WLTOG: An optimization approach for wild large-range target omnidirectional geolocation based on monocular PTZ camera

Xinxin Zhou1,2,3,4 | LinWang Yuan2,3 | Hua Tang5 | Yi Huang1 | Ningyu Feng2,3,4 | Changbin Wu2,3,4

Abstract
The pan-tilt-zoom (PTZ) cameras are increasingly popular in urban management and resource protection due to their omnidirectional coverage and reasonable cost. Traditionally, the research of monocular camera geolocation focuses on small-range ones with a central projection camera under ideal conditions, making it challenging to meet the requirements of omnidirectional coverage and large-range geolocation for PTZ cameras. In this article, a wild large-range target omnidirectional geolocation (WLTOG) algorithm was proposed. Firstly, the WLTOG algorithm is used to transform the target in a single image to the camera’s optical axis, thereby obtaining the camera posture value. After that, by using a high-order overdetermined equation, it fits the nonlinear mapping relationship between multiple pairs of PTZ camera postures and the geospatial coordinates. At last, it realizes the bidirectional calculation between pixel coordinates and geospatial coordinates. The experimental PTZ camera is mounted on a communication tower in a wild environment. Therefore, this method can be utilized to construct an omnidirectional geolocation relationship that meets the requirements of omnidirectional geolocation within 945 m based on unknown accurate internal and external orientation elements. The WLTOG algorithm is conducive to the
1 | INTRODUCTION

Surveillance videos, thought to be real-time geospatial data in cities (Li et al., 2020), contain much visual information (Chen et al., 2020; Kiran et al., 2018). However, most surveillance videos lack geospatial information, such as geographical location, attribute characteristics, and temporal characteristics, due to the absence of mapping relationships with the physical world (Xie et al., 2021), which limits the geospatial application of surveillance videos. Numerous methods have been proposed for target geolocation, including monocular vision (Lategahn & Stiller, 2014), binocular vision (Deng et al., 2020), and multi-vision (Brickwedde et al., 2019). Comparatively, targets can be located through one shot by monocular vision, thereby reducing hardware cost and deployment complexity in the wild environment. Nowadays, monocular vision is widely used in intelligent vehicles (Sivaraman & Trivedi, 2013), robot navigation (Mur-Artal et al., 2015), three-dimensional (3D) scene reconstruction (Saxena et al., 2007), industrial detection, and many other fields (Park et al., 2020). Monocular cameras can be classified into three types based on their coverage orientations, i.e., central perspective projection, pan-tilt-zoom (PTZ), and fisheye (Arroyo et al., 2020; Wang et al., 2015). Due to the omnidirectional coverage and low cost, omnidirectional PTZ cameras have been widely employed in urban governance, forest fire protection, traffic supervision, and other various application contexts. Additionally, the PTZ camera installed at a high altitude can be rotated to the region of interest with a wide observation range and is kind of in situ sensor considered to encompass proximal sensing (Grillo et al., 2021).

Limited by omnidirectional coverage, large-range, and the topographical complexity of the field environment, conventional monocular geolocation algorithms are unable to match the monocular PTZ camera geolocation criteria. Existing methods of target geolocation and mapping with monocular vision are mainly applied to central perspective projection cameras (Chen et al., 2016; Simonelli et al., 2019). Homography transformation can realize the transformation and inverse transformation of video into geographic space through perspective and inverse perspective transformation matrices, respectively. This method requires high accuracy of camera elements of interior and exterior orientation. As most cameras lack of geospatial coordinates and the fact that camera posture information is independent of the world coordinate system, it is challenging to meet the requirements of the wild environment. Furthermore, the homography transformation is hard to fulfill the geolocation of the PTZ camera in omnidirectional under the continuous rotating state due to each homography matrix corresponding to each direction. Pixel feature matching approaches realize the mapping between image pixels and geographical coordinates (Karlsson et al., 2005). For example, the SIFT+RANSAC method has to render 3D geographic scenes in real-time and then simulate camera framing based on 3D geographic scenes before making feature matches with actual collected video data (Goforth & Lucey, 2019). The advantage of this method is that it does not need the camera’s spatial coordinates and real-time posture, while the disadvantage is that the performance of 3D geographic scene computation will plummet suddenly when multiple cameras are used in parallel (Zhou et al., 2021). 3D object detection applications for monocular vision-based localization focus on short- or middle-range areas and indoor environments, including the 3DoF or even 6DoF pose of targets in indoor scenarios (Rad & Lepetit, 2017; Song et al., 2020), the short-range field of mobile robot navigation (Zhang et al., 2020), and middle-range 3D object detection applications (Li et al., 2019; Ma et al., 2019). The latest works primarily attempt to overcome the issue of large-scale monocular geolocation in SLAM and autonomous driving fields. A novel monocular geolocation algorithm was proposed by Gao et al. (2021), namely monocular global geolocation for outdoor long-range targets, which fails to consider the omnidirectional implications, is mainly applied in outdoor environments with targets located at an extended range of up to 150m. Hence, three challenges of monocular PTZ camera geolocation are as follows.
1. **It is difficult to geolocate a vast area.** Because of the absence of depth information provided by monocular PTZ vision, the large-range recovery from 2D to 3D of monocular vision using the perspective projection model geolocation is limited. Existing studies of monocular vision geolocation are easily influenced by camera optical distortion. Moreover, as the goal distance increases, the error rises exponentially.

2. **Omnidirectional geolocation is challenging in the outdoors.** There are also varying topography fluctuations and surface textures around the camera, which place more demands on the generalization capacity and robustness of omnidirectional geolocation mapping model of monocular PTZ vision.

3. **It is hard to calibrate precise elements of interior and outside orientation in the wild.** The underlying premise of current monocular vision is the specification of camera installation and the proper calibration of interior and exterior orientation elements. However, limited by artificial installation standardization and natural accessibility, it is impractical to calibrate the exact interior and exterior orientation elements of the PTZ camera promptly in the field environment.

This study proposes a wild large-range target omnidirectional geolocation method (WLTOG) to address the large-range and omnidirectional geolocation problem of monocular PTZ cameras. The WLTOG algorithm first transforms the target in a single image to the optical axis, and obtains camera posture values; it then fits the nonlinear mapping relation between camera posture values and world coordinates by using the overdetermined equation. Finally, the bidirectional calculation between the target pixel and world coordinates is achieved (image-to-world projection, world-to-image back-projection). The WLTOG method satisfies the requirements of large-range and multidirectional monocular PTZ camera geolocation in the wild environment, which overcomes the limitations of the perspective projection model geolocation and homography matrix transformation (Hartley & Zisserman, 2000). The contributions and innovations of this study are summarized in the following four points:

1. **Large-range, omnidirectional geolocation.** The WLTOG algorithm converts the classic binary homography relation-ship between image pixels and world coordinates to a triadic relation among image pixels, PTZ posture values, and world coordinates. The geolocation distance can exceed 945 m with an accuracy of <20 m.

2. **High adaptability.** The WLTOG algorithm has low requirements for camera hardware installation standards in the wild, and has strong adaptability for geolocation. Also, this geolocation method is well-suited for regions with varying topography fluctuations.

3. **Less world-image pairs.** Traditional visual transformation and homography matrix calculation methods need a mass of world-image pairs, which is a time-consuming and complex task, while this article describes a method that manages to simplify the world-image pairs data of the monocular vision. Taking advantage of the rotation characteristics of the PTZ camera, the target pixel coordinate can be converted to the camera PTZ posture value. Then, the high-order overdetermined equation is established based on the world-image pairs between the PTZ posture values and the world coordinates.

4. **Reducing the error distribution of the azimuth.** During the world-to-image back-projection, the mapping model ignores the symmetry of azimuth rotation. Consequently, this study proposes a strategy for grouping world-image pairs, which divides homogeneous terrain feature areas into one group, obtaining multiple groups of over-determined equations to reduce the error propagation of long-distance and omnidirectional geolocation.

Section 2 summarizes the related work in this article, Section 3 explains the proposed method, and Section 4 describes the experimental results. Finally, Section 5 summarizes the research findings.

## 2 | RELATED WORK

Monocular geolocation continues to arouse the interest of professionals in surveying and geoscience, as well as computer vision. The former involves geolocation and geometric measurement (including geometric measure-
ment of the ground target on a single image, single-point geolocation, and GIS augmented video surveillance). The latter includes object detection and geolocation (including 3D object detection algorithm, unmanned driving, and monocular robot vision geolocation). This work focuses on the field of geographic information so that we will present the systematic overview of ground target geometric measurement on a single image and GIS augmented video surveillance.

2.1 Geometric measurement of the target on a single image

In surveying and geoscience, geometric measurement of the target on a single image is equivalent to monocular geolocation. It primarily uses the scene structure information in the 2D image to determine the 3D geometric information in each image pixel. The methods of geometric measurement on a single image are classified into the direct approaches based on homology transformations and the indirect approaches based on invariants.

Criminisi et al. (1999) addressed the challenge of measuring in the world plane using central perspective images, and successfully predicted the uncertainty of these measurements. They employed a homography transformation to translate the locations from the image to the world plane, and then used the first-order model to forecast uncertainty. Considering the uncertainties of the image input points and the homography matrix, they utilized a linear distortion model and estimated projection parameters to minimize the projection error in the global plane, demonstrating the accuracy of the first-order analysis. Among these, the direct approaches based on the homography matrix may be used to determine the plane coordinates of a point on the plane by re-establishing the imaging relationship between the image and the space for geometric measurement. Depending on reference points and associated lines, these approaches need to build homography matrixes primarily. Typically, the homography matrix is solved using the direct linear transformation approach (Wang, 2011; Xie et al., 2017). The homography matrix has been extensively employed in visual measuring (Anagnostopoulos et al., 2017), 3D reconstruction (Zhu et al., 2020), and image mosaicking (Tian et al., 2020). This approach requires many reference points and a significant amount of human labor to achieve monocular PTZ camera geolocation, which is only suitable for the geolocation of targets in a limited number of specified PTZ rather than random postures. Given the multi-view capabilities of the monocular PTZ camera and its ability to cover various scenes, the direct measuring approach based on homology transformation is inefficient. For continuous PTZ cameras, Lisanti et al. (2016) suggested an adaptive calibration approach which is ideal for real-time target tracking due to its capability to collect keyframes from images with varying PTZ values, build world-image pairings in keyframes and 3D geographic scenes, and estimate the homography transition between keyframe images and 3D scenes. An image normalization approach based on map data is presented to address this issue. A nonlinear perspective correction model is established by creating a single link between video images and geographical space. We may calculate the homography matrix by using selected reference points between the picture and the world to establish the perspective correction model, which is used to determine the single-pixel actual size in the map. The suggested approach for perspective correction is applied to detect moving objects (Lin et al., 2020).

The indirect methods based on invariants, such as the single-point geolocation (SPG) algorithm, are mainly used to construct the nonlinear mapping relationship between the elements of interior and exterior orientation of the camera and the geographic scene information for geolocation. The SPG algorithm collects the real-time elements of the internal and external orientation of the PTZ camera and digital elevation terrain information and the geospatial coordinates of the target are determined by using the exhaustive search method (Ning et al., 2010). The SPG algorithm meets the accuracy requirements of forest fire prevention geolocation; however, for particular terrain, the calculation results appear to have excess prediction points, which still needed manual secondary screening, reducing geolocation efficiency. Besides, the SPG algorithm solves the problem of watchtower target geolocation, but the target is confined to the center sight of the camera, and fails to accomplish the complete mapping between image pixels and geospatial space. Karlsson et al. (2005) established an environment map according to local invariant features through target recognition and particle filter spatial localization based on SIFT to realize robot self-localization and relocation.
Arroyo et al. (2020) estimated the posterior probability of calibration data parameters by selecting reference points in satellite images and fisheye panorama. However, this method increases the error due to the breakdown of calibrated internal orientation elements as equipment ages.

2.2 | GIS-augmented video surveillance

The GIS-augmented video surveillance method aims to connect the real-world geographical scene viewed by users with a computer-generated virtual environment to enable real-time geolocation of various targets inside the monitoring system (Lewis et al., 2011; Milosavljević et al., 2010; Zhang et al., 2013). Milosavljević et al. (2010) proposed an approach for geographical data registration and developed the GeoScopeAVS system, combining geographic information systems with surveillance video. The system converts the PTZ camera image to the virtual geographic scene's position, direction, and field of view, establishes an observer viewpoint model, and aligns the 3D GIS with the camera view to identify geospatial objects in the camera image, registering the view parameters of multiple cameras with the 3D scene. Milosavljević et al. (2016) examined integrating spatial reference information and GIS data into surveillance footage. Xie et al. (2017) proposed the Video Geographic Information System (V-GIS), which, based on a moving object extraction method and a fusion pattern for GIS, can analyze the possible applications of GIS integration and moving objects. Shao et al. (2020) suggested an accurate matching approach for projecting vector data onto surveillance video by selecting a sufficient number of reference points from 2D GIS data. In addition, automatic feature matching is employed to align GIS data with PTZ video frames. Visual detection covers a two-dimensional vector region and monitors natural resource consumption. Zhang et al. (2021) proposed a multi-camera (OS-CDD) algorithm to map the objects in the multi-camera video with overlapping areas to geographic space and performed the object selection. Han et al. (2022) utilized surveillance videos and the digital surface models to solve the integration problem between 3D geographic information and surveillance video, providing a new approach for the multi-dimensional expression of geographic information and object measurement.

Based on the existing research, limited by the installation condition and omnidirectional computing complexity of the PTZ camera in the real-world, scholars conduct less research on monocular PTZ camera geolocation under natural conditions, instead, they focus on close-up and single-direction monocular geolocation of central perspective project cameras and fisheye cameras under ideal conditions.

3 | METHODOLOGY

3.1 | Technology roadmap

The transformation process of the PTZ camera’s image to world coordinate is nonlinear (Lin et al., 2020) and the pixel coordinates of the same target in different PTZ parameter images bring the challenge of an explosion of mapping related to the traditional single image to the solution of the homography matrix. Facing a large number of world-image pairs and homography matrices required by a single image in solving the homography transformation under multiple PTZ views, the traditional homography transformation is considered to be not suitable for the geolocation of the PTZ camera. Therefore, an omnidirectional world-image association approach is urgently required after dimension reduction and simplification.

In this article, the WLTOG method is proposed to address the large-range and omnidirectional geolocation of monocular PTZ cameras in the wild environment. In Figure 1, firstly, the PTZ camera is used to collect video data which is encoded and transformed by the video server, and transmitted to the monitoring client; secondly, the PTZ camera’s control command is sent through the computing server, and the bidirectional geolocation calculation is accomplished by using the WLTOG algorithm including four steps:
Step 1: Image pixels conversion. The camera posture can be rectified by transforming the target to the center of view based on the constraint relationship between the image pixel coordinates of the target and camera posture parameters.

Step 2: Picking world-image pairs. First, the computing server controls the PTZ camera's rotation to ensure that the target is in the center of view. Second, we search the geographical scene for the same target. And last, we manually choose several world-image pairs \((p, t, x, y, z)\).

Step 3: Building a bidirectional mapping model for world-image pairs. The bidirectional mapping model contains image-to-world projection \((p, t) \rightarrow (x, y, z)\) and world-to-image back-projection \((x, y, z) \rightarrow (p, t)\).

Step 4: The strategy for world-image pair grouping. When solving the world-to-image back-projection model, the reference points are divided into multiple groups according to the azimuth for higher accuracy of back-projection, considering the differences in omnidirectional terrain and the error accumulation characteristics in the rotation process of the PTZ camera.

3.2 Image pixels conversion

To reduce the complexity of omnidirectional geolocation, the traditional mapping relationship between image pixels and world coordinates of a single target can be transformed into the relationship among image pixels, PTZ postures.\(^2\)

**FIGURE 1** Technology roadmap.
and world coordinates. The first mapping step between image pixels and PTZ postures is called image pixels conversion. In addition, the bidirectional mapping model can be used to describe the mapping relationship between PTZ postures and world coordinates.

In this article, an image pixel conversion method was proposed to rotate the target’s image pixel into the optical axis and obtain the camera PTZ posture. According to the geometric constraint relationship between the image and elements of both interior and exterior orientation of the camera (Figure 2), when the target is at the arbitrary location of the image, the corrected PTZ posture of the camera, after converting the target to the image center can be calculated based on the horizontal and vertical viewing angle conversion.

The physical information related to the camera imaging chip includes $W$ and $H$, and $\mu$ (the physical size of a single pixel). The horizontal and vertical field of view ($\text{FOV}_H$, $\text{FOV}_V$) can be calculated as shown in Equations (1) and (2):

$$\text{FOV}_H = 2 \times \arctan \left( \frac{W \cdot \mu}{2f} \right)$$

$$\text{FOV}_V = 2 \times \arctan \left( \frac{H \cdot \mu}{2f} \right)$$

where $f$ stands for the focal length. In Figure 2, the target is within the view, the point pixel coordinates are $(i, j)$, and the pixel variation from the target point to the center is $(\Delta i, \Delta j)$ (Equations 3 and 4):

$$\Delta i = \frac{W}{2} - i \cdot \mu$$

$$\Delta j = j \cdot \mu - \frac{H}{2}$$

![Figure 2](image-url)  
**Figure 2** Geometric constraints relation between an image pixel and central image point. World coordinate system ($X - Y - Z$), PTZ camera coordinate system ($P - T - R$, azimuth-tilt-roll), and image coordinate system ($W - H$, width-height of imaging chip).
The horizontal and vertical fields of view are \((\text{FOV}_H, \text{FOV}_V)\). The equations used to calculate the change of the field of view \((\Delta a_h, \Delta a_v)\) based on \((\Delta i, \Delta j)\) are:

\[
\Delta \text{FOV}_H = \arctan\left(\frac{\Delta j^* W}{2 \tan\left(\frac{\text{FOV}_H}{2}\right)}\right)
\]

\[
\Delta \text{FOV}_V = \arctan\left(\frac{\Delta i^* H}{2 \tan\left(\frac{\text{FOV}_V}{2}\right)}\right)
\]

After angle transformation, the target turn to the center of view, the new camera posture \((p, t)\) is shown in Equations 7 and 8:

\[
p = p_0 + \Delta \text{FOV}_H
\]

\[
t = t_0 + \Delta \text{FOV}_V
\]

where \(p_0\) and \(t_0\) refer to the azimuth angle and the pitch angle of the camera in the camera coordinate system.

### 3.3 Picking world-image pairs

It is essential to manually choose the world-image pairs to establish the bidirectional mapping model between PTZ posture and world coordinates. The primary data for choosing world-image pairs include the image of the camera, the geographic scene composed of the remote sensing image with very high resolution and the digital elevation model (DEM). Building inflection points and crossroads, for example, should be prioritized as reference objects because they are unobstructed with apparent textural elements, and stay still for an extended of time. The world-image pairs are recorded as \((p, t, x, y, z)\), and the steps to select the reference points are illustrated in Figure 3.

Step 1: We control the PTZ camera’s movement, aligning the center point of the camera with a target visible on the ground, and obtain the current camera posture value \((p, t)\).

Step 2: We also locate the corresponding target in the geographical scene, raise the query, and obtain the world coordinates of target A \((x_A, y_A)\).

Step 3: We obtain the elevation \(z_A\) of target A from DEM data in the geographic scene.

Step 4: When the visual field is located in regions where the texture's structural characteristics are not readily apparent, reference points can be supplemented by using the rotating reference points method, which is to rotate the specified reference points to a reference-free area. The application premise is that the area’s terrain fluctuates slightly, and the construction process is shown in Part 4 of Figure 3.

Step 4.1: As shown in Step 4 of Figure 3, the geographic coordinates \(O(x_O, y_O)\) of the camera and those of the initial reference point A \((x_A, y_A)\) are known. Point B is the reference point after rotating from point A. Firstly, we have to calculate the distance \(d\) (Equation (9)) between two points and the angle \(\alpha\) (Equation (10)) between the OA line and the north direction:

\[
d = \sqrt{(x_A - x_O)^2 + (y_A - y_O)^2}
\]

\[
\alpha = \begin{cases} 
\arcsin \frac{|y_A - y_O|}{d}, & x_O - x_A \geq 0, y_O - y_A \geq 0 \\
180 - \arcsin \frac{|y_A - y_O|}{d}, & x_O - x_A \leq 0, y_O - y_A \geq 0 \\
180 + \arcsin \frac{|y_A - y_O|}{d}, & x_O - x_A \leq 0, y_O - y_A \leq 0 \\
360 - \arcsin \frac{|y_A - y_O|}{d}, & x_O - x_A \geq 0, y_O - y_A \leq 0
\end{cases}
\]
Step 4.2: Set the rotation angle as positive in the clockwise direction. The rotation angle $\beta$ is known. Then rotate the angle $\beta$ clockwise from Point A to reach Point B in the fourth quadrant. The quadrant $q$ of the rotated point and the corresponding angle in the first quadrant $\gamma$ can be obtained through calculation. The calculation equations are expressed in Equations (11) and (12):

$$q = \left\lfloor \frac{(\alpha + \beta)}{90} \right\rfloor \mod 4$$  \hspace{1cm} (11)

$$\gamma = (\alpha + \beta) \mod 90$$  \hspace{1cm} (12)

Step 4.3: $q = 0, 1, 2, 3$ corresponds to the first, fourth, third, and second quadrants, respectively. According to the above equation, Point B is in the fourth quadrant, and the corresponding Point $B'$ is in the first quadrant. $y_1, y_2, y_3$ and $y_4$ refer to the corresponding $\gamma$ when rotating to different quadrants. According to angle $\gamma$ and distance
**3.4 Building bidirectional mapping model for world-image pairs**

The bidirectional mapping model between PTZ posture and world coordinates is established by using world-image pairings, which can be regressed to generate a mapping relation. In the equations $AX = b$, $A = (a_i)$, $b$ refers to a known $m$-dimension vector, and $x$ stands for the solution of the $n$-dimension vector. In the case $m > n$, the number of equations is higher than that of the independent variables, and this kind of equation is overdetermined. Because there is no exact solution for overdetermined equations, we find the least square or closest solution instead (Grafarend & Awange, 2012). The high-order overdetermined equation has a strong fitting ability, and the changing hyperplane relationship can still be fitted in the case of omnidirectional and a small number of world-image pairs.

### 3.4.1 Image-to-world projection

The transformation from the pixel coordinates of the target in the image into geographical coordinates through high-order overdetermined equation, or from $(p, t)$ to $(x, y, z)$, is defined as image-to-world projection. The equations for a world-image pair $(p_0, t_0, x_0, y_0, z_0)$, for example, are as follows:

$$a_0 t_0^4 + b_0 p_0^3 + c_0 t_0^3 p_0 + d_0 p_0^3 t_0 + e_0 t_0^2 p_0^2 + f_0 t_0^3 p_0^3 + g_0 p_0^5 + h_0 t_0^7 p_0 + i_0 t_0^7 p_0 + j_0 t_0^7 p_0 + k_0 p_0^7 + l_0 t_0 p_0 + m_0 t_0 + n_0 p_0 = x_0 \quad (16)$$

$$a_1 t_0^4 + b_1 p_0^3 + c_1 t_0^3 p_0 + d_1 p_0^3 t_0 + e_1 t_0^2 p_0^2 + f_1 t_0^3 p_0^3 + g_1 p_0^5 + h_1 t_0^7 p_0 + i_1 t_0^7 p_0 + j_1 t_0^7 p_0 + k_1 p_0^7 + l_1 t_0 p_0 + m_1 t_0 + n_1 p_0 = y_0 \quad (17)$$

$$a_2 t_0^4 + b_2 p_0^3 + c_2 t_0^3 p_0 + d_2 p_0^3 t_0 + e_2 t_0^2 p_0^2 + f_2 t_0^3 p_0^3 + g_2 p_0^5 + h_2 t_0^7 p_0 + i_2 t_0^7 p_0 + j_2 t_0^7 p_0 + k_2 p_0^7 + l_2 t_0 p_0 + m_2 t_0 + n_2 p_0 = z_0 \quad (18)$$

with 4-order that is determined by the fitting effect of trail reference points (Figures 4a1,a2,b1,b2). We obtain the matrix from Equation (19) by expanding the above equations into multiple world-image pairs:
FIGURE 4 Order fitting analysis diagram of the high-order overdetermined equation. This figure shows the data distribution of \((x, y, z)\) and \((p, t)\) and the cross-section of the fitting curve equation between them. From (a1,a2,b1,b2), it can be seen that \((p, t)\) and \((x, y, z)\) can be well reduced when the order of the structural equation is greater than or equal to 3. Considering the fitting effect when the pan value is too small or too large, the 4-order structural equation is selected. Similarly, combined with (c1,c2,d1,d2), although the fitting effect of each order structural equation is similar, only the 3-order structural equation can well fit the pan value and tilt value when x is too large or too small.
\[
\begin{bmatrix}
t_0^4 & p_0^4 & \cdots & \cdots & t_0^1 & p_0^1 \\
t_1^4 & p_1^4 & \cdots & \cdots & t_1 & p_1 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
p_{i-4}^4 & t_{i-4}^4 & \cdots & \cdots & p_{i-1} & t_{i-1} \\
p_i^4 & t_i^4 & \cdots & \cdots & p_i & t_i \\
\end{bmatrix}
\]

\[
X_1 = \begin{bmatrix}
a_0 & a_1 & a_2 \\
b_0 & b_1 & b_2 \\
\vdots & \vdots & \vdots \\
m_0 & m_1 & m_2 \\
n_0 & n_1 & n_2 \\
\end{bmatrix}
\]

among them,

\[
X_1 = \begin{bmatrix}
x_0^3 & y_0^3 & \cdots & \cdots & y_0 & z_0 \\
x_1^3 & y_1^3 & \cdots & \cdots & y_1 & z_1 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
x_{i-4}^3 & y_{i-4}^3 & \cdots & \cdots & y_{i-1} & z_{i-1} \\
x_i^3 & y_i^3 & \cdots & \cdots & y_i & z_i \\
\end{bmatrix}
\]

For the 42 unknown variables, at least 14 sets of world-image pairs are required to identify the solution \(X_1\). When the number of world-image pairs exceeds 14, the matrix becomes overdetermined.

### 3.4.2 World-to-image back-projection

World-to-image back-projection refers to the transformation from the geographic coordinates of the target into the center of the image, that is, from \((x, y, z)\) to \((p, t)\), and constructing a high-order overdetermined equation. Taking a world-image pair \((p_0, t_0, x_0, y_0, z_0)\) as an example, and the equations are as follows:

\[
a_3 x_0^3 + b_3 y_0^3 + c_3 z_0^3 + d_3 x_0^2 y_0 + e_3 x_0^2 z_0 + f_3 y_0^2 z_0 + g_3 z_0^2 x_0 + h_3 z_0^2 y_0 + i_3 z_0^2 t_0 + j_3 x_0 y_0 z_0 + k_3 x_0 + l_3 y_0 + m_3 z_0 = p_0 \tag{20}
\]

\[
a_4 x_0^3 + b_4 y_0^3 + c_4 z_0^3 + d_4 x_0^2 y_0 + e_4 x_0^2 z_0 + f_4 y_0^2 y_0 + g_4 y_0^2 x_0 + h_4 z_0^2 x_0 + i_4 z_0^2 y_0 + j_4 x_0 y_0 z_0 + k_4 x_0 + l_4 y_0 + m_4 z_0 = t_0 \tag{21}
\]

with 3-order that is determined by the fitting effect of trail reference points (Figures 4c1,c2,d1,d2). We also obtain the matrix from Equation 22 by expanding world-image pairs:

\[
X_2 = \begin{bmatrix}
x_0^3 & y_0^3 & \cdots & \cdots & y_0 & z_0 \\
x_1^3 & y_1^3 & \cdots & \cdots & y_1 & z_1 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
x_{i-4}^3 & y_{i-4}^3 & \cdots & \cdots & y_{i-1} & z_{i-1} \\
x_i^3 & y_i^3 & \cdots & \cdots & y_i & z_i \\
\end{bmatrix}
\]
among them, 

\[
X_2 = \begin{bmatrix}
    a_3 & a_4 \\
    b_3 & b_4 \\
    \vdots & \vdots \\
    \vdots & \vdots \\
    l_3 & l_4 \\
    m_3 & m_4
\end{bmatrix}
\]

Similarly, to solve the 26 unknown variables in \(X_2\), at least 13 sets of world-image pairs are required.

### 3.5 The strategy for world-image pairs grouping

In the world-to-image back-projection process, the symmetry of azimuth rotation is ignored in the high-order overdetermined equation. Consequently, this article proposes a world-image pairs grouping approach to optimize the world-to-image back-projection mapping model to boost the calculation accuracy of value \(p\).

As shown in Figure 5a, the \(y\)-axis corresponds to the geographic due north direction, and the camera azimuth corresponding to the due north direction is set as 0°. Besides, \(O\) stands for the PTZ camera position, points 1, 2, 3, and 4 refer to the reference points, and Point 5 is the prediction point. The geographical location distribution and corresponding azimuth are shown in Figure 5a. For Point 5, the points 1, 2, and 3 are close, but the difference of the values of \(p\) is 147°, 185°, and 117°, respectively, while Point 4 is the furthest, and the difference of value \(p\) is only 0°. Because the geographical coordinates are used to establish the model when predicting the value \(p\) by using the high-order overdetermined equation, the prediction results of Point 5 are affected by points 1, 2, and 3, and errors exist in the prediction results of value \(p\). To solve this problem, we developed a world-image pairs grouping strategy in which the reference points are divided into groups according to azimuth when establishing the world-to-image back-projection mapping model. As illustrated in Figure 5a, after the dotted line division, the reference points on both sides establish the mapping model, respectively, such that the prediction Point 5 is unaffected by Points 1, 2, and 3.
The grouping method can be divided into various groups, i.e., 0~180° and 180~360°, or more groups according to the azimuth. Moreover, it can be divided in conformity with the terrain.

Because of the reference point grouping strategy, it is impossible to determine which group of reference points should be selected to build the mapping model with only knowing the geographical coordinates of the prediction points in the construction of the world-to-image back-projection. To identify which bidirectional mapping model should be utilized, the rough value $p$ corresponding to the prediction points should be computed first. This method is used to calculate the horizontal and vertical angle components through the camera position Point $O (x_0, y_0, z_0)$ and the coordinates of the reference Point $S (x_a, y_a, z_a)$, and then the value $p$ of the reference point through the azimuth $p_0$ corresponding to the due north direction, as shown in Figure 5b.

As shown in Figure 4b, we calculate the distance $d$ between Point $O$ and Point $S$ by using Equation (23):

$$d = \sqrt{(x_a - x_0)^2 + (y_a - y_0)^2 + (z_a - z_0)^2}$$

(23)

In addition, we calculate the vertical angle component $\alpha$ by using Equation (24):

$$\alpha = \arctan \left(\frac{z_a - z_0}{d}\right)$$

(24)

We calculate the horizontal angle component $\beta$, as expressed in Equation (25):

$$\beta = \frac{y_a - y_0}{\cos(\alpha) \cdot d}$$

(25)

and then calculate the value $p$ corresponding to Point $s$, as shown in Equation (26):

$$p = p_0 + \beta$$

(26)

where $p_0$ is the azimuth corresponding to the due north direction.

4 | EXPERIMENT AND RESULT

4.1 | Research area and equipment

In this research experiment, a Hikvision DS-2DE7220YD-A spherical camera is taken as the PTZ camera (Figure 6d), which has in situ sensors, featuring a high-performance optical sensor and real-time crisp image output at 1920*1080@30fps. It is characterized by a magnification range of 20 times opticwise and 16 times digitwise, as well as a horizontal field of view angle of 2~57° and a vertical field of view angle of 1~32°. The network-controllable camera can rotate 360° continuously in the horizontal direction and —15~90° in the vertical direction. The equipment is mounted on a communication tower (118.9763, 31.9664,52) at Chunhua Street, Nanjing, China (Figure 6a) and the PTZ camera on the top of the communication tower with a height of 35 m in the center of the experimental area (Figure 6d). The total experimental area is about 4 km², located in the middle east of Nanjing, with geographical coordinates between 31°56’~31°59’N and 118°56’~118°59’E. The experimental data contain high-resolution remote sensing images with 0.05 m spatial resolution (Figure 6b), and the elevation data with 0.5 m spatial resolution from an aerial survey (Figure 6c). From the profile lines in three directions to the insight of the experimental area’s topographic change (Figure 6c), we can find that the experimental area is mainly in a hilly location with an elevation range of 30 to 70m, and the terrain is high in the north and low in the south. The mountains are located in the northwest, reaching a maximum elevation of 200m. All tests were performed on a standard PC with an Intel Core I7 6700 CPU running at 3.4 GHz and an NVIDIA GeForce GTX 1080 GPU with 16 GB of RAM.
4.2 | Experiment results

4.2.1 | Picking world-image pairs

Picking world-image pairs is the fundamental step in building the WLTOG algorithm. We constructed 64 world-image pairs with relevant properties within a 945 m radius around the communication tower (Figure 7). Forty-six pairs of
world images were assigned as reference points, and 20 pairs of world images served as verification points. The shadow portion of Figure 7 is rendered invisible due to the tower’s occlusion. First, adjust the camera’s PTZ posture to ensure the optical axis is aligned with observed ground objects. Secondly, locate the matched object on the remote

FIGURE 7  The spatial locations of reference points and verification points.
sensing image with high spatial resolution, and then record the camera’s $p$ and $t$ values as well as the geographical coordinates.

### 4.2.2 Nonlinear fitting of the bidirectional mapping model

By inputting values of 46 world-image pairs into Equations (19) and (22), we established the bidirectional mapping model and gained the overdetermined equation’s fitting parameters of $X_1$ and $X_2$ (Table 1 before grouping). Simultaneously, to address the issue that the high-order overdetermined equation mapping model does not account for the rotational symmetry of value $p$ during the world-to-image back-projection process, we adopted the world-image pairs grouping strategy (Figure 6 division line) and obtained two groups of overdetermined equations (Table 2 after grouping), reducing the error propagation associated with multidirectional geolocation.

### 4.2.3 Results analysis for the strategy of grouping

This section will quantify the prediction error of $p$ ($p_{\text{error}}$) induced by using the grouping strategy during the world-to-image back projection process, calculate the prediction $p$ for verification points and the absolute difference between the predicted and true $p$ to obtain $p_{\text{error}}$ and create a correlation line chart between $p_{\text{error}}$ and the true value $p$ (Figure 8), thereby assessing the effect of grouping graphically. As can be seen in Figure 8, before grouping, the $p_{\text{error}}$ fluctuates significantly with the increase of the true $p$. However, after grouping, the distribution of $p_{\text{error}}$ is stable, and the value swings slightly with the increase of the true $p$. We used Root Mean Square Error (RMSE) (Equation (27)) to evaluate the degree of dispersion across $p_{\text{error}}$ groups quantitatively before and after grouping. The RMSE of the $p_{\text{error}}$ is 21.489 before and 3.357 after grouping, respectively, demonstrating that the strategy for grouping is available to reduce $p_{\text{error}}$ successfully.

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2}$$ (27)

![Figure 8](image-url) $p_{\text{error}}$ error diagram before and after grouping.
Comparison of geolocation algorithms

To illustrate the efficiency of our method on large-range, omnidirectional, and less world-image pairs, the WLTOG algorithm was compared with the SPG algorithm and the random forest geolocation (RFG) algorithm (Table 2). The SPG algorithm widely employed in forest fire prevention serves as a typical example of video geolocation. Through

FIGURE 9 Error distribution of prediction results: (a) $p_{\text{error}}$ distribution of world-to-image back-projection; (b) $d_{\text{error}}$ distribution of image-to-world projection. In (a), the points are numbered from left to right as follows: 12, 15, 6, 14, 13, 5, 10, 9, 3, 4, 11, 8, 18, 17, 20, 16, 19, 7, 2, 1, and I, II in (a) correspond to Figure 7. In (b), the points are numbered from left to right: 14, 6, 12, 15, 10, 2, 13, 7, 1, 8, 5, 9, 4, 3, 11, 16, 17, 18, 19, 20.

4.2.4 Comparison of geolocation algorithms

To illustrate the efficiency of our method on large-range, omnidirectional, and less world-image pairs, the WLTOG algorithm was compared with the SPG algorithm and the random forest geolocation (RFG) algorithm (Table 2). The SPG algorithm widely employed in forest fire prevention serves as a typical example of video geolocation. Through
Before grouping

| $X_1$ | $X_2$ |
|-------|-------|
| 0     | 0     |
| 1.00e-06 | 1.00e-06 |
| 9.60e-04 | 5.87e-04 |
| 5.00e-06 | 6.00e-06 |
| 0     | 0     |
| 0     | 0     |
| 1.00e-06 | 1.00e-06 |
| 5.00e-06 | 5.00e-06 |
| 3.00e-05 | 1.70e-05 |
| 0     | 0     |
| 0     | 0     |
| 1.00e-06 | -1.00e-06 |
| -2.60e-05 | 2.33e-04 |
| 4.00e-06 | 4.10e-05 |
| 5.58e-04 | 1.98e-04 |
| 4.00e-06 | 2.80e-05 |
| 1.21e-04 | 7.30e-05 |
| 3.89e-04 | 3.31e-04 |
| 4.00e-06 | 1.65e-03 |
| 5.00e-02 | 5.85e-03 |
| 1.52e-01 | 1.55e-02 |
| 1.52e-02 | 6.74e-02 |
| 3.00e-01 | 4.04e-01 |
| 4.66e-01 | 6.30e-02 |
| 2.23 | 2.04 |
| 7.86 | 1.50 |

**TABLE 1**  A nonlinear fitting result for the bidirectional mapping model
|     | $X_1$          | $X_2$          |     |     |
|-----|---------------|---------------|-----|-----|
|     | After grouping |               |     |     |
|     | 0             | 0             | 0   | 0   |
|     | $1.00e-06$    | 0             | 0   | $3.00e-06$ |
|     | 0             | 0             | 0   | $-7.45e-04$ |
|     | $1.00e-06$    | 0             | 0   | $2.00e-06$ |
|     | 0             | $-1.00e-06$   | 0   | $4.00e-06$ |
|     | $-2.20e-05$  | $-2.18e-04$   | $2.00e-06$ |
|     | $-5.56e-04$  | $-2.19e-04$   | $4.20e-05$ |
|     | $1.38e-04$   | $4.90e-05$    | $6.00e-06$ |
|     | $-3.71e-04$  | $4.34e-04$    | $2.70e-05$ |
|     | $4.90e-03$   | $4.90e-02$    | $-1.56e-03$ |
|     | $1.53e-01$   | $-1.43e-02$   | $-1.29e-02$ |
|     | $1.93e-02$   | $-7.07e-02$   | $-5.22e-03$ |
|     | $-1.16e$     | $-2.97$       | $4.06e-01$ |
|     | $-9.84e$     | $7.78$        | $1.50$ |

**TABLE 1 (Continued)**
| Point number | p   | t   | x   | y   | z   | Target distance | \( p_{error} \) | \( t_{error} \) | \( d_{error} \) | \( p_{error} \) | \( t_{error} \) | \( d_{error} \) | \( p_{error} \) | \( t_{error} \) | \( d_{error} \) | \( p_{error} \) | \( t_{error} \) | \( d_{error} \) |
|--------------|-----|-----|-----|-----|-----|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| 1            | 308.5 | 15  | 118.9639 | 31.9613 | 51.57 | 228.804 | 6.498 | 0.515 | 21.019 | 4.366 | 0.256 | 10.715 | 12.845 | 1.755 | 13.734 | 3.111 | 3.125 | 18.554 |
| 2            | 293.1 | 18.7 | 118.9645 | 31.9615 | 52.11 | 184.863 | 7.391 | 0.804 | 19.492 | 10.186 | 0.213 | 12.979 | 3.311 | 3.125 | 18.554 | 2.845 | 0.512 | 37.954 |
| 3            | 175.5 | 9.1  | 118.9671 | 31.9636 | 50.05 | 279.384 | 1.335 | 0.23  | 11.598 | 0.234 | 0.033 | 5.345 | 2.845 | 0.512 | 37.954 | 4.366 | 0.256 | 10.715 |
| 4            | 200.6 | 10.8 | 118.9671 | 31.9627 | 48.83 | 271.213 | 1.039 | 0.999 | 3.126 | 0.147 | 0.065 | 3.378 | 2.985 | 0.585 | 54.606 | 1.976 | 0.166 | 11.902 |
| 5            | 70.9  | 6.8  | 118.9635 | 31.9645 | 62.01 | 247.911 | 0.305 | 0.269 | 19.865 | 2.035 | 2.321 | 93.483 | 4.751 | 0.911 | 18.977 | 3.058 | 0.241 | 21.957 |
| 6            | 59.7  | 19.2 | 118.9639 | 31.9636 | 55.05 | 131.941 | 4.144 | 0.361 | 6.886 | 0.157 | 1.929 | 34.474 | 2.115 | 0.312 | 21.957 | 2.115 | 0.312 | 21.957 |
| 7            | 275.5 | 16.9 | 118.965  | 31.9613 | 52.01 | 208.183 | 4.604 | 0.158 | 10.863 | 8.649 | 0.166 | 11.092 | 3.355 | 2.415 | 44.672 | 1.071 | 0.275 | 69.108 |
| 8            | 251   | 14.1 | 118.9659 | 31.9614 | 49.1  | 240.114 | 3.023 | 0.491 | 6.74  | 0.033 | 0.498 | 21.399 | 1.071 | 0.275 | 69.108 | 2.676 | 0.226 | 21.365 |
| 9            | 135.6 | 8    | 118.966  | 31.9646 | 52.84 | 259.751 | 1.867 | 1.38  | 9.819 | 2.633 | 1.988 | 33.929 | 1.271 | 0.241 | 13.447 | 0.025 | 0.612 | 12.176 |
| 10           | 101.4 | 11.7 | 118.9647 | 31.9642 | 60.79 | 172.843 | 1.587 | 0.403 | 16.903 | 2.36  | 2.756 | 131.692 | 5.321 | 0.461 | 10.854 | 2.676 | 0.226 | 21.365 |
| 11           | 218.9 | 9.6  | 118.9672 | 31.9618 | 49.13 | 315.028 | 0.708 | 0.378 | 8.756 | 0.778 | 0.028 | 11.984 | 1.223 | 0.975 | 56.957 | 4.191 | 1.279 | 42.858 |
| 12           | 36.2  | 19   | 118.9635 | 31.9633 | 54.14 | 142.925 | 1.867 | 1.38  | 9.819 | 2.633 | 1.988 | 33.929 | 1.271 | 0.241 | 13.447 | 0.025 | 0.612 | 12.176 |
| 13           | 69.9  | 10   | 118.9637 | 31.9642 | 58.5  | 207.253 | 2.676 | 2.147 | 25.749 | 1.984 | 1.66  | 69.056 | 0.025 | 0.612 | 12.176 | 6.245 | 0.201 | 22.321 |
| 14           | 69.3  | 19.7 | 118.9641 | 31.9637 | 56.99 | 122.943 | 0.911 | 3.541 | 34.118 | 4.191 | 1.279 | 42.858 | 6.245 | 0.201 | 22.321 | 0.791 | 1.05  | 12.437 |
| 15           | 52.9  | 14.5 | 118.9635 | 31.9637 | 55.75 | 170.924 | 0.773 | 1.051 | 15.263 | 0.077 | 0.907 | 55.462 | 0.791 | 1.05  | 12.437 | 2.676 | 0.226 | 21.365 |
| 16           | 265.1 | 353.8 | 118.9665 | 31.9596 | 55.32 | 405.833 | 2.113 | 0.226 | 21.365 | 1.954 | 0.047 | 28.533 | 2.209 | 0.344 | 23.337 | 2.752 | 1.129 | 21.568 |
| 17           | 260.2 | 354.3 | 118.9662 | 31.9582 | 55.12 | 585.876 | 0.346 | 0.879 | 13.221 | 2.752 | 1.129 | 21.568 | 2.743 | 2.475 | 33.689 | 5.532 | 2.094 | 157.43 |
| 18           | 258.4 | 354.9 | 118.9659 | 31.9570 | 54.82 | 639.585 | 2.335 | 1.083 | 17.771 | 5.532 | 2.094 | 157.43 | 3.436 | 3.871 | 58.339 | 4.323 | 1.976 | 21.033 |
| 19           | 270.7 | 356.1 | 118.9672 | 31.9561 | 55.73 | 819.235 | 4.398 | 2.201 | 18.576 | 4.323 | 1.976 | 21.033 | 5.25  | 1.773 | 37.337 | (Continues) |
TABLE 2 (Continued)

| Point number | p   | t   | x     | y    | z     | Target distance | WLTOG World-to-image back-projection | Image-to-world projection | SPG World-to-image back-projection | Image-to-world projection | RFG World-to-image back-projection | Image-to-world projection |
|--------------|-----|-----|-------|------|-------|-----------------|---------------------------------------|----------------------------|-------------------------------|-------------------------------|-------------------------------|---------------------------------|
| 20           | 261.8 | 356.4 | 118.9662 | 31.9552 | 55.01 | 945.933         | 6.187 | 2.297 | 23.35       | 8.037 | 2.904 | 35.533       | 7.521 | 2.286 | 57.87  |

Note: The number in the table corresponds to the number in Figure 7 for the verification point. Where \( p \) and \( t \) denote the camera’s decimal degrees of freedom, unit dimension (°); \( x \) and \( y \) denote the target’s decimal latitude and longitude in WGS84, unit dimension (°); \( z \) denotes the target’s altitude in WGS84, unit dimension (m); target distance denotes the Euclidean distance between the target and communication tower, unit dimension (m); image-to-world projection denotes the process of converting the target’s image pixel into the corresponding geospatial coordinate. The error associated with image-to-world projection is denoted by the distance \( d_{\text{error}} \) between the predicted and actual geospatial coordinates, unit dimension (m). \( d_{\text{error}} \) is the absolute distance error between predicting \( d \) and true \( d \); World-to-image back-projection refers to the process of converting the geospatial coordinates of the target into the corresponding camera posture when the target is in the center of the camera image. The error associated with world-to-image back-projection is defined as the difference between the target’s actual camera posture value and the anticipated camera posture value denoted by the variables \( p_{\text{error}} \) and \( t_{\text{error}} \), unit dimension (°). \( p_{\text{error}} \) is the absolute azimuth error between predicting \( p \) and true \( p \), and \( t_{\text{error}} \) is the absolute tilt error between predicting \( t \) and true \( t \).

In Table 2, the \( d \) errors of verification points (ID 5, ID 10, ID 13 and ID 18) solved by the SPG method deviate significantly. Combined with the field investigation and analysis, we found that these points were located in the area where the terrain fluctuates greatly or were covered by vegetation and buildings, and the coordinates of surrounding reference points are quite different from the actual value. When SPG is used to locate, the value of point \( P \) and \( t \) is first verified according to the \( t \) offset value of reference points interpolation is used to correct, 2-order interpolation method, selected as interpolation method, amplifies the error further. It can be seen that large topographic relief or plants and building shelter will be able to generate large fluctuations in the geolocation effect of verification points.
rigorous intervisibility analysis of DEM, the SPG method is used to identify the relationship between the visual line of the target and the camera parameters, thereby solving the target’s geographical coordinates. The RFG algorithm is a nonlinear fitting algorithm based on the random forest fit. We could objectively estimate each decision tree's internal generation error by combining numerous decision trees. It can be seen from Table 2 that the bidirectional geolocation error of the WLTOG algorithm ($p_{error} = 3.357^\circ, t_{error} = 1.323^\circ, d_{error} = 17.305$ m) is lower than that of the SPG algorithm ($p_{error} = 3.997^\circ, t_{error} = 1.324^\circ, d_{error} = 53.009$ m) and RFG algorithm ($p_{error} = 4.963^\circ, t_{error} = 1.356^\circ, d_{error} = 34.243$ m).

Figure 9 demonstrates the WLTOG algorithm's superiority in fitting both the world-to-image back-projection and image-to-world projection. Figure 9a illustrates the correlation line charts between $p_{error}$ and the true $p$ for three algorithms used in the world-to-image back-projection process. In all directions, the $p_{error}$ of the WLTOG algorithm is lower than that of the SPG and RFG algorithms (Figure 9a). Due to the limited number of reference points, the $p_{error}$ of the RFG algorithm is much higher than that of the algorithms of WLTOG and SPG, indicating that the RFG algorithm is better suited to the prediction scenario, including multi-world-image pairs data. Figure 9b illustrates the correlation line charts between prediction distance error ($d_{error}$) and actual distance for the three algorithms used in the image-to-world projection process. As shown in Figure 9b, as the distance between verification points and the camera increases, the RFG algorithm's error progressively increases, and the SPG algorithm's error distribution becomes unstable. In contrast, the WLTOG algorithm’s error distribution remains stable, with a slightly downward tendency. This phenomenon occurs because most reference and verification points are dispersed between 200 and 300 m, and there are few reference points beneath the communication tower.

It is possible to add reference points in the visible long-distance region to improve remote geolocation accuracy. Therefore, we rely on experiments to verify the usefulness of this approach for long-distance localization. Selecting

![Figure 10](image-url)

**FIGURE 10** (a1) A typical target example (an engineering truck) in the PTZ camera image; (a2) the world coordinates result for the engineering truck in Figure 10a1; (b1) the electronic fence around a pond with the turning points indicated on the PTZ camera image; (b2) the world coordinates results for the turning points of the electronic fence and in Figure 10b2. The target box and electronic fence are indicated by the green border in (a1) and (b1), respectively, while the world coordinates determined by the WLTOG algorithm are given by the yellow crosses in (a2) and (b2).
a visible long-distance direction, as illustrated in Figure 7 (ID 18, 19, and 20), we pick reference points in the local region within 500–1000 m. It is unnecessary to investigate the back-projection in this experiment because reference points are added for visible camera alignment. After updating the image-to-world projection model, we calculate the geolocation error of the specified verification points. The verification points (ID 18, 19, and 20) are 679.6, 819.2, and 945.9 m away from the tower, respectively, with geolocation errors of 12.2, 18.6, and 23.3 m (Figure 9b). On the other hand, it can be seen from the experimental results that picking reference points in the visible long-distance region is available to increase the accuracy of long-distance geolocation results. In addition, the localization range of the algorithm can be extended to 945 m by the addition of reference points.

FIGURE 11 The target region on the map is denoted by the green boundaries, while the WLTOG algorithm's estimated image pixel results are denoted by the yellow crosses.
4.3 | The application experiments

4.3.1 | Image-to-world projection

Image-to-world projection refers to the transformation from the target's pixel coordinates in the image into geographical coordinates to augment the target with 3D geological information. It is primarily used in two applications, i.e., target detection geolocation and electronic fence geolocation. Figures 10a1, a2 illustrate the geolocation of target detection, where an engineering truck can be seen through the monocular PTZ camera screen (Figure 10a1). After computation, the vehicle's geographical coordinates were determined (Figure 10a2). As shown in Figure 10b1, the pond's electronic fence (made of seven drawing points) was drawn on the camera image, with the drawing point positioned at the pond's turning point. The prediction results for the electronic fence's drawing points are given in Figure 10b2. What's more, the detailed experiment demonstration video of the WLTOG algorithm is presented at https://doi.org/10.6084/m9.figshare.19513693.v1. It is worth noting that when the distance between the drawing points and the camera increases or the density of the reference points decreases, the predicted geolocation of some points may result in a slight deviation.

4.3.2 | World-to-image back-projection

World-to-image back-projection refers to converting the target's geographical coordinates to the posture values of the camera. It is widely used in practice, including target tracking and fixed scene monitoring. In the world-to-image back-projection experiment, 16 points were randomly selected on the map, the target's predicted camera posture values were calculated using the bidirectional mapping model, and the PTZ camera was controlled to complete the rotation (Figure 11). The expanded view of the 16 points displays the prediction results for each positioning point following the computation of the world-to-image back-projection model. The yellow cross represents the targets' actual camera posture. Based on the experimental findings, the prediction results for the 16 points were found in the image, confirming that the model could precisely convert map coordinates to image pixel coordinates.

5 | DISCUSSION AND CONCLUSION

The world is growing increasingly complex and dynamic (Yuan, 2017). Meanwhile, video surveillance from PTZ cameras can be regarded as the most effective way to record dynamic geographical scenes. So, a new creative GIS research direction is to study on large-range and omnidirectional geolocation for monocular PTZ cameras. The WLTOG algorithm is proposed in this article to perform bidirectional calculations between the target pixel of a single image and the world coordinates. The experimental findings demonstrate that the WLTOG algorithm satisfies the omnidirectional geolocation requirement within 945 m and examines the process to improve the geolocation error distribution. The WLTOG algorithm also applies to large-range and omnidirectional monocular PTZ camera geolocation in the wild environment. It overcomes the limitations of perspective projection model geolocation and omnidirectional home matrix transformation geolocation, which does not require accurate calibration of interior and exterior orientation elements. It is applicable to a wide range of video surveillance and geographic application situations, such as urban security, forest fire-prevention, and video GIS.

The existing studies focus on the binary homography relationship between image pixels and world coordinates. However, the monocular PTZ camera has the characteristic of multi-angle rotation and multi-focal zoom, making the geolocation of the PTZ camera develop into a triadic relation among image pixels, PTZ postures, and world coordinates, thereby satisfying the application requirements of omnidirectional geolocation. This is also the ingenious design of our method in this article, which was not fully paid attention to in the previous studies.
In the future, our focus will be on the improvement of the positioning accuracy while considering center eccentricity. Due to the instability of the PTZ camera's physical structure caused by external impact-induced deformation, long-term mechanical wear, etc, there is a problem of apparent center eccentricity during multiple focusing of the PTZ camera, posing significant challenges to the bidirectional mapping models with multiple focal lengths. To further improve geolocation accuracy, it is necessary to improve the bidirectional mapping model to adapt to the eccentricity of the PTZ camera.

ACKNOWLEDGMENTS
The authors would like to thank the anonymous referees and editor for their valuable comments, which significantly improved this article.

FUNDING INFORMATION
This research was supported by National Natural Science Foundation of China (Grant Nos. 42201504, 41471318), Open Foundation of Key Lab of Virtual Geographic Environment of Ministry of Education (Grant No. 2021VGE02), Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications (Grant No. NY221143).

CONFLICT OF INTEREST
No potential conflict of interest was reported by the authors.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are openly available in data and code supporting at https://doi.org/10.6084/m9.figshare.19513693.v1.

ENDNOTES
1 This term, refers to the reference point of the image and that of the geographical scene, and has been translated as 'world-image pairs' (Arroyo et al., 2020).
2 This term, containing pan angles, tilt angles and the zooming level, has been transformed as PTZ parameters (Lisanti et al., 2016).
3 Since the ground objects positioned in the optical axis are not subject to optical distortion induced by focal length shift. To simplify calculation, we do not discuss the camera focal length and are used to estimate the values of the intrinsic parameters for different pan and tilt angles at the initial zooming level.

REFERENCES
Anagnostopoulos, V., Havlena, M., Kiefer, P., Giannopoulos, I., Schindler, K., & Raubal, M. (2017). Gaze-informed location-based services. International Journal of Geographical Information Science, 31(9), 1770–1797. https://doi.org/10.1080/13658816.2017.1334896
Arroyo, S. I., Bussi, U., Safar, F., & Oliva, D. (2020). A monocular wide-field vision system for geolocation with uncertainties in urban scenes. Engineering Research Express, 2(2), 025041. https://doi.org/10.1088/2631-8695/ab9b36
Brickwedde, F., Abraham, S., & Mester, R. (2019). Mono-sf: Multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes. IEEE/CVF International Conference on Computer Vision, Long Beach, CA (pp. 2780–2790). IEEE. https://doi.org/10.1109/ICCV.2019.00287
Chen, C., Wang, B., Lu, C. X., Trigoni, N., & Markham, A. (2020). A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence. arXiv preprint arXiv:12567. https://doi.org/10.48550/arXiv.2006.12567
Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., & Urtasun, R. (2016). Monocular 3D object detection for autonomous driving. IEEE Conference on Computer Vision & Pattern Recognition, Las Vegas, NV (pp. 2147–2156). IEEE. https://doi.org/10.1109/CVPR.2016.236
Shao, Z., Li, C., Li, D., Altan, O., Zhang, L., & Ding, L. (2020). An accurate matching method for projecting vector data into surveillance video to monitor and protect cultivated land. ISPRS International Journal of Geo-Information, 9(7), 448. https://doi.org/10.3390/ijgi9070448

Simonelli, A., Bulo, S. R., Porzi, L., López-Antequera, M., & Kontschieder, P. (2019). Disentangling monocular 3D object detection. IEEE/CVF International Conference on Computer Vision, Long Beach, CA (pp. 1991–1999). IEEE. https://doi.org/10.1109/TPAMI.2020.3025077

Sivaraman, S., & Trivedi, M. M. (2013). Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. IEEE Transactions on Intelligent Transportation Systems, 14(4), 1773–1795. https://doi.org/10.1109/TITS.2013.2266661

Song, C., Song, J., & Huang, Q. (2020). Hybridpose: 6d object pose estimation under hybrid representations. IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 431–440). IEEE. https://doi.org/10.1109/CVPR42600.2020.00051

Tian, Y., Sun, A., Luo, N., & Gao, Y. (2020). Aerial image mosaicking based on the 6-DoF imaging model. International Journal of Remote Sensing, 41(1), 74–89. https://doi.org/10.1080/01431161.2019.1657602

Wang, M. (2011). Research on geometric measurement for man-made features based on single images. [PhD Dissertation]. Nanjing Normal University.

Wang, W., Gee, T., Price, J., & Qi, H. (2015). Real time multi-vehicle tracking and counting at intersections from a fisheye camera. 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI (pp. 17–24). IEEE. https://doi.org/10.1109/WACV.2015.10

Xie, Y., Wang, M., Liu, X., Wang, Z., Mao, B., Wang, F., & Wang, X. (2021). Spatiotemporal retrieval of dynamic video object trajectories in geographical scenes. Transactions in GIS, 25(1), 450–467. https://doi.org/10.1111/TGIS.12696

Xie, Y., Wang, M., Liu, X., & Wu, Y. (2017). Integration of GIS and moving objects in surveillance video. ISPRS International Journal of Geo-Information, 6(4), 94. https://doi.org/10.3390/ijgi6040094

Yuan, M. (2017). 30 years of UGIS: The changing landscape of geographical information science and the road ahead. International Journal of Geographical Information Science, 31(3), 425–434. https://doi.org/10.1080/13658816.2016.1236928

Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., & Liu, G. (2020). A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 166, 183–200. https://doi.org/10.1016/j.isprsjprs.2020.06.003

Zhang, X., Liu, X., & Song, H. (2013). Video surveillance GIS: A novel application. 21st International Conference on Geoinformatics, Henan, China (pp. 1–4). IEEE. https://doi.org/10.1109/Geoinformatics.2013.6626079

Zhang, X., Shi, X., Luo, X., Sun, Y., & Zhou, Y. (2021). Real-time web map construction based on multiple cameras and GIS. ISPRS International Journal of Geo-Information, 10(12), 803. https://doi.org/10.3390/ijgi10120803

Zhou, X., Wu, C., Lv, D., Qi, H., & Han, P. (2021). Object detection and spatial location method for monocular camera based on 3D virtual geographical scene. 6th International IEEE Conference on Cloud Computing and Big Data Analytics, Chengdu, China (pp. 488–495). IEEE. https://doi.org/10.1109/ICCCBDA51879.2021.9442542

Zhu, R., Zhu, R., Yang, X., Hold-Geoffroy, Y., Perazzi, F., Eisenmann, J., Sunkavalli, K., & Chandraker, M. (2020). Single view metrology in the wild. 16th European Computer Vision Conference, Glasgow, UK (pp. 316–333). https://doi.org/10.1007/978-3-030-58598-3

How to cite this article: Zhou, X., Yuan, L., Tang, H., Huang, Y., Feng, N., & Wu, C. (2022). WLTOG: An optimization approach for wild large-range target omnidirectional geolocation based on monocular PTZ camera. Transactions in GIS, 26, 2975–3002. https://doi.org/10.1111/tgis.12991