

Article

# Video Geospatial Mapping of Large-Scale Tower-Based Cameras Based on 3D GIS and Gradient Descent

Xianguo Ling, Xingguo Zhang \*, Xin Li and Xiangfei Meng

School of Geographic Sciences, Xinyang Normal University, Xinyang 464000, China; lxg006@xynu.edu.cn (X.L.); lixin619@xynu.edu.cn (X.L.); mengxiangfei@xynu.edu.cn (X.M.)

\* Correspondence: zhangxingguo@xynu.edu.cn

## Abstract

To address the challenges of the large-scale georeferencing of tower-based cameras and the limited capability of video-based spatial analysis, we proposed a geospatial mapping method integrating 3D GIS and gradient descent optimization. Using a Digital Elevation Model (DEM), high-resolution remote sensing imagery, and tower-based video data as the primary data sources, the proposed method first estimates the intrinsic parameters of the tower-based camera by aligning a 3D GIS virtual camera with the video imagery. Subsequently, the initial camera extrinsic parameters are estimated using the PnP algorithm based on the previously estimated intrinsic matrix  $K$  and the corresponding control point pairs. Building upon these initial estimates, the camera intrinsic and extrinsic parameters are jointly optimized using a constrained L-BFGS-B framework that incorporates prior knowledge of the tower planar location, explicit box constraints, and a semi-constrained parameterization scheme with bounded parameter ranges. Furthermore, an outlier-removal and re-optimization strategy is employed to further improve the accuracy of parameter estimation. Finally, the optimized parameters are employed to transform image coordinates into three-dimensional world coordinates, and video geospatial mapping is achieved through the integration of colored point clouds with the 3D GIS scene. The results showed the following: (1) The 3D GIS scene constructed from publicly available DEM and high-resolution remote sensing imagery met the requirements for the initial estimation of intrinsic and extrinsic camera parameters. (2) Compared with PnP, RANSAC-PnP, SQPnP, and DLT, the proposed method achieves lower reprojection and 3D spatial errors. For the independent check points, the RMSE of the reprojection error is reduced by 66.4%, 73.6%, 68.0%, and 48.3%, respectively, while the RMSE of the 3D spatial error is reduced by 84.6%, 86.2%, 83.1%, and 69.4%, respectively. These results demonstrate that the proposed method provides reliable camera parameter estimates for video geospatial mapping. (3) Using the estimated camera parameters, image coordinates are transformed into 3D world coordinates to generate a georeferenced colored point cloud, which facilitates integrated analysis with existing geospatial datasets. The proposed method provides a feasible solution for tower-based camera georeferencing and three-dimensional visualization under conditions without field calibration. It offers a theoretical and technical basis for geospatial monitoring and related applications.

**Keywords:** tower-based video; 3D GIS; constrained L-BFGS-B; video geospatial mapping



Academic Editors: Wolfgang Kainz, Zengli Wang, Hanlin Zhou, Wenxi Liao and Yongyang Xu

Received: 6 May 2026

Revised: 6 July 2026

Accepted: 10 July 2026

Published: 12 July 2026

**Copyright:** © 2026 by the authors.

Published by MDPI on behalf of the

International Society for

Photogrammetry and Remote Sensing.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

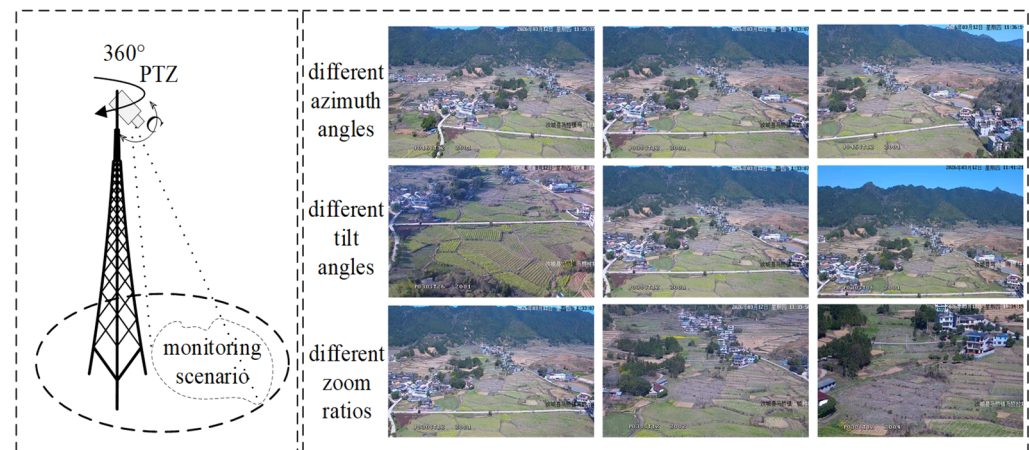
conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

## 1. Introduction

With the rapid development of information and communication technologies, Earth observation approaches have undergone profound transformations, and the representation of geographic information is evolving from traditional static maps toward dynamic, real-time, and fine-grained paradigms [1]. In recent years, large-scale camera networks have been widely deployed in fields such as smart cities, cultivated land protection, precision agriculture, and disaster emergency responses, gradually becoming a new type of infrastructure for geospatial information acquisition. Meanwhile, video monitoring applications have expanded from conventional security surveillance to natural scene and geographic process observation, providing new data support for the fine-grained and real-time management of natural resources [2].

Large-scale deployments of tower-based cameras are typically equipped with optical zoom capabilities of several tens of times, enabling effective long-range observation and supporting real-time video monitoring of natural resources over spatial extents of several kilometers, as illustrated in Figure 1. However, surveillance video adopts a perspective imaging model, which differs substantially from the orthographic view representation commonly used in traditional geospatial data. This discrepancy poses significant challenges for the integration and synergistic analysis of the two data types. To address this issue, many researchers have explored the integration of video and GIS, aiming to fuse video data with traditional maps and remote sensing imagery. In general, this line of research has evolved from symbolic representation to geospatial mapping, and further to three-dimensional scene fusion. Early studies primarily displayed video data on maps as point symbols or field-of-view symbols, enabling positioning, retrieval, and management [3]. Subsequently, research attention shifted toward the mutual mapping between video and geospatial data, aiming to achieve bidirectional transformation between image coordinates and two-dimensional and three-dimensional geospatial coordinates. Representative methods mainly include the homography method, the multi-plane constraint method, and the camera model-based method. Among them, the homography and multi-plane constraint methods can achieve the transformation of one or multiple planar regions in video imagery into geospatial coordinates [4,5]. The camera model-based method requires known intrinsic and extrinsic camera parameters, and performs bidirectional conversion between pixel coordinates and geospatial coordinates through the intersection of viewing rays with a three-dimensional model [6]. The key to this approach lies in camera calibration, for which the direct linear transformation (DLT) [7] and perspective-n-point (PnP) methods are commonly used [8,9]. This process is highly dependent on high-precision camera parameters. Large-scale tower-based cameras are usually installed at elevated positions under diverse terrain conditions. Owing to historical deployment practices, such large-scale devices generally lack high-precision parameters, making field georeferencing and calibration difficult to implement, while substantially increasing the workload and cost of on-site measurements. Therefore, if publicly available geospatial data can be fully utilized to replace field surveys for large-scale camera georeferencing, the associated costs can be effectively reduced. Meanwhile, existing studies that project video textures onto 3D GIS environments [10] have primarily focused on orthorectified video map generation and real-time visualization. However, they generally do not facilitate a systematic estimation and optimization of the camera intrinsic and extrinsic parameters for tower-based cameras, nor do they establish a complete calibration framework tailored to video geospatial mapping. This limitation restricts the support for subsequent high-precision spatiotemporal analysis and the reuse of video-derived geospatial data.



**Figure 1.** Schematic illustration of tower-based video monitoring.

To this end, in response to the demand for the collaborative integration of large-scale tower-based camera video and GIS, this study proposes a video geospatial mapping method integrating 3D GIS and gradient descent optimization. The proposed method consists of three major components, namely intrinsic parameter estimation of the tower-based camera, joint optimization of intrinsic and extrinsic camera parameters, and video geospatial mapping. The core of the method lies in achieving optimized camera calibration based on the correspondence between video image coordinates and 3D GIS world coordinates. On this basis, video frames are further transformed into colored point clouds, forming a new type of geospatial data that can be jointly analyzed with existing geospatial datasets.

## 2. Related Work

### 2.1. Camera Calibration

Camera calibration is the process of estimating a camera's intrinsic and extrinsic parameters and constitutes a critical step in video geospatial mapping. The accuracy of parameter estimation directly determines the accuracy and stability of the mapping results.

Traditional camera calibration methods mainly rely on geometric imaging models and explicit constraint solving, and can generally be categorized into template-based methods, active vision methods, and self-calibration methods. Template-based methods establish the correspondence between three-dimensional control points and two-dimensional image points by using calibration references with known geometric parameters, and perform joint or stepwise estimation of the camera's intrinsic and extrinsic parameters by minimizing the reprojection error. Representative methods include Zhang's calibration method based on a checkerboard calibration board [11], the Tsai two-step method [12], and the Heikkilä–Silvén four-step method [13]. Active vision methods estimate the intrinsic and extrinsic parameters by actively controlling the motion of the camera or platform to generate modellable movements [14], such as methods based on self-calibration theory for moving cameras [15] and calibration based on three orthogonal translational motions [16]. Self-calibration methods estimate camera parameters solely by exploiting the geometric consistency constraints among multi-view images in the absence of known calibration objects or external measurement references. Representative examples include the self-calibration method based on the absolute quadric [17], the Kruppa equation-based self-calibration method [18], and the stratified self-calibration method under varying intrinsic parameters [19]. Although traditional camera calibration methods generally exhibit strong interpretability and high accuracy, they still suffer from several limitations, including a strong dependence on calibration conditions and prior information, high sensitivity to

noise and initial values, limited cross-scene transferability, and high maintenance costs in large-scale dynamic applications.

The above limitations have driven camera calibration to evolve from a static, manually guided geometric solving paradigm toward an intelligent paradigm characterized by data-driven, targetless, and online optimization approaches. These approaches mainly include deep learning-based parameter regression calibration methods, targetless geometry–learning hybrid calibration methods, and joint optimization methods based on differentiable rendering and neural scene representations. In 2018, Schneider et al. [20] proposed the RegNet learning-based extrinsic calibration framework, which directly regressed the LiDAR-camera extrinsic parameters through a convolutional neural network, thereby demonstrating the feasibility of data-driven approaches for cross-modal calibration under targetless conditions. In 2021, Lv et al. [21] proposed the LCCNet method, which introduced feature correspondences and cost-volume modeling into the LiDAR-camera extrinsic estimation process, thereby improving calibration accuracy and robustness in complex scenes. In 2023, Liao et al. [22] conducted a systematic review of learning-based camera calibration and established a unified task classification framework covering the pinhole model, distortion model, cross-view, and cross-sensor calibration tasks, thereby providing a unified analytical paradigm for data-driven calibration research. In 2024, Tan et al. [23] proposed a classification framework for LiDAR-camera learning-based extrinsic calibration methods, dividing them into AEPE and REPP categories, and summarized the development trends of targetless extrinsic estimation in terms of methodological mechanisms, applicable scenarios, and performance boundaries. In 2026, Jung et al. [24] introduced a joint optimization mechanism integrating neural Gaussian scene representation and differentiable rendering under targetless conditions, and proposed the TLC-Calib framework to improve the robustness and generalization capability of extrinsic calibration in complex scenes. Despite the advantages of intelligent camera calibration in terms of automation and adaptability to complex scenarios, such methods still face common challenges, including a strong dependence on high-quality data and annotations, insufficient cross-device and cross-scene generalization, and limited model interpretability. Meanwhile, targetless cross-modal alignment remains constrained by limited robustness in dynamic and weak-texture scenes, while the high computational cost of joint optimization methods, such as differentiable rendering, further restricts their engineering deployment and large-scale applications.

## 2.2. Video Geospatial Mapping

The essence of video geospatial mapping lies in establishing the mapping relationship between image pixel coordinates and two-dimensional and three-dimensional geospatial coordinates, thereby enabling the positioning, overlay, and spatiotemporal analysis of video information within a GIS environment. At present, video geospatial mapping methods mainly include the homography matrix method, automatic matching method, multi-plane constraint method, camera model-based method, line-of-sight and DEM intersection method, and three-dimensional mapping method.

The homography matrix method and the automatic matching method do not require the estimation of camera parameters. Zhang Xingguo et al. [25] researched the mutual mapping between surveillance video and two-dimensional geospatial data, and constructed the transformation relationship between the image plane and the geographic plane using a homography matrix solution based on corresponding control points. The results showed that this method achieved satisfactory mapping efficiency and practical applicability in scenes with relatively small terrain variations. Shao et al. [26] proposed a matching method for projecting vector data onto surveillance video in the context of cultivated land monitoring. By establishing a mapping model based on the feature correspondences between



image and geospatial data, the method enabled the automatic localization of geographic features in video imagery. The findings indicated that this approach reduced the need for manual point selection and improved the level of automation in practical applications. The homography matrix method relies on no fewer than four pairs of corresponding control points and is suitable for approximately planar scenes; however, errors tend to accumulate when the terrain relief is significant or when the camera tilt angle varies considerably. The automatic matching method can reduce labor costs, but it still suffers from mismatching under cross-view, cross-modal, and weak-texture conditions, and its stability is highly dependent on the matching quality and the outlier rejection strategy.

By contrast, multi-plane constraint methods, camera model-based methods, line-of-sight and DEM intersection methods, and three-dimensional mapping methods require accurate intrinsic and extrinsic camera parameters. Zhang et al. [27] conducted research on orthorectified video map generation using a multi-plane constraint and 3D GIS view-matching method. The results showed that, compared with single-plane mapping, this method more effectively alleviated geometric distortions in complex scenes and improved mapping consistency. Milosavljević et al. [28] researched surveillance video georegistration using a camera model-based parameter estimation method. Their results demonstrated that explicitly solving the camera position, orientation, and field-of-view parameters could effectively improve the stability and accuracy of the transformation from pixel coordinates to geographic coordinates. Han et al. [29] investigated the extraction of spatiotemporal information from surveillance video targets using a line-of-sight and DSM/DEM intersection method, and the findings indicated that terrain elevation constraints enhanced target geolocation capability, although the method performance was highly sensitive to the accuracy and resolution of elevation data. Liu Zhendong et al. [30] researched the integration of 3D GIS scenes and multi-source video data using a three-dimensional mapping and multi-coordinate transformation method. The results showed that this method enabled the rapid positioning and collaborative visualization of multiple video objects within a 3D scene, although it required high-quality camera parameters and three-dimensional models.

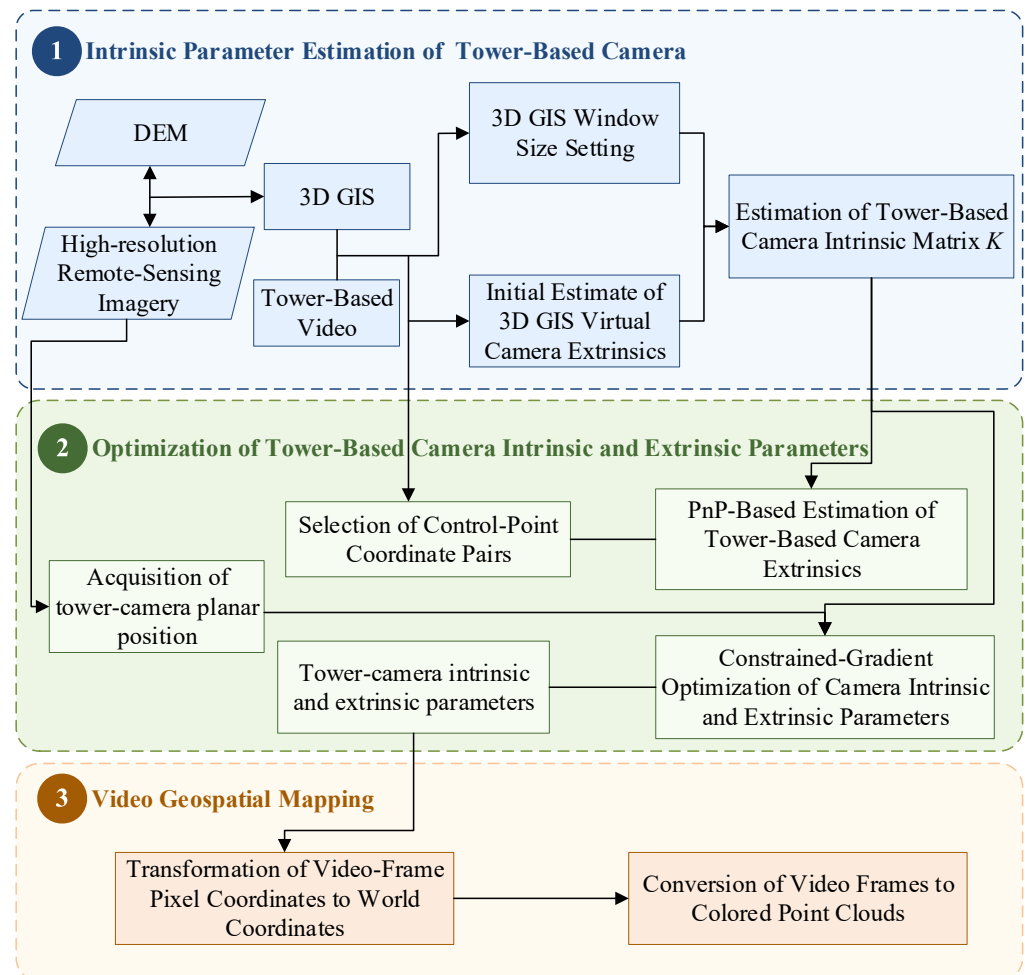
In summary, with respect to camera calibration, neither traditional geometric methods nor intelligent approaches can be readily extended in a low-cost and stable manner to complex terrains and large-scale networks of elevated tower-based cameras. How to achieve high-precision, low-cost, and sustainably updatable parameter estimation for large-scale tower-based cameras under small-sample conditions remains a critical issue that urgently needs to be addressed. With regard to video texture projection, although existing methods can achieve satisfactory visualization performance, they are still not well suited for spatiotemporal analysis.

### 3. Research Methodology

To meet the demand for the high-precision geospatial mapping of large-scale tower-based camera videos, this study proposes a tower-based camera geospatial mapping method integrating 3D GIS and gradient descent optimization. The overall technical workflow is shown in Figure 2.

The proposed method consists of three major components, namely intrinsic parameter estimation of the tower-based camera, joint optimization of intrinsic and extrinsic parameters, and video geospatial mapping. The primary data sources used in this method include digital elevation model (DEM), high-resolution remote sensing imagery, and tower-based video data. First, the 3D GIS view window size is proportionally configured according to the resolution of the tower-based video frames, and the intrinsic parameter matrix  $K$  of the tower-based camera is calculated in combination with the estimated extrinsic parameters of the 3D GIS virtual camera. Second, under the condition that  $K$  is known, the initial values

of the extrinsic parameters are solved using the PnP method based on the corresponding control point pairs selected from the 3D GIS scene and the tower-based video. In addition, planar position constraints of the tower-based camera are introduced, and the intrinsic and extrinsic parameters are further optimized using a constrained gradient-based method. Finally, the optimized intrinsic and extrinsic parameters are used to configure the synchronized 3D GIS view, thereby achieving the spatial mapping and georeferencing from video frames to colored point clouds.



**Figure 2.** Overall methodological framework of the proposed video geospatial mapping approach. The blue, green, and orange shaded regions denote different stages of the framework, and arrows indicate the processing flow among the modules.

### 3.1. Intrinsic Parameter Estimation of Tower-Based Camera

The intrinsic parameters of the tower-based camera include the principal point, the principal distance, and the distortion coefficients  $d$ , as defined in Equation (1), where  $k_1$ ,  $k_2$ , and  $k_3$  denote the radial distortion coefficients, and  $p_1$  and  $p_2$  denote the tangential distortion coefficients.

$$d = \{k_1, k_2, k_3, p_1, p_2\} \quad (1)$$

Lens distortion is inherently present in practical imaging systems. However, for the natural-scene monitoring applications considered in this study, tower-based cameras are typically installed on high towers (generally exceeding 20 m in height). Owing to historical deployment conditions, accurate information regarding the camera model, lens specifications, and factory calibration parameters is often unavailable. In addition, conducting

on-site intrinsic calibration using standard calibration targets is generally impractical for such installations, making the reliable estimation of lens distortion parameters difficult. Therefore, this study focuses on engineering scenarios in which prior camera information is unavailable. Under this assumption, a zero-distortion pinhole camera model is adopted as an approximation, whereby all distortion coefficients are set to zero and are excluded from the subsequent optimization process. Based on this model, the camera intrinsic parameters are estimated by integrating the 3D GIS scene with tower-based video observations. The resulting intrinsic matrix is primarily parameterized by the principal point coordinates and focal length and is represented in the form of a  $3 \times 3$  matrix. This intrinsic matrix is subsequently used as the initial intrinsic parameter estimate for both the PnP-based pose initialization and the constrained gradient-based optimization procedure.

### 3.1.1. 3D GIS Window Size Setting

3D GIS refers to a geographic information system that integrates multi-source geospatial information, such as terrain elevation and remote sensing imagery, into a three-dimensional visualization scene, thereby supporting the representation, retrieval, and analysis of the spatial locations, geometries, and their relationships in the real world. Let the resolution of the current frame of the tower-based video be  $W \times H$ , and its aspect ratio be denoted as  $\alpha$ , as defined in Equation (2). By proportionally configuring the 3D GIS view window according to  $\alpha$ , the aspect ratio of the 3D GIS view window can be kept consistent with that of the tower-based video frame, thereby avoiding the interference of geometric distortions introduced by non-uniform scaling, as shown in Figure 3.

$$\alpha = W/H \quad (2)$$



**Figure 3.** 3D GIS view size configuration. (a) Video frame; (b) 3D GIS view.

### 3.1.2. 3D GIS Virtual Camera Extrinsic Parameter Estimation

The extrinsic parameters of the virtual camera in the 3D GIS environment mainly include the camera position  $(X_{vir}, Y_{vir}, Z_{vir})$ , tilt angle (Tilt), azimuth angle (Azimuth), and roll angle (Roll). First, the planar coordinates of the camera  $(X_{vir}, Y_{vir})$  are obtained from high-resolution remote sensing imagery. Let  $D_h$  denote the tower height and  $H_0$  represent the ground elevation at the camera location derived from the DEM. The camera elevation  $Z_{vir}$  is then calculated using Equation (3). Subsequently, the extrinsic parameters of the virtual camera are adjusted to align the current tower-based video frame with the 3D GIS view, thereby achieving an initial coarse registration between the two. Specifically, stable features that are clearly identifiable in both the video frame and the 3D GIS scene, such as road intersections and building edges, are selected as reference targets. Under the constraint of the tower location, the virtual camera position is further refined within a limited range, while the azimuth, tilt, and roll angles are iteratively adjusted to ensure that the spatial relationships of the reference features in the 3D GIS scene correspond as closely

as possible to those observed in the video frame. This procedure establishes an initial alignment between the video and the 3D GIS view.

$$Z_{\text{vir}} = H_0 + D_h \quad (3)$$

### 3.1.3. Tower-Based Camera Intrinsic Parameter Estimation

The intrinsic parameters of the tower-based camera mainly include the principal point and the principal distance  $f$ . According to the image dimensions  $(W, H)$  of the tower-based video frame, the principal point  $(C_x, C_y)$  is located at the image center. After the video frame had been matched with the 3D GIS scene as described in Section 3.1.2, the field of view (FOV) was obtained, and the initial value of the principal distance  $f$  was calculated using Equation (8), where  $f_x \approx f_y$  was generally assumed. Subsequently, the camera intrinsic matrix  $K$  was constructed from the principal distance  $f$  and the principal point  $(C_x, C_y)$ . This intrinsic matrix served only as the initial parameter for the subsequent PnP solution and constrained optimization, while the final camera intrinsic parameters were jointly estimated during the optimization process.

$$C_x = W/2 \quad (4)$$

$$C_y = H/2 \quad (5)$$

$$d = \sqrt{(W)^2 + (H)^2} \quad (6)$$

$$\text{FOV}_{\text{rad}} = \text{FOV} \cdot \pi / 180 \quad (7)$$

$$f_x \approx f_y = f = d/2 \tan(\text{FOV}_{\text{rad}}/2) \quad (8)$$

## 3.2. Optimization of Tower-Based Camera Intrinsic and Extrinsic Parameters

### 3.2.1. Acquisition of the Planar Position of the Tower-Based Camera and Control Point Coordinate Pairs

The acquisition of coordinate data mainly includes the extraction of the planar position coordinates of the tower-based camera and the construction of corresponding control point pairs. The former can be obtained by determining the planar location of the tower-based camera in the coordinate system from high-resolution remote sensing imagery, while the latter is achieved by selecting corresponding feature points from tower-based video frames and extracting their image coordinates, which are then matched with the corresponding three-dimensional geographic coordinates in the 3D GIS scene, thereby establishing the correspondence between image coordinates and 3D world coordinates, as shown in Figure 4. For a given point in the tower-based video frame, its pixel coordinates are denoted as  $P_i(u_i, v_i)$ . The corresponding 3D geographic coordinate is denoted as  $G_i(X_i, Y_i, Z_i)$ , which is retrieved from the 3D GIS environment.



**Figure 4.** Acquisition of corresponding control point coordinate pairs.



### 3.2.2. Tower-Based Camera Extrinsic Parameter Estimation Using the PnP Method

Under the condition that the intrinsic parameters of the tower-based camera are known, this study adopts the Perspective-n-Point (PnP) method to estimate the extrinsic parameters of the camera. First, based on the 3D GIS scene and the tower-based video frame, six or more pairs of control points are selected to obtain the corresponding 3D world coordinates and pixel coordinates, respectively. Let the 3D world coordinate of the  $i$ -th control point be denoted as  $G_i = [X_i, Y_i, Z_i]^T$ , and its corresponding pixel coordinate be denoted as  $P_i = [u_i, v_i]^T$ . The standard optimization form of the PnP method can be formulated as the minimization of the sum of squared reprojection errors over all control points, as defined in Equation (9).

$$\min_{R,t} \sum_{i=1}^N \| P_i - \hat{P}_i \|^2_2 \quad (9)$$

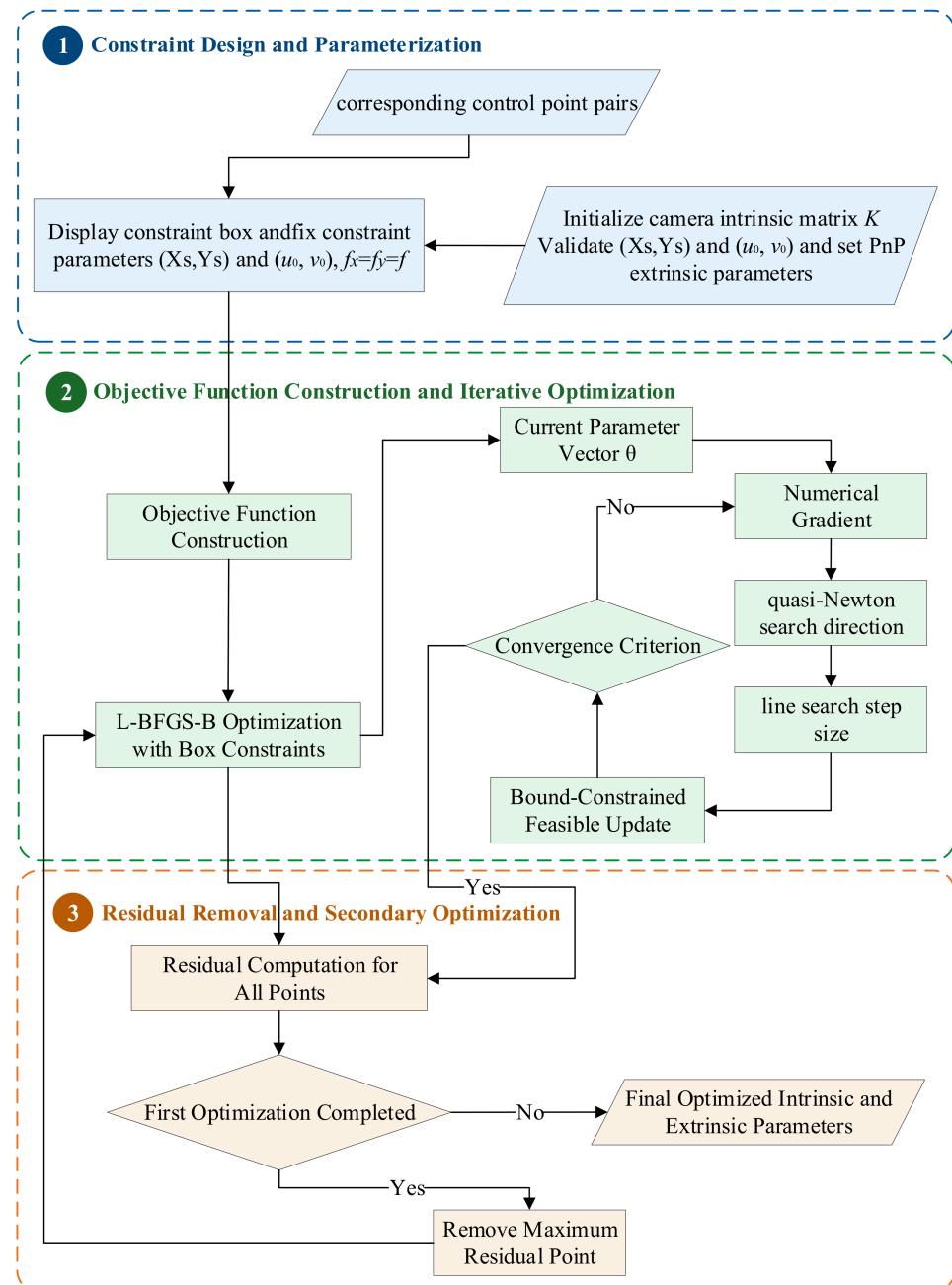
### 3.2.3. Constrained Gradient Optimization of Camera Intrinsic and Extrinsic Parameters

Although the traditional PnP calibration method can estimate the camera extrinsic parameters from known corresponding control point pairs, it is difficult to obtain accurate solutions for the intrinsic and extrinsic parameters in large-scale scenes with complex terrain. To improve the estimation accuracy and robustness, this study employs a scene-constrained L-BFGS-B optimization framework initialized with PnP estimates. By jointly optimizing the camera intrinsic and extrinsic parameters and incorporating an outlier-removal and re-optimization strategy, the proposed method further enhances the accuracy of parameter estimation for tower-based cameras. The proposed method mainly consists of three major processes, namely constraint formulation and parameterization, objective function construction and iterative optimization, and residual outlier removal with secondary optimization, as shown in Figure 5.

The first step is constraint design and parameterization. Under the condition that the correspondence between the corresponding control point pairs and reasonable initial values, such as the intrinsic matrix  $K$ , the planar position priors  $(X_s, Y_s)$ , the principal point  $(u_0, v_0)$ , and the PnP-based initial extrinsic parameters are known, a constrained nonlinear optimization model is constructed with the minimization of reprojection error as its core objective. The boundary constraint mechanism consists of two levels. The first imposes constraints on the focal length and camera orientation parameters. Specifically, the focal length is subject to box constraints, with its allowable range defined to cover plausible imaging scales while preventing non-physical solutions. The orientation angles are constrained within bounded intervals to ensure the validity of the rotation parameters and to avoid ambiguities arising from angular periodicity. The specific parameter bounds are listed in Table 1. During optimization, the L-BFGS-B algorithm enforces these boundary constraints, ensuring that all iterations remain within the feasible parameter space. The second component adopts a semi-constrained parameterization strategy. The camera planar position  $(X_s, Y_s)$  and principal point coordinates  $(u_0, v_0)$  are fixed throughout the optimization process. The camera planar position is obtained from high-resolution remote sensing imagery, while the principal point is assumed to coincide with the image center. In addition, the constraint  $f_x = f_y = f$  is imposed in accordance with the square-pixel assumption of the pinhole camera model. By incorporating reliable scene priors while reducing the number of unknown parameters, this strategy decreases parameter correlations and alleviates the ill-posedness of the optimization problem, thereby improving solution stability and robustness under conditions with a limited number of control points. Under the aforementioned constraints and parameterization scheme, the parameter vector to be estimated is denoted by  $\theta$ , as defined in Equation (10), where  $Ze$  denotes the optimized

camera elevation in the Web Mercator coordinate system, and  $A$ ,  $T$ , and  $R$  represent the Azimuth, Tilt, and Roll angles, respectively.

$$\theta = [Ze, f, A, T, R]^T \quad (10)$$



**Figure 5.** Workflow of the joint optimization of intrinsic and extrinsic parameters for tower-based cameras using constrained L-BFGS-B and outlier re-optimization.

The second step is objective function construction and iterative optimization. The optimization is performed using the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints (L-BFGS-B), where the normalized Euclidean norm of the pixel reprojection residuals over all control points is adopted as the objective function, as defined in Equation (11), where  $\pi(G_i; \theta)$  denotes the projection function jointly defined by the intrinsic matrix, the distortion model, and the attitude rotation parameters. As

described in Section 3.1, the distortion coefficients are fixed to zero and excluded from the optimization throughout the entire parameter estimation process.

$$\min_{\theta} F(\theta) = \frac{1}{N} \sqrt{\sum_{i=1}^N \|\pi(G_i; \theta) - u_i\|_2^2} \quad (11)$$

**Table 1.** Parameter bounds and constraints for the constrained optimization.

| Parameter                  | Description      | Lower Bound | Upper Bound |
|----------------------------|------------------|-------------|-------------|
| $Z_e(\text{m})$            | Camera elevation | None        | None        |
| $f_x = f_y = f(\text{px})$ | Focal length     | 500         | 5000        |
| Azimuth (rad)              | Azimuth angle    | $-\pi$      | $\pi$       |
| Tilt (rad)                 | Tilt angle       | $-\pi$      | $\pi$       |
| Roll (rad)                 | Roll angle       | $-\pi$      | $\pi$       |

At the  $k$ -th iteration, the algorithm constructs a low-rank approximation of the inverse Hessian matrix based on the current parameter  $\theta^k$ , its numerical gradient, and the historical gradient difference information, thereby obtaining the quasi-Newton search direction  $d^k$ . After the step size  $\alpha^k$  is determined through a line search procedure, the boundary feasibility operator  $\Pi_B(\cdot)$  is applied to ensure that the parameters satisfy the box constraints, as defined in Equation (12). Accordingly, the parameter vector  $\theta$  is estimated and updated collaboratively along a unified search direction rather than by component-wise descent with fixed step sizes, until the gradient norm and the variation of the objective function satisfy the convergence criteria.

$$\theta^{k+1} = \Pi_B(\theta^k + \alpha^k d^k) \quad (12)$$

Numerical gradients are computed using the finite-difference method, and the default line-search strategy of the L-BFGS-B solver is employed. To ensure sufficient convergence of the optimization process, the maximum number of iterations is set to 10,000, while the gradient convergence tolerance and objective function convergence tolerance are set to  $1 \times 10^{-8}$  and  $1 \times 10^{-12}$ , respectively. Because all the optimization parameters are confined to physically reasonable and finite ranges through box constraints, no additional parameter normalization or scaling transformation is applied during the optimization process. The optimization settings are summarized in Table 2.

**Table 2.** L-BFGS-B optimization settings.

| Parameter            | Value                      |
|----------------------|----------------------------|
| Gradient computation | Finite-difference gradient |
| Maximum iterations   | 10,000                     |
| Gradient tolerance   | $1 \times 10^{-8}$         |
| Function tolerance   | $1 \times 10^{-12}$        |
| Scaling              | None                       |

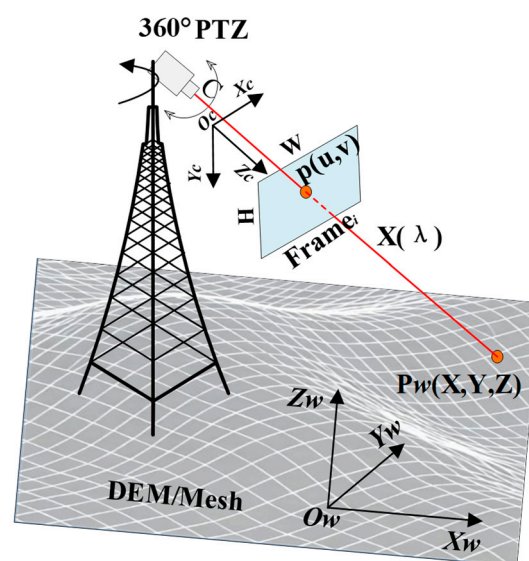
The third step is residual removal and secondary optimization. After the first optimization converges, the reprojection residuals of all control points are calculated. To suppress the influence of gross errors, the control point corresponding to the maximum residual is removed, and a second optimization is performed under the same initial conditions. This process further improves the stability and accuracy of the intrinsic and extrinsic parameter estimation, and outputs the final optimized intrinsic and extrinsic parameters after the secondary optimization. It should be noted that the control point exhibiting the largest

reprojection residual is not necessarily an erroneous observation. Rather than explicitly classifying such a point as an outlier, the objective of this strategy is to mitigate the influence of high-residual observations on the optimization process, thereby enhancing the robustness of the estimated camera parameters.

### 3.3. Video Geospatial Mapping

#### 3.3.1. Transformation from Video Frame Pixel Coordinates to 3D World Coordinates

Based on the optimized intrinsic and extrinsic parameters of the tower-based camera, let a pixel point  $p = (u, v)$  in a video frame  $Frame_i$  with width  $W$  and height  $H$ . Under the camera coordinate system  $O_c - X_c Y_c Z_c$  and the world coordinate system  $O_w - X_w Y_w Z_w$ , a spatial ray  $X(\lambda)$  is emitted from the camera optical center and passes through the pixel point. The intersection of this ray with the DEM or Mesh model yields the corresponding 3D world point  $P_w$ , thereby establishing a mapping relationship between pixel coordinates and 3D coordinates based on line-of-sight projection, as shown in Figure 6.



**Figure 6.** Video frame pixel-to-3D coordinate transformation.

Let the homogeneous pixel coordinate corresponding to the pixel point  $p$  be denoted as  $p' = [u, v, 1]^T$ ; according to the zero-distortion pinhole imaging model,  $p$  can be regarded as the intersection point of a viewing ray with the image plane under the camera coordinate system  $O_c - X_c Y_c Z_c$ . By applying the inverse intrinsic transformation to  $p'$ , the direction vector of the corresponding viewing ray in the camera coordinate system can be obtained and subsequently normalized, as defined in Equation (13).

$$d_c = \frac{K^{-1}p'}{\|K^{-1}p'\|} \quad (13)$$

where  $K$  denotes the optimized intrinsic matrix, and  $d_c$  represents the unit direction vector from the camera optical center to the pixel point  $p$  in the camera coordinate system. Next, the rotation matrix  $R$  is used to transform  $d_c$  into the world coordinate system  $O_w - X_w Y_w Z_w$ , thereby obtaining the corresponding viewing direction vector  $d_w$  of the pixel point  $p$  in the world coordinate system. Taking the optimized camera position as the starting point and  $d_w$  as the direction, a spatial ray parametric equation  $X(\lambda)$  is constructed, which originates from the camera optical center and passes through the pixel point  $p$ , as defined in Equations (14) and (15), where  $\lambda$  denotes the ray parameter along the viewing direction,  $C$  represents the camera center position in the world coordinate system, and



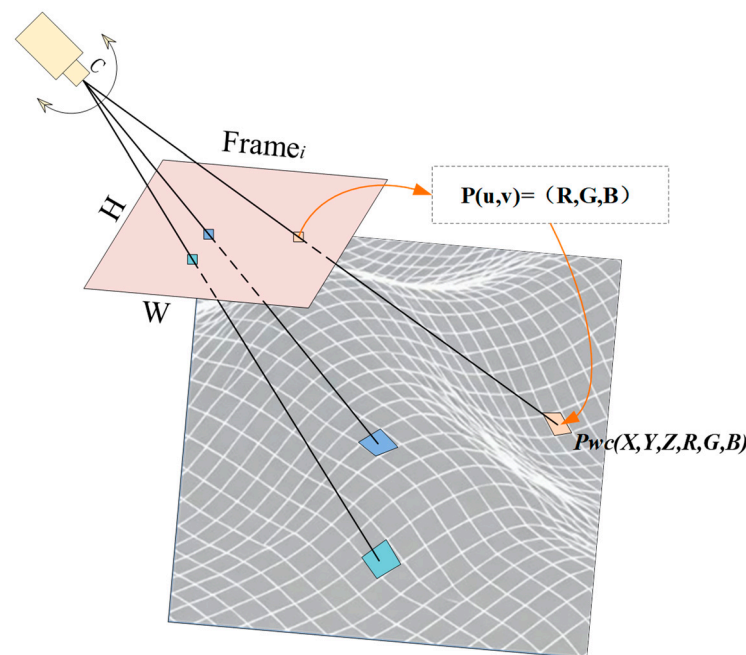
$R$  denotes the rotation matrix constructed from the camera attitude angles. Finally, the intersection point between the ray and the ground surface is computed in the 3D GIS scene using the DEM or Mesh model, yielding  $P_w = (X, Y, Z)$ , which represents the 3D coordinate corresponding to the pixel point  $p$ . By applying the above transformation to each pixel of the entire video frame, the mapping from full-frame pixel coordinates to 3D coordinates can be completed.

$$d_w = R^T d_c \quad (14)$$

$$X(\lambda) = C + \lambda d_w, \lambda > 0 \quad (15)$$

### 3.3.2. Transformation of Colored Point Clouds from Video Frames

The transformation of colored point clouds from video frames is achieved by projecting each pixel along its camera viewing ray onto the 3D ground surface and assigning the corresponding color attributes. Let a video frame be denoted as  $Frame_i$ . For each pixel point, the first valid intersection point  $P_w$  between its spatial ray and the DEM or Mesh model is computed. The RGB color values of the corresponding pixel  $P(u, v)$  from the video frame are then assigned to  $P_w$ , thereby generating a colored point  $P_{wc}$ . After traversing all pixel points in the entire frame, a colored point cloud is obtained, as shown in Figure 7.



**Figure 7.** Generation of colored point clouds from video frames.

## 4. Experiments and Analysis

### 4.1. Experimental Environment and Data

The experimental hardware environment of this study consisted of a 13th Gen Intel(R) Core (TM) i7-13620H processor, 32 GB RAM, and an NVIDIA GeForce RTX 4060 Laptop GPU (8 GB). The software and programming environment included ArcGIS Pro 3.5.2 and Python 3.11.

In terms of data, the tower-based camera video used in this study was collected from a region in Hunan Province, China, where the terrain is mainly characterized by mountainous areas and flat land. The 3D GIS scene was constructed using ArcGIS Pro 3.5.2, integrated with the Esri terrain service, 0.5 m resolution Esri online high-resolution remote sensing imagery, and a 30 m resolution digital elevation model (DEM). The video frame was selected from a tower-based video sequence captured from a single viewing

perspective without zoom operations, with a spatial resolution of  $1920 \times 1080$  pixels. Based on this dataset, 29 sets of corresponding control point pairs were collected. Combined with prior information, including the planar coordinates of the tower location, these data were used to perform camera intrinsic and extrinsic parameter optimization and subsequent video geospatial mapping experiments.

The experimental data were collected under stable meteorological conditions to minimize the effects of illumination variations and visibility fluctuations on camera calibration accuracy. In addition, the Web Mercator projection was adopted as the horizontal coordinate reference system, while elevations were referenced to the geoid and expressed as orthometric heights. This configuration ensured that the control points, camera poses, and 3D scene were processed and analyzed consistently within a unified spatial reference.

#### 4.2. Experimental Analysis

##### 4.2.1. Intrinsic Parameter Estimation of the Tower-Based Camera

For the tower-based camera A used in the experiment, the intrinsic parameters were estimated according to the method described in Section 3.1, where the main parameters include the focal length  $f$  and the principal point  $(C_x, C_y)$ . Given that the resolution of the video frame is  $1920 \times 1080$ , its aspect ratio  $\alpha$  is 1.777. To ensure consistency of the field of view, the size of the 3D GIS view window was adjusted to  $960 \times 540$ , thereby maintaining the same aspect ratio as the video frame. Subsequently, the pose parameters of the virtual camera were adjusted so that the 3D GIS view window was synchronized with the video frame view, as shown in Figure 8.



**Figure 8.** Synchronization between the camera video frame and the 3D GIS view: (a) video frame; (b) 3D GIS view.

After view synchronization was completed, the extrinsic parameters and field of view (FOV) of the virtual camera corresponding to the current 3D GIS scene were obtained. The detailed parameters are listed in Table 3.

**Table 3.** Estimated parameters of the virtual camera.

| $X_{\text{vir}}$ (m) | $Y_{\text{vir}}$ (m) | $H_0$ (m) | $D_h$ (m) | FOV (deg) | Tilt (deg) | Azimuth (deg) | Roll (deg) |
|----------------------|----------------------|-----------|-----------|-----------|------------|---------------|------------|
| 12,645,882.33        | 2,946,735.07         | 615.98    | 83.00     | 60.13     | −10.00     | 344.87        | 0.00       |

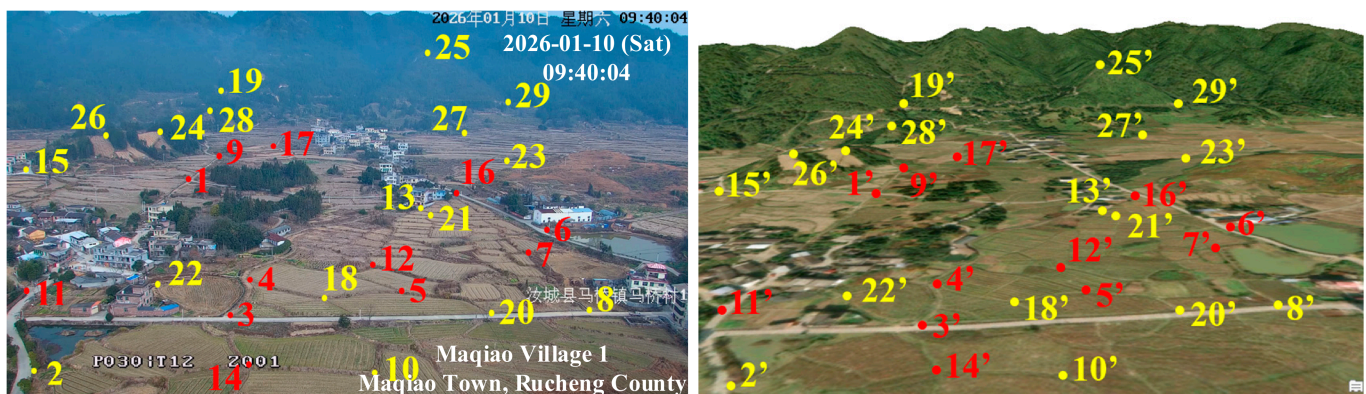
In this experiment, the principal point was initialized as the image center, i.e.,  $(C_x, C_y) = (960.00, 540.00)$ . According to the FOV values listed in Table 3, the focal length was calculated as  $f = 1902.70$  px based on the zero-distortion pinhole imaging model. Accordingly, the estimated intrinsic matrix of the tower-based camera is denoted as  $K$ , as given in Equation (16).

$$K = \begin{bmatrix} 1902.70 & 0 & 960.00 \\ 0 & 1902.70 & 540.00 \\ 0 & 0 & 1 \end{bmatrix} \quad (16)$$

#### 4.2.2. Intrinsic and Extrinsic Parameter Optimization of the Tower-Based Camera

To estimate the intrinsic and extrinsic parameters of the tower-based camera, corresponding control point pairs were established using 2D image points extracted from the current video frame and their corresponding 3D world points, yielding a total of 29 control points. Specifically, the image points  $P_i(u_i, v_i)$  were manually selected from the video frame, while the corresponding 3D world points  $G_i(X_i, Y_i, Z_i)$  were obtained from the 3D GIS scene.

The selection of control points followed the principles of stability, distinctiveness, and spatial uniformity. Feature points with clear geometric characteristics were selected within the area covered by the tower-based video, including road intersections and intersections of field boundaries. Their spatial distribution is shown in Figure 9.



**Figure 9.** Corresponding control points in the video frame and 3D GIS view.

Among the 29 control points, 12 were selected as calibration points for optimizing the intrinsic and extrinsic parameters of the tower-based camera, while the remaining 17 were used as independent check points for accuracy assessment. The complete control point dataset and its partitioning scheme are presented in Table 4.

**Table 4.** Pixel and 3D world coordinates of corresponding control points.

| Pt | Set | Pixel $u$ | Pixel $v$ | Web Mercator X Coordinate (m) | Web Mercator Y Coordinate (m) | Elevation (m) |
|----|-----|-----------|-----------|-------------------------------|-------------------------------|---------------|
| 1  | CAL | 507       | 477       | 12,645,543.96                 | 2,947,292.20                  | 600.96        |
| 2  | CHK | 75        | 1009      | 12,645,713.27                 | 2,946,929.08                  | 589.82        |
| 3  | CAL | 632       | 858       | 12,645,757.43                 | 2,946,999.92                  | 591.29        |
| 4  | CAL | 681       | 759       | 12,645,738.34                 | 2,947,052.38                  | 592.42        |
| 5  | CAL | 1103      | 785       | 12,645,821.54                 | 2,947,062.77                  | 592.56        |
| 6  | CAL | 1527      | 620       | 12,645,896.69                 | 2,947,212.60                  | 598.85        |
| 7  | CAL | 1467      | 681       | 12,645,888.16                 | 2,947,157.65                  | 596.43        |
| 8  | CHK | 1636      | 841       | 12,645,918.06                 | 2,947,062.04                  | 593.80        |
| 9  | CAL | 598       | 411       | 12,645,486.46                 | 2,947,452.05                  | 607.84        |
| 10 | CHK | 1033      | 1017      | 12,645,833.91                 | 2,946,968.98                  | 590.65        |
| 11 | CAL | 55        | 787       | 12,645,646.13                 | 2,946,991.44                  | 591.53        |
| 12 | CAL | 1033      | 719       | 12,645,799.76                 | 2,947,097.17                  | 594.21        |
| 13 | CHK | 1159      | 555       | 12,645,793.53                 | 2,947,256.33                  | 598.94        |
| 14 | CAL | 672       | 995       | 12,645,783.50                 | 2,946,961.57                  | 590.70        |
| 15 | CHK | 56        | 451       | 12,645,349.06                 | 2,947,280.10                  | 601.67        |

**Table 4.** *Cont.*

| Pt | Set | Pixel $u$ | Pixel $v$ | Web Mercator X Coordinate (m) | Web Mercator Y Coordinate (m) | Elevation (m) |
|----|-----|-----------|-----------|-------------------------------|-------------------------------|---------------|
| 16 | CAL | 1263      | 513       | 12,645,809.91                 | 2,947,334.33                  | 600.90        |
| 17 | CAL | 752       | 383       | 12,645,519.61                 | 2,947,551.23                  | 611.95        |
| 18 | CHK | 893       | 807       | 12,645,788.99                 | 2,947,038.17                  | 592.31        |
| 19 | CHK | 598       | 222       | 12,645,218.30                 | 2,947,921.93                  | 685.02        |
| 20 | CHK | 1364      | 849       | 12,645,873.05                 | 2,947,044.46                  | 592.59        |
| 21 | CHK | 1191      | 576       | 12,645,805.35                 | 2,947,239.21                  | 598.51        |
| 22 | CHK | 422       | 767       | 12,645,698.17                 | 2,947,024.27                  | 592.13        |
| 23 | CHK | 1405      | 425       | 12,645,841.34                 | 2,947,582.82                  | 608.42        |
| 24 | CHK | 428       | 342       | 12,645,400.51                 | 2,947,446.30                  | 629.89        |
| 25 | CHK | 1182      | 119       | 12,645,577.97                 | 2,948,309.25                  | 777.32        |
| 26 | CHK | 274       | 353       | 12,645,346.63                 | 2,947,406.17                  | 629.30        |
| 27 | CHK | 1283      | 345       | 12,645,747.20                 | 2,947,858.34                  | 624.59        |
| 28 | CHK | 568       | 281       | 12,645,302.29                 | 2,947,734.42                  | 649.02        |
| 29 | CHK | 1402      | 257       | 12,645,786.71                 | 2,948,225.78                  | 672.85        |

Based on the estimated camera intrinsic matrix  $K$  and the 12 control point pairs used for tower-based camera calibration, the initial camera extrinsic parameters were first estimated using the PnP method. The resulting parameters are listed in Table 5.

**Table 5.** Initial camera parameters used for optimization.

| Parameter      | Value   |
|----------------|---------|
| $Z_{init}(m)$  | 707.76  |
| $f_{init}(px)$ | 1902.70 |
| Azimuth (deg)  | −18.89  |
| Tilt (deg)     | 79.11   |
| Roll (deg)     | 0.79    |

Subsequently, the planar coordinates of the tower location ( $X_s, Y_s$ ) obtained from the Esri high-resolution online imagery in ArcGIS Pro were incorporated as prior information, with coordinates of (12,645,889.09, 2,946,768.35). Meanwhile, following the constraint mechanisms described in Section 3.2, a constrained L-BFGS-B optimization approach, together with an outlier rejection strategy, was employed to jointly optimize the camera intrinsic and extrinsic parameters. The final optimized parameters are presented in Table 6.

**Table 6.** Optimization results of tower-based camera parameters.

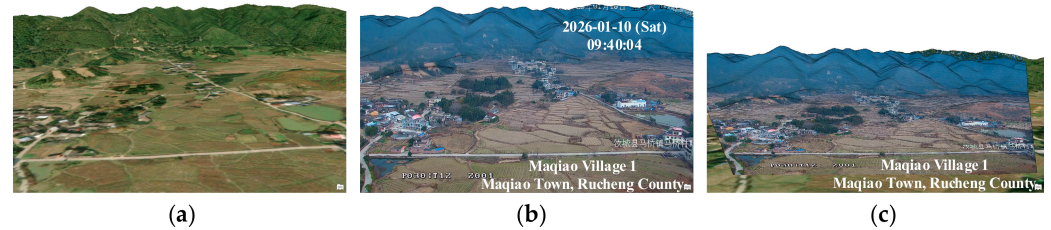
| $Z_e$ (m) | Azimuth (deg) | Tilt (deg) | Roll (deg) | $f$ (px) |
|-----------|---------------|------------|------------|----------|
| 695.97    | −18.19        | 79.28      | 1.22       | 1680.60  |

#### 4.2.3. Video Geospatial Mapping Results

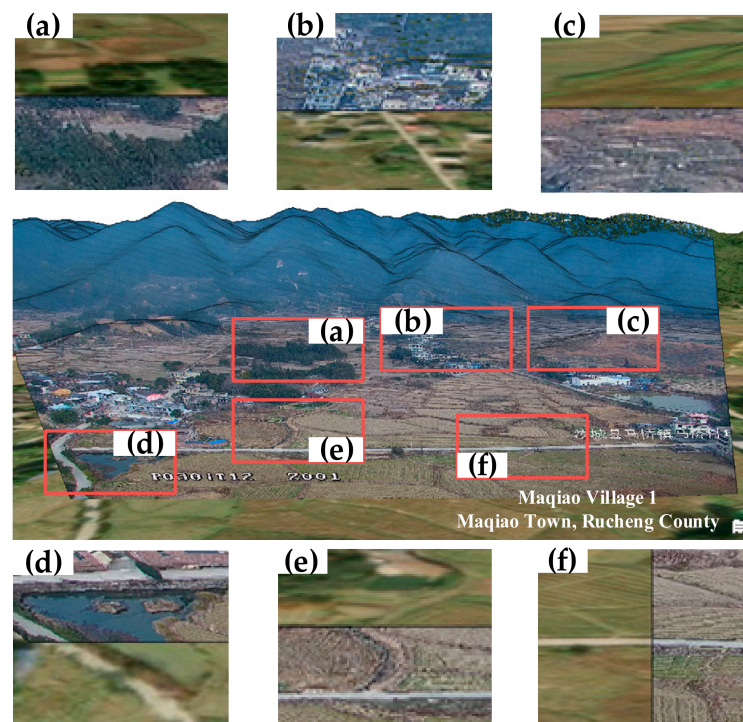
After obtaining the optimized intrinsic and extrinsic parameters of the tower-based camera, the camera settings were applied to further adjust the 3D GIS view. By integrating the current video frame with the corresponding 3D GIS scene, a colorized point cloud was generated, as shown in Figure 10. The point cloud coordinates were derived from the 3D GIS scene, while the color information was assigned from the video frame.

To further evaluate the spatial consistency of the video geospatial mapping results at the local scale, several representative regions were selected from the overlay of the generated colored point cloud and the 3D GIS scene for detailed visual comparison, as shown in Figure 11.





**Figure 10.** Visualization of the 3D GIS view, colored point cloud, and their overlay. (a) 3D GIS view; (b) colored point cloud generated from the corresponding video frame; (c) overlay of the two. All visualizations were rendered from a local 3D scene in ArcGIS Pro using a perspective projection; therefore, a fixed map scale does not accurately represent the spatial extent within the three-dimensional visualization.



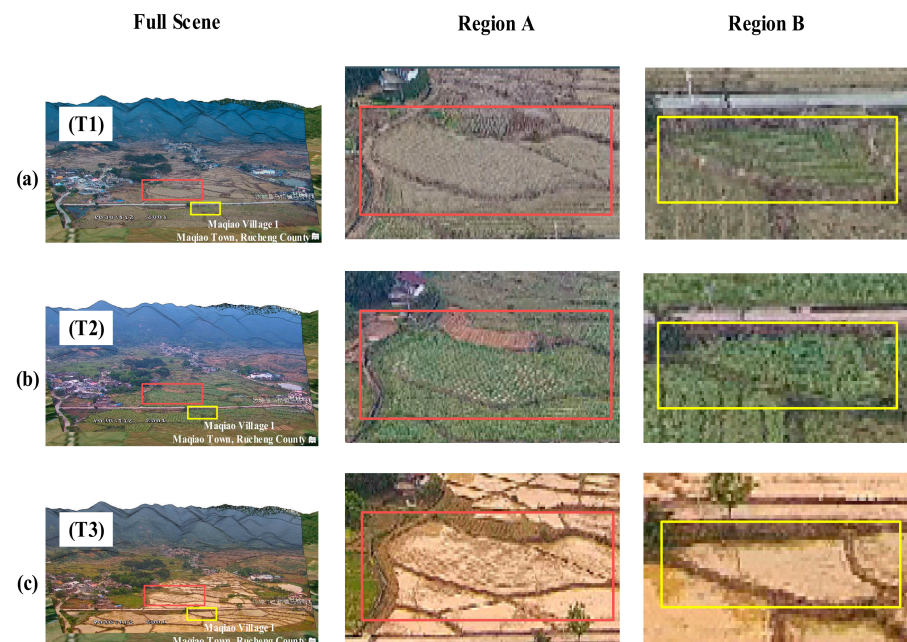
**Figure 11.** Local comparison of overlay results. The center panel shows the overall overlay result, with six representative regions highlighted by red rectangles. Subfigures (a–f) present enlarged views of representative areas, including (a) forest areas, (b) buildings, (c) slopes, (d) water bodies, (e) farmland, and (f) roads. All visualizations were rendered from a local 3D scene in ArcGIS Pro using a perspective projection; therefore, a fixed map scale does not accurately represent the spatial extent within the three-dimensional visualization.

Each subfigure shows a comparison between the colored point cloud and the corresponding 3D GIS scene within the same spatial extent. The qualitative results indicate that linear and areal features with strong geometric constraints, such as roads, field boundaries, and building outlines, exhibit good spatial correspondence in both position and overall orientation across the study area. Compared with roads and field boundaries, the mapping consistency of forested areas and buildings is relatively weaker at finer spatial scales. This discrepancy may be attributed to the complex and repetitive texture patterns of tree canopies, the limited representation of height and façade information in the 3D scene, and the restricted spatial resolution of the video imagery. Overall, at a coarse spatial scale, the distribution of mapped features derived from the video frame remains largely consistent with the corresponding regions in the 3D GIS scene. These observations provide



intuitive evidence supporting the reliability of the estimated camera parameters and the effectiveness of the proposed video geospatial mapping approach.

Video geospatial mapping not only is used for the visualization of individual frames but, more importantly, enables the real-time projection of dynamic image information acquired by tower-based cameras into a unified three-dimensional geographic coordinate framework. This capability allows results from different acquisition times to be analyzed and compared within a common spatial reference system. Figure 12 presents a temporal comparison of local details from the video geospatial mapping results for the same area, providing an intuitive representation of parcel-scale temporal variations and generating multi-temporal datasets with explicit geospatial references.



**Figure 12.** Temporal comparison of local details in video geospatial mapping results for the same area. Rows (a–c) correspond to three different time periods (T1, T2, and T3). Columns show the full scene (left), Region A (middle), and Region B (right). Red and yellow boxes in the full scene indicate the locations of the corresponding enlarged regions. All visualizations were rendered from a local 3D scene in ArcGIS Pro using a perspective projection; therefore, a fixed map scale does not accurately represent the spatial extent within the three-dimensional visualization.

To a certain extent, these results demonstrate that the proposed camera parameter estimation and video geospatial mapping workflow can effectively organize multi-temporal video information within a GIS environment and facilitate its integration with other geospatial datasets. Consequently, the proposed approach provides a foundation for applications such as crop monitoring, field management, and land-cover dynamics analysis by supporting the generation, organization, and visualization of georeferenced multi-temporal video data.

#### 4.2.4. Comparison of Video Geospatial Mapping Accuracy Among Different Methods

To compare the performance of different calibration methods in camera parameter estimation and video geospatial mapping, comparative experiments were conducted using a common initial intrinsic matrix  $K$  and the same set of 12 calibration points. The evaluated methods included PnP, EPnP, RANSAC-PnP, DLT, SQPnP, and the proposed method. The evaluation metrics comprised reprojection error and 3D spatial error, for which the maximum error, mean error, and root mean square error (RMSE) were calculated at both calibration points and check points. For all methods, camera parameters were estimated

exclusively from the 12 calibration points, whereas the 17 independent check points were reserved solely for validation. The camera parameters estimated from the 12 calibration points are presented in Table 7, and the corresponding comparison of reprojection errors among different methods is reported in Table 8.

**Table 7.** Camera parameters estimated by different methods.

| Method     | K Required? | Initial $f$ (px) | Initial Principal Point (px) | Solved $f$ (px)                      | Solved Principal Point (px) | Estimated Camera Position (Xe, Ye, Ze) (m) | Estimated Camera Pose (Azimuth, Tilt, Roll) (deg) |
|------------|-------------|------------------|------------------------------|--------------------------------------|-----------------------------|--------------------------------------------|---------------------------------------------------|
| PnP        | Yes         | 1902.7           | (960,540)                    | 1902.7                               | (960,540)                   | 12,645,910.57,<br>2,946,724.42, 707.77     | −18.90°, 79.12°,<br>0.79°                         |
| RANSAC-PnP | Yes         | 1902.7           | (960,540)                    | 1902.7                               | (960,540)                   | 12,645,918.60,<br>2,946,707.55, 714.50     | −19.15°, 78.92°,<br>0.10°                         |
| EPnP       | Yes         | 1902.7           | (960,540)                    | 1902.7                               | (960,540)                   | 12,645,585.51,<br>2,947,660.48, 711.68     | 160.48°, 75.27°,<br>179.54°                       |
| DLT        | No          | <b>x</b>         | <b>x</b>                     | 1725.5 ( $f_x$ )<br>1460.8 ( $f_y$ ) | (1065,350)                  | 12,645,934.81,<br>2,946,764.32, 723.38     | −17.17°, 84.34°,<br>1.61°                         |
| SQPnP      | Yes         | 1902.7           | (960,540)                    | 1902.7                               | (960,540)                   | 12,645,914.84,<br>2,946,721.75, 707.56     | −19.26°, 79.25°,<br>0.59°                         |
| Ours       | Yes         | 1902.7           | (960,540)                    | 1680.6                               | (960,540)                   | 12,645,889.10,<br>2,946,768.36, 695.97     | −18.19°, 79.28°,<br>1.22°                         |

“**x**” indicates unavailable or non-estimable parameters under the corresponding method. For DLT, the focal lengths  $f_x$  and  $f_y$  are derived from the decomposition of the estimated projection matrix P and are therefore reported separately. For the remaining methods, the square-pixel assumption ( $f_x = f_y = f$ ) is adopted, and a single focal length parameter  $f$  is reported.

**Table 8.** Comparison of reprojection errors (in pixels) among different methods.

| Method     | RMSE_CAL | Mean_CAL | Max_CAL | RMSE_CHK | Mean_CHK | Max_CHK |
|------------|----------|----------|---------|----------|----------|---------|
| PnP        | 14.73    | 12.85    | 25.14   | 31.40    | 27.78    | 57.61   |
| EPnP       | 747.07   | 463.13   | 2023.33 | 2018.02  | 1451.90  | 4834.39 |
| RANSAC-PnP | 25.33    | 16.77    | 71.83   | 40.05    | 31.09    | 106.01  |
| DLT        | 7.34     | 6.32     | 13.80   | 20.40    | 14.05    | 65.58   |
| SQPnP      | 15.90    | 12.61    | 33.20   | 32.99    | 28.03    | 61.94   |
| Ours       | 6.10     | 5.22     | 11.88   | 10.56    | 8.84     | 24.67   |

As shown in Table 8, substantial differences in reprojection accuracy are observed among the evaluated methods. Among them, EPnP exhibits the largest errors, reaching the order of  $10^2$ – $10^3$  pixels. Combined with the camera parameter estimates presented in Table 7, these results indicate that the pose estimated by EPnP deviates significantly from the actual camera configuration, causing most reprojected points to fall outside the image boundaries. This suggests that EPnP fails to produce reliable parameter estimates under the current experimental conditions; EPnP was therefore excluded from the subsequent accuracy analysis.

Among the remaining methods, the proposed method consistently achieves the lowest or near-lowest errors across all evaluation metrics. DLT achieves better performance, while PnP and SQPnP exhibit comparable error levels. In contrast, RANSAC-PnP produces larger reprojection errors. For the independent check points, its RMSE is reduced by 66.4%, 73.6%, 68.0%, and 48.3% compared with those of PnP, RANSAC-PnP, SQPnP, and DLT, respectively. For the calibration points, the proposed method achieves an RMSE of 6.10 px, which is lower than that of all competing methods. Furthermore, all methods exhibit larger errors at the check points than at the calibration points, indicating greater uncertainty for samples that are not involved in parameter estimation. Overall, the proposed method demonstrates superior accuracy under the current experimental conditions.

To further evaluate the impact of different calibration methods on video geospatial mapping performance, 17 independent check points were selected for validation. The maximum error, mean error, and root mean square error (RMSE) of the 3D spatial error were computed. Table 9 compares the 3D spatial errors obtained by different methods at both the calibration points and the independent check points.

**Table 9.** Comparison of 3D geospatial errors (in meters) among different methods.

| Method     | MaX_CAL | Mean_CAL | RMSE_CAL | Max_CHK | Mean_CHK | RMSE_CHK |
|------------|---------|----------|----------|---------|----------|----------|
| PnP        | 20.60   | 6.61     | 9.02     | 287.75  | 36.54    | 80.85    |
| RANSAC-PnP | 28.21   | 7.59     | 10.70    | 335.33  | 42.38    | 90.02    |
| EPnP       | ✗       | ✗        | ✗        | ✗       | ✗        | ✗        |
| DLT        | 12.42   | 4.44     | 5.41     | 107.86  | 23.36    | 40.70    |
| SQPnP      | 16.83   | 6.33     | 8.41     | 259.91  | 35.72    | 73.85    |
| Ours       | 6.28    | 2.78     | 3.32     | 36.19   | 8.12     | 12.45    |

“✗” indicates that the method fails to produce valid results. EPnP is excluded from further quantitative comparisons.

As shown in Table 9, the proposed method consistently achieves relatively low errors on both types of points, demonstrating favorable performance in video geospatial mapping. For the calibration points, the proposed method yields overall lower error levels compared with the competing approaches. Specifically, the RMSE is reduced by 63.2%, 69.0%, 38.6%, and 60.5% relative to PnP, RANSAC-PnP, DLT, and SQPnP, respectively. For the independent check points, the proposed method also achieves the lowest RMSE of 12.45 m, outperforming all baseline methods. These results indicate that the camera parameters estimated by the proposed method not only improve fitting accuracy at the control points but also maintain satisfactory accuracy at the check points that are not involved in parameter estimation.

#### 4.2.5. Ablation Study and Analysis of the Impact of Outlier Removal

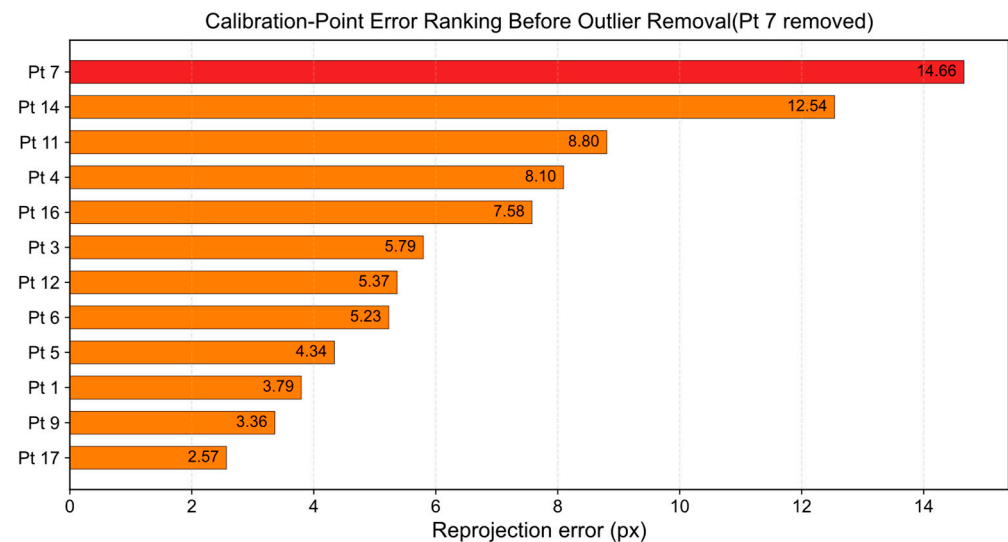
To investigate the effects of the constrained optimization and outlier removal strategies on calibration accuracy, three ablation experiments were designed. These experiments sequentially compare the PnP baseline, PnP combined with constrained L-BFGS-B optimization, and the full framework incorporating both constrained optimization and outlier rejection. The detailed experimental configurations are provided in Table 10.

**Table 10.** Ablation settings used for evaluating the contribution of each module.

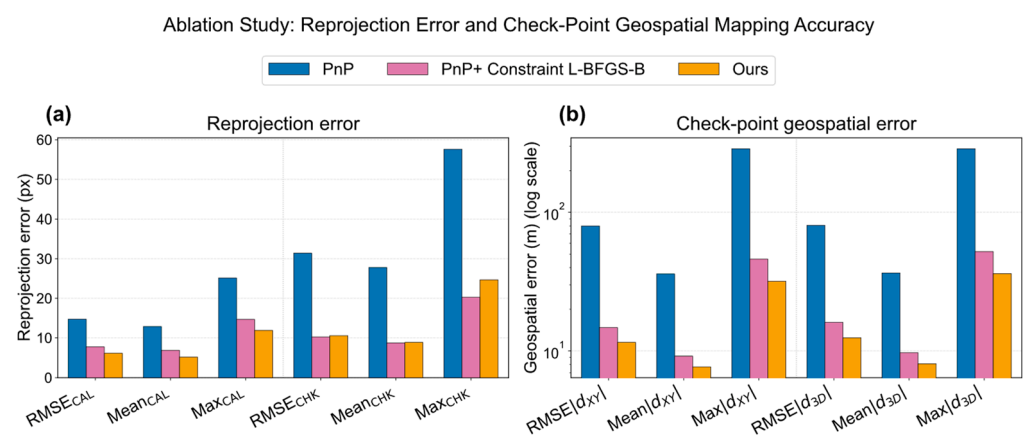
| Configuration   | PnP Initialization | Constrained Optimization | Outlier Removal | Number of Calibration Points |
|-----------------|--------------------|--------------------------|-----------------|------------------------------|
| PnP             | ✓                  | ✗                        | ✗               | 12                           |
| PnP+ Constraint | ✓                  | ✓                        | ✗               | 12                           |
| L-BFGS-B        | ✓                  | ✓                        | ✓               | 11                           |
| Ours            | ✓                  | ✓                        | ✓               | 11                           |

“✓” indicates that the corresponding module is enabled, while “✗” indicates it is not used.

Figure 13 illustrates the distribution and ranking of reprojection errors for the 12 calibration points after the first stage of constrained optimization, based on which potential outliers are identified. Meanwhile, Figure 14 further compares the reprojection errors and georeferencing errors at independent check points under different experimental settings, in order to evaluate the impact of each strategy on camera parameter estimation accuracy and video geospatial mapping performance.



**Figure 13.** Reprojection error ranking of calibration points before outlier removal. The control points are sorted in descending order of reprojection error, and Pt7 is identified as a potential outlier based on the highest residual.



**Figure 14.** Ablation study on reprojection error and geospatial mapping accuracy under different configurations. (a) Reprojection error comparison on the control-point set and independent check-point set. (b) Geospatial error comparison of independent check points under different methods (logarithmic scale).

As shown in Figure 14, compared with the PnP baseline method, the introduction of constrained L-BFGS-B joint optimization significantly reduces the reprojection errors on both the calibration set and the check set, while also improving the horizontal and 3D spatial errors of the independent check points. On this basis, further incorporation of an outlier removal strategy leads to a continued decrease in reprojection error on the calibration set, whereas the reprojection error on the check set shows no consistent improvement, indicating that this strategy mainly enhances local fitting capability with limited impact on unseen samples. Nevertheless, both the horizontal error and the 3D spatial error are further reduced and achieve the best performance among the three experimental configurations. Overall, these results indicate that, under the current experimental conditions, the proposed optimization and outlier handling strategies effectively improve the overall accuracy of video geospatial mapping.



## 5. Conclusions and Discussion

Aiming at the need for high-precision georeferencing of tower-based video, this study addresses the difficulty of obtaining intrinsic and extrinsic camera parameters and the impracticality of conventional field calibration in tower-based monitoring scenarios. To this end, we proposed a 3D GIS and gradient-based optimization integrated geospatial mapping method. Within a unified 3D GIS framework, the method focuses on the optimization of camera parameters and the representation of video geospatial mapping, with corresponding algorithm design and experimental validation carried out.

The main conclusions of this study are summarized as follows. (1) A camera parameter optimization method is developed for tower-based scenarios under small-sample control point conditions. Starting from the estimated intrinsic matrix  $K$  and corresponding control point pairs, the PnP method is first used to obtain initial extrinsic parameters. This is followed by a constrained L-BFGS-B optimization incorporating planar priors of the tower location, explicit box constraints, and a semi-constrained parameterization strategy. In addition, an outlier removal and re-optimization scheme is introduced to reduce the influence of gross errors, enabling more reliable estimation of camera parameters. (2) Under the current experimental conditions, the proposed method helps improve the accuracy of camera parameter estimation and mitigate the influence of outlier control points on the optimization results, thereby providing effective support for the implementation of video geospatial mapping. (3) Based on the tower-based camera parameters and the digital elevation model (DEM), colorized point clouds are generated through the transformation from image coordinates to 3D world coordinates. These point clouds are then overlaid onto the 3D GIS environment, providing a fundamental basis for subsequent collaborative analysis with existing geospatial datasets.

However, several limitations of the proposed method should be acknowledged. First, the experimental validation was conducted in a single study area with a single observation viewpoint, and its applicability and stability under multi-region, multi-view, and heterogeneous tower-based camera scenarios remain to be further investigated. In addition, the proposed method adopts a zero-distortion pinhole camera model and relies on DEM and remote sensing imagery to construct the 3D GIS environment. Therefore, several factors, including lens distortion, control point selection errors, meteorological conditions, the lack of high-resolution 3D scene information, and the quality of the underlying geospatial data, may introduce uncertainties and affect the final results. Future work will incorporate more diverse experimental scenarios and systematic uncertainty analyses to further evaluate the influence of these factors on camera parameter estimation and video geospatial mapping performance, and to assess the generalization ability and robustness of the proposed method.

**Author Contributions:** Xianguo Ling was responsible for writing the original draft, conducting experiments and method validation, performing data analysis, and visualization. Xingguo Zhang contributed to conceptualization and methodology design, and provided writing guidance and supervision. Xin Li and Xiangfei Meng were responsible for data processing and manuscript revision and editing. All authors participated in the editing of the paper. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (Grant No. 41401436), the Innovation and Development Special Project of China Meteorological Administration (CXFZ2025Q009), the Natural Science Foundation of Henan Province (Grant No. 202300410345), and the Graduate Scientific Research Innovation Fund of Xinyang Normal University (Grant No. 2025KYJJ84).

**Data Availability Statement:** The original data presented in the study are openly available at <https://github.com/Lxg006/TowerCameras> (accessed on 9 July 2026).

**Acknowledgments:** The author appreciates the editors and reviewers for their comments, suggestions, and valuable time and efforts in reviewing this manuscript.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

- Li, D.; Xu, X.; Shao, Z. On Geospatial Information Science in the Era of IoE. *Acta Geod. Cartogr. Sin.* **2022**, *51*, 1–8.
- Chen, J.; Zhai, X.; Wang, Y.H.; Ruan, Y.Z.; Liu, W.Z.; Li, X.B.; Wu, H.T.; Cao, C.H.; Wang, D.H.; Zhao, P.J.; et al. The Basic Connotation and Development Directions of Smart Territory. *J. Spatio-Temporal Inf.* **2024**, *31*, 157–167. [\[CrossRef\]](#)
- Lewis, P.; Fotheringham, S.; Winstanley, A. Spatial video and GIS. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 697–716. [\[CrossRef\]](#)
- Malis, E.; Vargas, M. *Deeper Understanding of the Homography Decomposition for Vision-Based Control*; INRIA Research Report RR-6303; HAL: Lyon, France, 2007; p. 90. Available online: <https://hal.science/inria-00174036v3> (accessed on 9 July 2026).
- Xie, Y.; Wang, M.; Liu, X.; Mao, B.; Wang, F. Integration of Multi-Camera Video Moving Objects and GIS. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 561. [\[CrossRef\]](#)
- Zhang, X.; Shi, X.; Luo, X.; Sun, Y.; Zhou, Y. Real-Time Web Map Construction Based on Multiple Cameras and GIS. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 803. [\[CrossRef\]](#)
- Wang, P.; Jiao, B.; Yao, P.; Wei, X.; Zhang, A. A Robust Direct Linear Transformation for Camera Pose Estimation Using Points. *Image Vis. Comput.* **2024**, *141*, 104883. [\[CrossRef\]](#)
- Lepetit, V.; Moreno-Noguer, F.; Fua, P. EPnP: An Accurate O(n) Solution to the PnP Problem. *Int. J. Comput. Vis.* **2009**, *81*, 155–166. [\[CrossRef\]](#)
- Gao, X.-S.; Hou, X.-R.; Tang, J.; Cheng, H.F. Complete Solution Classification for the Perspective-Three-Point Problem. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 930–943. [\[CrossRef\]](#)
- Zhang, X.; Meng, X.; Zhang, L.; Ling, X.; Yang, S. Orthographic Video Map Generation Considering 3D GIS View Matching. *ISPRS Int. J. Geo-Inf.* **2025**, *14*, 398. [\[CrossRef\]](#)
- Zhang, Z. A Flexible New Technique for Camera Calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [\[CrossRef\]](#)
- Tsai, R.Y. A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses. *IEEE J. Robot. Autom.* **1987**, *3*, 323–344. [\[CrossRef\]](#)
- Heikkila, J.; Silven, O. A Four-Step Camera Calibration Procedure with Implicit Image Correction. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Juan, PR, USA, 17–19 June 1997; pp. 1106–1112. [\[CrossRef\]](#)
- Yang, C.J.; Wang, W.; Hu, Z. An Active Vision Based Camera Intrinsic Parameters Self-Calibration Technique. *Chin. J. Comput.* **1998**, *21*, 428–435. Available online: <https://robot.sia.cn/cn/article/id/2343> (accessed on 9 July 2026).
- Hartley, R.; Hayman, E.; de Agapito, L.; Reid, I. Camera Calibration and the Search for Infinity. In Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV), Kerkyra, Greece, 20–27 September 1999; pp. 510–517. [\[CrossRef\]](#)
- Caprile, B.; Torre, V. Using Vanishing Points for Camera Calibration. *Int. J. Comput. Vis.* **1990**, *4*, 127–139. [\[CrossRef\]](#)
- Triggs, B. Autocalibration and the Absolute Quadric. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Juan, PR, USA, 17–19 June 1997; pp. 609–614. [\[CrossRef\]](#)
- Xu, G.; Zhang, Z. Algebraic Derivation of the Kruppa Equations and a New Algorithm for Self-Calibration of Cameras. *J. Opt. Soc. Am. A* **1999**, *16*, 2419–2424. [\[CrossRef\]](#)
- Pollefeys, M.; Koch, R.; Van Gool, L. Self-Calibration and Metric Reconstruction Inspite of Varying and Unknown Intrinsic Camera Parameters. *Int. J. Comput. Vis.* **1999**, *32*, 7–25. [\[CrossRef\]](#)
- Schneider, N.; Piewak, F.; Stiller, C.; Franke, U. RegNet: Multimodal Sensor Registration Using Deep Neural Networks. In 2017 *IEEE Intelligent Vehicles Symposium (IV)*; IEEE: New York, NY, USA, 2017; pp. 1803–1810. [\[CrossRef\]](#)
- Lv, X.; Wang, B.; Dou, Z.; Ye, D.; Wang, S. LCCNet: LiDAR and Camera Self-Calibration Using Cost Volume Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Nashville, TN, USA, 19–25 June 2021; pp. 2894–2901. [\[CrossRef\]](#)
- Liao, K.; Nie, L.; Huang, S.; Lin, C.; Zhang, J.; Zhao, Y.; Gabbouj, M.; Tao, D. Deep learning for camera calibration and beyond: A survey. *arXiv* **2023**, arXiv:2303.10559. [\[CrossRef\]](#)
- Tan, Z.; Zhang, X.; Teng, S.; Wang, L.; Gao, F. A Review of Deep Learning-Based LiDAR and Camera Extrinsic Calibration. *Sensors* **2024**, *24*, 3878. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jung, H.; Kim, N.; Kim, J.; Park, J. Targetless LiDAR-Camera Calibration with Neural Gaussian Splatting. *IEEE Robot. Autom. Lett.* **2026**, *11*, 4777–4784. [\[CrossRef\]](#)

25. Zhang, X.; Liu, X.; Wang, S.; Yang, L. Mutual Mapping between Surveillance Video and 2D Geospatial Data. *Geomat. Inf. Sci. Wuhan Univ.* **2015**, *40*, 1130–1136. [[CrossRef](#)]
26. Shao, Z.; Li, C.; Li, D.; Altan, O.; Zhang, L.; Ding, L. An Accurate Matching Method for Projecting Vector Data into Surveillance Video to Monitor and Protect Cultivated Land. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 448. [[CrossRef](#)]
27. Zhang, X.; Li, Q.; Li, X.; Shi, X.; Li, H. A New Method for Orthographic Video Map Construction Based on Video and GIS. *Geocarto Int.* **2023**, *38*, 2289450. [[CrossRef](#)]
28. Milosavljević, A.; Rančić, D.; Dimitrijević, A.; Predić, B.; Mihajlović, V. A Method for Estimating Surveillance Video Georeferences. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 211. [[CrossRef](#)]
29. Han, S.; Dong, X.; Hao, X.; Miao, S. Extracting Objects' Spatial–Temporal Information Based on Surveillance Videos and the Digital Surface Model. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 103. [[CrossRef](#)]
30. Liu, Z.; Dai, Z.; Li, C.; Xuebao, C. A Fast Fusion Object Determination Method for Multi-Path Video and Three-Dimensional GIS Scene. *Acta Geod. Cartogr. Sin.* **2020**, *49*, 632. [[CrossRef](#)]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.