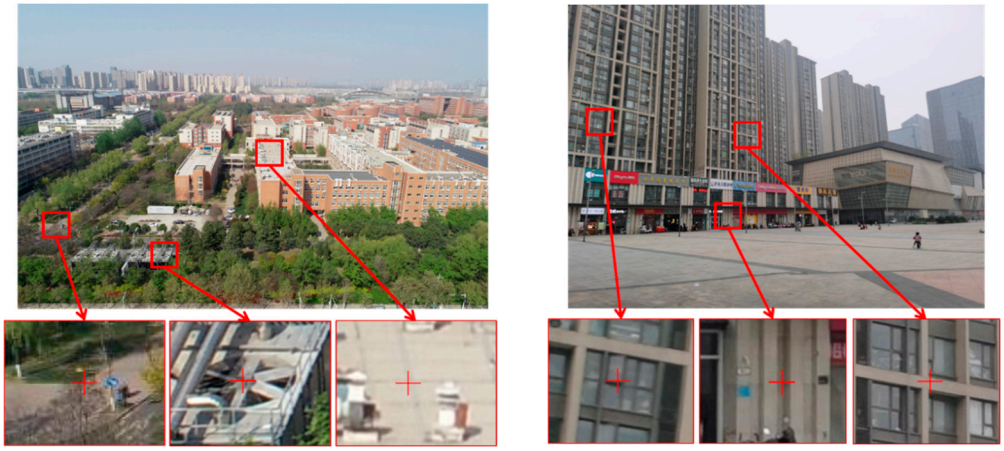


Figure 10. Target object points in P1 and P2.

Table 4 lists the basic information for each target point, and Table 5 lists the different experimental results obtained. In the target point information, target distance (TD) refers to the distance between the target point and the camera in meters, and radial distance (RD) refers to the radial distance of the target point on the image from the principal point in pixels. Base coordinates (BC) refer to the coordinates of the target point in the base geographic data, which are used to evaluate the errors of the coordinates of the target point calculated in different trials. The evaluation index includes the position error (PE), absolute position error (APE), and relative position error (RPE). The PE refers to the error of the target point in the ECEF system. The APE is the L2 distance of the PE. The RPE is the relative error calculated using RPE = APE/TD.

Table 4. Basic Information for Each Target Point.

	T1	T2	T3	T4	T5	T6
TD(m)	225.08	172.25	119	103	82	64
RD	(728,435)	(111,655)	(423,755)	(696,534)	(138,433)	(751,863)
BC	−2093175.39	−2093268.12	−2093291.67	−2093828.59	−2093861.08	−2093831.16
	4806240.40	4806242.66	4806203.83	4806463.96	4806437.85	4806415.05
	3621093.20	3621000.58	3621049.34	3620461.02	3620453.39	3620457.80

Table 5. Experimental Results of Multi-Object Positioning.

Target	Index	Trial 1	Trial 4	Trial 5
T1	PE(m)	−0.31/−0.36/0.13	−0.27/0.64/−0.41	0.18/−0.38/0.5
	APE(m)	0.50	0.80	0.65
	RPE	0.25%	0.39%	0.32%
T2	PE(m)	−0.25/−0.32/0.31	−0.98/0.88/−1.68	−2.88/0.21/−2.06
	APE(m)	0.68	2.13	3.54
	RPE	0.4%	1.2%	2.0%
T3	PE(m)	−1.52/0.19/−0.66	−0.72/0.72/1.97	−1.31/−1.64/1.56
	APE(m)	1.66	2.21	2.61
	RPE	1.5%	1.90%	2.37%
T4	PE(m)	0.1/0.07/0.1	0.07/0.12/0.01	0.01/0.12/−0.05
	APE(m)	0.16	0.15	0.13
	RPE	0.15%	0.14%	0.12%

Table 5. *Cont.*

Target	Index	Trial 1	Trial 4	Trial 5
T5	PE(m)	0.27/0.12/0.11	0.21/0.14/0.05	−0.32/0.22/0.02
	APE(m)	0.32	0.25	0.40
	RPE	0.4%	0.49%	0.35%
T6	PE(m)	−0.05/1.54/−1.71	−0.1/2.32/−2.39	1.01/2.27/−2.86
	APE(m)	2.30	3.33	3.79
	RPE	3.5%	5.2%	5.9%

Experiments show that this method can alleviate the uneven distribution of target position errors. In Trials 4 and 5, targets T1, T4, and T5 are located near the image center with an average relative error of 0.3%, while T3 and T6 are located near the edge of the image with an average relative error of 3.84%. This is because the camera pose error is larger in the vertical direction, resulting in a target position error in the image, which also presents the problem of uneven distribution in the vertical direction. Because of the error in the calculated camera position, the calculated principal optical axis direction differs significantly from the true principal optical axis pitch angle, which results in a small position error near the image principal point and a large position error at the image edge in the vertical direction. In Trial 1, the target position error at the edge of the image is reduced because of the constraints of inverse depth and structure semantics, which alleviates the problem.

It also can be seen from the experiment that the absolute error of object registration accuracy of this method is 0.66 m, while the average error of the camera's own pose is 3.36 m. This is because, although the camera pose error is larger, after position correction, the two main optical axes can intersect near the image center target point, so the target positioning accuracy is higher than the camera's pose. This indicates that this method does not need to rely on the camera's own high-precision pose when locating the object, and our method can achieve high-precision registration.

4.3. Simulation Experiments and Analysis

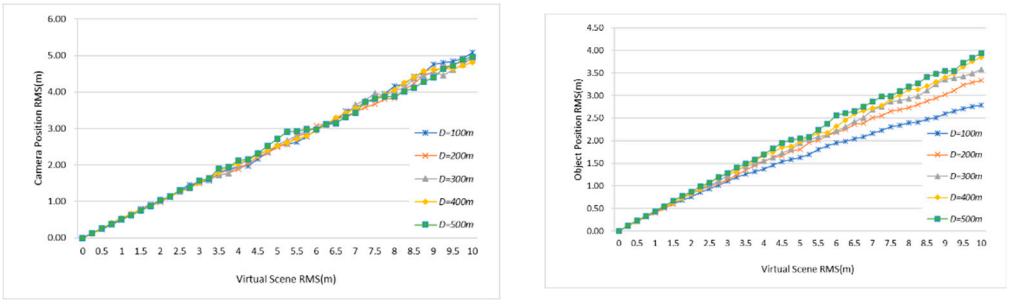
Due to differences in modeling methods and data sources, the accuracy of 3D models used in digital twin applications is often different, and different devices and environments have an impact on 2D-3D registration. To further analyze the impact of 2D-3D registration accuracy, simulation experiments were carried out in this study. The error factors affecting the registration algorithm are divided into systematic and random errors. The systematic error can be compensated for in the algorithm by measuring the camera calibration in the laboratory; therefore, the random error from various sensor measurements was the main factor causing the localization error. Table 6 lists the commonly used error terms and their distributions.

Table 6. Parameters affecting target localization accuracy.

Data	Error Model	Unit
Camera Position Error	Normal Distribution	$\varphi_A/(^\circ)$
Camera Orientation Error	Normal Distribution	$\psi/(^\circ)$
Pixel Measurement Error	Normal Distribution	$u/(pixel)$
Feature Matching Error	Normal Distribution	/
Image Distortion Error		/
Virtual Scene RMS Error	Normal Distribution	m
Feature Count	/	/
Field of View	/	/

The impact of the parameters in Table 5 on registration accuracy requires in-depth analysis. The literature [35–37] has analyzed and performed experiments on parameters

such as image distortion and camera pose. Hence, we adopted the Monte Carlo [38] method to analyze the impact of virtual scene errors on camera positioning and object positioning accuracy. The Monte Carlo method, also known as the stochastic simulation method, uses computers to generate qualified random data that can replace data that are difficult to obtain in experiments. Figure 11a shows the relationship between virtual scene data errors and calculated camera position root mean square (RMS) errors when the camera distance D is from 100 m to 500 m. Taking the center point of the image as the object position, Figure 11b describes the relationship between the RMS object position error and the virtual scene data error.



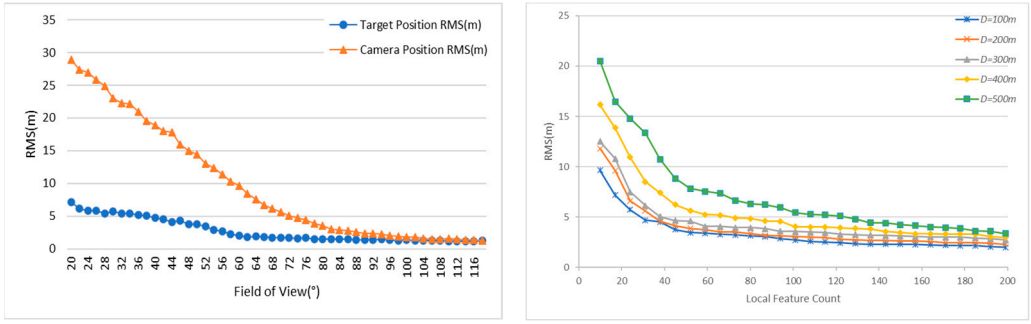
(**a**) (**b**)

Figure 11. Influence of virtual scene error on 2D-3D registration accuracy: (**a**) influence on camera pose accuracy; and (**b**) influence on object positioning accuracy.

The results show that the camera position error increases with the increase in virtual scene errors, but at different distances D, the RMS camera position errors remain approximately the same. When analyzing the object position error, the virtual scene data error increases when the RMS target position error increases. Additionally, when D increases, the incremental proportion of the RMS object position error is smaller than the incremental proportion of the distance D.

These experimental results show that: (1) The camera position can be used as a constraint condition for image matching. If only the virtual scene data error needed to be considered, when the virtual scene data error was 10 m and the object distance was 500 m, the calculated RMS camera position error was only 5 m. (2) The result of object localization was better than the calculated result for the camera position. When D = 100 m, the calculated RMS object position error was only half of the camera position error. This is because when the camera direction was aimed at the target, a more accurate target localization result could be obtained.

Figure 12 illustrates the impact of the field of view and feature count on the calculation accuracy of the camera pose and target position. The results show that when the field of view increases, the camera position error and target position error gradually decrease and that the field of view has a greater impact on the accuracy of the camera position. The feature count also has an impact on the target localization results. We used random sampling to uniformly select different numbers of feature points from an image for comparative experiments. The figure shows that when the feature count decreases, the RMS error of target localization increases continuously. Notably, when the feature count is less than 50, the RMS error of localization greatly increases.



(**a**) (**b**)

Figure 12. Analysis of the influence of the field of view (**a**) and the feature count (**b**).

5. Conclusions

This paper proposes a virtual–real 2D-3D registration method under the constraints of inverse depth and structural semantics. First, the perspective and inverse depth images of the virtual scene are obtained by using PDNT technology, and the structural semantic features are extracted by the two-line MSS plumb line extraction method. Then, we estimate the camera pose under the constraints of inverse depth information and structural semantics and accurately achieve the registration of the real image and the virtual scene. The experimental results show that the proposed method can achieve high-precision vision sensor registration in a digital twin scene and solves the large vertical error problem effectively. The application experiment of monocular image multi-object spatial positioning proves the practicability of this method.

This study obtained the 2D-3D registration of static images. Future studies will include ways to use digital twin scenes to achieve dynamic target tracking and real-time localization based on monocular images. Because the manual selection of measurement targets has a negative impact on user experience, achieving target capture in an outdoor environment is an important issue that will likely be included in future research.

Author Contributions: Conceptualization, X.H. and Y.Z.; methodology, X.H. and Q.S.; software, X.H.; validation, Q.S.; formal analysis, X.H. and Y.Z.; investigation, Q.S.; resources, X.H. and Q.S.; data curation, X.H.; writing—original draft preparation, X.H. and Y.Z.; writing—review and editing, X.H.; visualization, Q.S.; supervision, Y.Z.; project administration, Q.S. All authors have read and agreed to the published version of the manuscript.

Funding: The Natural Science Foundation of Henan Province (No. 202300410536).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors appreciate the editors and reviewers for their comments, suggestions, and valuable time and effort in reviewing this manuscript.

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

References

1. Deng, T.; Zhang, K.; Shen, Z.-J. A systematic review of a digital twin city: A new pattern of urban governance toward smart cities. *J. Manag. Sci. Eng.* **2021**, *6*, 125–134. [CrossRef]
2. Gao, Z.; Song, Y.; Li, C.; Zeng, F.; Wang, P. Research on the Application of Rapid Surveying and Mapping for Large Scare Topographic Map by UAV Aerial Photography System. *Remote Sens.* **2017**, *42*, 121. [CrossRef]
3. Billinghurst, M.; Clark, A.; Lee, G.J. A survey of augmented reality. *Found. Trends Hum.-Comput. Interact.* **2015**, *8*, 73–272. [CrossRef]
4. Li, C.; Liu, Z.; Zhao, Z.; Dai, Z. A fast fusion method for multi-videos with three-dimensional GIS scenes. *Multimed. Tools Appl.* **2020**, *80*, 1671–1686. [CrossRef]

5. Xu, Z.; Wang, G. Research on the Method of 3D Registration Technology. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *440*, 032139. [CrossRef]
6. Coughlan, J.M.; Yuille, A.L. Manhattan World: Compass Direction From a Single Image by Bayesian Inference. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; pp. 941–947.
7. Ma, W.; Xiong, H.; Dai, X.; Zheng, X.; Zhou, Y. An Indoor Scene Recognition-Based 3D Registration Mechanism for Real-Time. *AR-GIS Vis. Mob. Appl.* **2018**, *7*, 112.
8. Li, J.; Wang, C.; Kang, X.; Zhao, Q. Camera localization for augmented reality and indoor positioning: A vision-based 3D feature database approach. *Int. J. Digit. Earth* **2019**, *13*, 727–741. [CrossRef]
9. Wu, Y.; Che, W.; Huang, B.G.T. An Improved 3D Registration Method of Mobile Augmented Reality for Urban Built Environment. *Int. J. Comput. Games Technol.* **2021**, *2021*, 1–8. [CrossRef]
10. Yue, L.; Li, H.; Zheng, X.J.S. Distorted Building Image Matching With Automatic Viewpoint Rectification and Fusion. *Sensors* **2019**, *19*, 5205. [CrossRef]
11. Huang, W.; Sun, M.; Li, S. A 3D GIS-Based Interactive Registration Mechanism for Outdoor Augmented Reality System. *Expert Syst. Appl.* **2016**, *55*, 48–58. [CrossRef]
12. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
13. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superpoint: Self-Supervised Interest Point Detection and Description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 224–236.
14. Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; Sattler, T. D2-Net: A Trainable cnn for Joint Description and Detection of Local Features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8092–8101.
15. Li, J.; Hu, Q.; Ai, M. 4FP-Structure: A Robust Local Region Feature Descriptor. *Photogramm. Eng. Remote Sens.* **2017**, *83*, 813–826. [CrossRef]
16. Liu, Z.; Marlet, R. Virtual Line Descriptor and Semi-local Matching Method for Reliable Feature Correspondence. In Proceedings of the Machine Vision Conference, Surrey, UK, 3–7 September 2012; Volume 2012, pp. 16.1–16.11.
17. Chum, O.; Matas, J. Matching With PROSAC-Progressive Sample Consensus. In Proceedings of the Computer Vision and Pattern Recognition (CVPR05), San Diego, CA, USA, 20–26 June 2005; pp. 220–226.
18. Sattler, T.; Leibe, B.; Kobbelt, L. SCRAMSAC: Improving RANSAC's Efficiency With a Spatial Consistency Filter. In Proceedings of the International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 2090–2097.
19. Jiang, S.; Jiang, W. Reliable Image Matching via Photometric and Geometric Constraints Structured by Delaunay Triangulation. *ISPRS J. Photogramm. Remote Sens.* **2019**, *153*, 1–20. [CrossRef]
20. Bian, J.; Lin, W.; Liu, Y.; Zhang, L.; Yeung, S.; Cheng, M.; Reid, I. GMS: Grid-Based Motion Statistics for Fast, Ultra-Robust Fea-ture Correspondence. *Int. J. Comput. Vis.* **2020**, *128*, 1580–1593. [CrossRef]
21. Pai, P.; Naidu, V.P.S. Target Geo-localization Based on Camera Vision Simulation of UAV. *J. Opt.* **2017**, *46*, 425–435. [CrossRef]
22. Fu, Q.; Quan, Q.; Cai, K.-Y. Robust Pose Estimation for Multirotor UAVs Using Off-Board Monocular Vision. *IEEE Trans. Ind. Electron.* **2017**, *64*, 7942–79511. [CrossRef]
23. Zhang, L.; Deng, F.; Chen, J.; Bi, Y.; Phang, S.K.; Chen, X.; Chen, B.M. Vision-Based Target Three-Dimensional Geolocation Using Unmanned Aerial Vehicles. *IEEE Trans. Ind. Electron.* **2018**, *65*, 8052–8061. [CrossRef]
24. Roig, G.; Boix, X.; Shitrit, H.B.; Fua, P. Conditional Random Fields for Multi-camera Object Detection. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; Volume 2011, pp. 563–570.
25. Shi, J.; Zou, D.; Bai, S.; Qian, Q.; Pang, L. Reconstruction of dense three-dimensional shapes for outdoor scenes from an image sequence. *Opt. Eng.* **2013**, *52*, 123104. [CrossRef]
26. Sánchez, A.; Naranjo, J.M.; Jiménez, A.; González, A. Analysis of Uncertainty in a Middle-Cost Device for 3D Measurements in BIM Perspective. *Sensors* **2016**, *16*, 1557. [CrossRef]
27. Ma, J.; Bajracharya, M.; Susca, S.; Matthies, L.; Malchano, M. Real-Time Pose Estimation of a Dynamic Quadruped in GPS-Denied Environments for 24-Hour Operation. *Int. J. Robot. Res.* **2016**, *35*, 631–653. [CrossRef]
28. Tekaya, S.B. *Distance Estimation Using Handheld Devices*; Naval Postgraduate School: Monterey, CA, USA, 2013.
29. Akinlar, C.; Topal, C. EDLines: A real-time line segment detector with a false detection control. *Pattern Recognit. Lett.* **2011**, *32*, 1633–1642. [CrossRef]
30. Lu, X.; Yaoy, J.; Li, H.; Liu, Y.; Zhang, X. 2-Line Exhaustive Searching for Real-Time Vanishing Point Estimation in Manhattan World. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 31 March 2017; Volume 2017, pp. 345–353.
31. Mukhopadhyay, P.; Chaudhuri, B.B. A survey of Hough Transform. *Pattern Recognit.* **2015**, *48*, 993–1010. [CrossRef]
32. Von Gioi, R.G.; Jakubowicz, J.; Morel, J.M.; Randall, G. LSD: A Line Segment Detector. *Image Processing Line* **2012**, *2*, 35–55. [CrossRef]
33. Fotiou, I.A.; Rostalski, P.; Parrilo, P.A.; Morari, M. Parametric Optimization and Optimal Control Using Algebraic Geometry Methods. *Int. J. Control* **2006**, *79*, 1340–1358. [CrossRef]

34. Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superglue: Learning Feature Matching With Graph Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4938–4947.
35. Deng, F.; Zhang, L.; Gao, F.; Qiu, H.; Gao, X.; Chen, J. Long-Range Binocular Vision Target Geolocation Using Handheld Electronic Devices in Outdoor Environment. *IEEE Trans. Image Process.* **2020**, *29*, 5531–5541. [CrossRef]
36. Cai, Y.; Ding, Y.; Xiu, J.; Zhang, H.; Qiao, C.; Li, Q. Distortion Measurement and Geolocation Error Correction for High Altitude Oblique Imaging Using Airborne Cameras. *J. Appl. Remote Sens.* **2020**, *14*, 014510. [CrossRef]
37. Qiao, C.; Ding, Y.; Xu, Y.; Xiu, J. Ground Target Geolocation Based on Digital Elevation Model for Airborne Wide-Area Reconnaissance System. *J. Appl. Remote Sens.* **2018**, *12*, 016004. [CrossRef]
38. Collings, B.J.; Niederreiter, H. Random Number Generation and Quasi-Monte Carlo Methods. *J. Am. Stat. Assoc.* **1993**, *88*, 699. [CrossRef]