From High-resolution Remote Sensing to Tower-mounted Video Surveillance System: Dual Feature Image Matching-Based Precise Positioning Method in Complex Suburban Areas

Yunong Chen[1], Zhiqing Tang[2], Zhijun Wen[2], Boshen Chang[1], Zhizheng Zhang[1], Pengcheng Wei[1], Deren Li[1,*]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China – yunong_chen@whu.edu.cn
[2] The Second Surveying and Mapping Institute of Hunan Province, Hunan 410000, China
* Corresponding author: drli@whu.edu.cn (Deren Li)

Keywords: Complex Suburban Areas, Remote Sensing, Tower-mounted Video Surveillance, Precision Positioning, Coordinate Matching.

Abstract

With the continuous advancement of urbanization, the demand for comprehensive monitoring of diverse suburban areas is increasing, which simultaneously raises the need for high-precision positioning technologies especially for the automatic positioning for video surveillance imagery. The current challenges in high-precision positioning for surveillance data include: (1) The difficulty of matching wide-angle images; (2) The significant scale differences between Tower-mounted Video Surveillance (TVS) and Remote Sensing (RS) ortho-images complicate multi-scale automatic focusing; (3) The positioning accuracy decreases due to the physical factors in suburban scene areas. To address these issues, this study proposes a Dual-Feature Image Matching-based Method for High-Resolution RS and TVS Collaboration method for precise positioning in complex suburban areas, referred to as DFMC. Through the processes of geo-referencing, dual-feature image matching, homography transformation and geographic coordinate computation, the proposed DFMC overcomes the challenges of high-precision matching between wide-angle video images and RS images, facilitating the projection of any point coordinates from the pixel coordinate system of TVS frames to the projected Cartesian coordinate system of RS ortho-images. Experiments are conducted using over 28,000 video frames of tower-mounted surveillance system across 210,000 km^2 in Hunan Province of China. The results indicate the proposed DFMC achieves a positioning accuracy error of less than 1.5 meters in flat areas and less than 3 meters in complex suburban areas. Therefore, DFMC enables rapid monitoring and positioning in complex suburban areas, providing valuable informational support for relevant authorities.

1. Introduction

Suburban areas are typically regarded as part of the city due to their unique geographical locations and important functions in urban development planning. In recent years, with the acceleration of the urbanization process, research on land use, management, security assurance, and infrastructure maintenance in suburban areas has gradually increased (Wu et al. 2023, Yang et al. 2013). However, traditional positioning technologies struggle to meet the comprehensive demands of the geomorphologically diverse suburban areas. In particular, Remote Sensing (RS) imagery is close to vertical photography, the gained orthoimage has high positioning accuracy and can be easily matched with map coordinates, while cannot be captured in real time. The real time Tower-mounted Video Surveillance (TVS) data obtaining is widely used in suburban areas, while cannot be directly matched with the map coordinates, thus the positioning accuracy is low. Using the advantages of RS data to get high-resolution image and accurate positioning, combining the advantage of real-time TVS data to monitor the change and potential risk can effectively enhance positioning accuracy, making it more suitable for application in functionally complex and geomorphologically varied suburban areas. The application of the above method will not only aid in tackling the management and development challenges of suburban areas but will also provide critical intelligent management support in fields such as resource monitoring, smart city construction, urban intelligent transportation, and emergency response, thereby laying a solid foundation for the long-term development of the nation (Haight et al. 2023).

The key issue of TVS-RS collaboration lies in how to integrate and process multi-source data to achieve effective data management. Multi-source data primarily includes remote sensing images data from satellite or Unmanned Aerial Vehicle UAV (such as latitude and longitude coordinates) and ground target information (such as images and videos provide pixel coordinates); data processing refers to the rapid and high-precision transformation of these two types of data. RS images are acquired through non-contact methods and have become effective tools for monitoring urban expansion and early warning of geological disasters (Casagli et al. 2023). However, RS method has several drawbacks such as slow update frequency by using satellite and small coverage area and require trained staffs to operate by using UVA. In contrast, ground video monitoring technology, particularly TVS systems, may be affected by weather conditions and have a limited coverage area; however, they offer advantages such as resistance to natural factors like cloud cover, real-time video capture, and the ability to achieve remote control. Therefore, the combination of these data and technologies effectively addresses the limitations of remote sensing technology regarding timeliness and the shortcomings of TVS concerning the coverage area. Furthermore, ground monitoring not only provides intuitive visual information that enhances the validation capability of precise positioning system data but also accelerates emergency response in critical situations, improving the efficiency of rescue or law enforcement operations and providing important evidence for post-event analysis. Therefore, effectively combining remote sensing with ground video monitoring can fully leverage their respective strengths and significantly enhance the efficiency of suburban management.

Recently, researchers have attempted to integrate TVS imagery with RS images in a nested manner, applying this approach to specific scenario-based tasks. For instance, Milosavljević et al. (Milosavljević et al. 2016) proposed a method that integrates

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-173-2025 | © Author(s) 2025. CC BY 4.0 License.

Augmented Reality (AR) technology with three-dimensional (3D) Geographic Information Systems (GIS) into video surveillance systems. Shao et al. (Shao, 2020) developed a precise matching method for vector data and surveillance videos based on dense matching techniques. This method enables the identification of a sufficient number of control points from two-dimensional (2D) GIS data and selected reference video images, subsequently achieving alignment of remote sensing imagery with PTZ (Pan-Tilt-Zoom) video image frames through automated feature matching techniques. However, georeferencing is required at each stage of the process and imposes high demands on the video scene being utilized. To connect surveillance videos with a virtual 3D GIS environment, Milosavljević et al. (Milosavljević et al. 2017) proposed a parameter estimation method that jointly matches video feature points with 3D coordinates. This method relies on high-resolution Digital Orthophoto Maps (DOM) and Digital Elevation Models (DEM), employing the Levenberg-Marquardt iterative optimization algorithm to ultimately determine the most suitable camera position and orientation information. However, due to variations in camera focal length and angle, this method struggles to obtain accurate and reliable physical information of the imagery, thus it is rarely used in practical applications.

To ensure national security and support global sustainable development strategies, many provinces and cities in China have established real-time land monitoring systems. For instance, Jiangsu Province's ‘Guarding the Land with Insightful Eyes Surveillance system’ system and Ningbo City's arable land protection supervision system. The former has deployed over 18,000 video surveillance points in key areas of Jiangsu Province, allowing for the overlay of real-time video monitoring images with fundamental databases such as cadastral survey data, overall land use planning, and mineral resource distribution. This system can automatically compare and analyse discrepancies to issue early warnings and monitor suspected illegal activities (Cao et al. 2019, Ouyang et al. 2022). The latter leverages Ningbo's tower video monitoring system to overlay existing arable land map layer data and create electronic fences for agricultural land. This system can monitor illegal occupation of arable land within a 15-kilometer radius around tower base stations in real time, supporting the automatic identification of 33 target types, including red brick piles, excavators, and prefabricated houses, with an average accuracy of 90% (Wang et al. 2022, Li et al. 2022). Additionally, Jiaxing City, Zhejiang Province, has recently established the ’National Land Sky Eye Surveillance System’ (Zhang et al. 2020). It achieves real-time monitoring of land use and illegal construction within a 3-5 square kilometre area around communication towers through 101 communication tower stations and video monitoring equipment, with platform operational efficiency exceeding 99%.

Although these existing TVS-RS collaborative systems have initially realized monitoring and analysis functions in the area, they still exhibit the following deficiencies: 1) The difficulty of matching wide-angle images: Conventional global affine transformations lead to geometric distortions that may extend to kilometres, failing to meet the necessary supervision standards; 2) Inaccurate image feature matching: There is a discrepancy in feature point matching between TVS images and RS imagery, preventing precise image alignment; 3) Weak anti-interference capability: High-altitude cameras on towers are easily affected by external geographic factors (such as complex suburban areas) and variations in focal length, which in turn impact high-precision positioning. Consequently, the current TVS-RS collaborative systems struggle to meet the requirements posed by the diverse geomorphology of suburban areas, particularly in hilly terrains, making efficient management and monitoring of these areas challenging.

In this work, we propose a dual-feature image matching method for the collaboration of high-resolution RS and TVS for precise positioning in complex suburban areas, termed DFMC. The proposed method first establishes a mapping relationship between the pixel coordinate system of TVS images and the projected Cartesian coordinate system of RS orthoimages to achieve georeferencing. Subsequently, DFMC performs feature detection and feature description of the TVS image and the aligned standard image, and achieves image matching at arbitrary angles by dual feature matching. This dual feature matching contains two parts, coarse matching establishes a pixel-by-pixel dense matching between the tower image and the standard image, and then refines the matching accuracy of both images at fine matching. Finally, DFMC utilizes homothetic transformation to project the coordinates of any point in the current TVS image frame onto the reference image coordinate system, achieving rapid coordinate conversion and positioning. Practical case studies demonstrate that the proposed DFMC method can achieve high-precision and rapid positioning in flat areas, as well as accurate positioning in complex suburban hilly terrains. The collaborative remote sensing precise positioning method introduced in this paper offers new perspectives and theoretical support for planning, management, environmental monitoring, and emergency response in suburban areas.

2. Methodology

In addressing the complexities of suburban areas, the proposed DFMC method can rapidly match a specific image frame from TVS with the corresponding RS imagery, map the image coordinates of various points, lines, and surfaces in the TVS image to the geographic coordinate system of the satellite orthoimage. As illustrated in Figure 1, the DFMC method consists of three main stages: georeferencing, dual feature image matching, homography transformation, and positional information conversion. Specifically, georeferencing establishes a mapping relationship between the TVS reference image and the satellite orthoimage by selecting corresponding feature points, thereby eliminating discrepancies caused by factors such as tilt angle, shooting height, and color texture, while creating a standardized image sample library. Subsequently, through feature detection and description of the heterogeneous images, DFMC performs dual feature matching. This matching process initially establishes dense pixel-wise matching at a coarse level, followed by refinement of feature matching at a detailed level. Finally, the homography transformation projects any point coordinates from the TVS image frame into the RS imagery coordinate system, facilitating rapid location conversion.

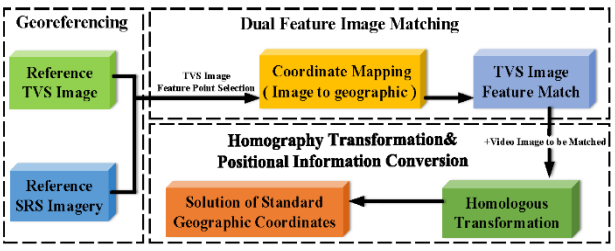

Figure 1. The Proposed Framework of DFMC.

2.1 Georeferencing

RS imagery employs a projected Cartesian coordinate system, while the TVS images utilize a pixel coordinate system. To

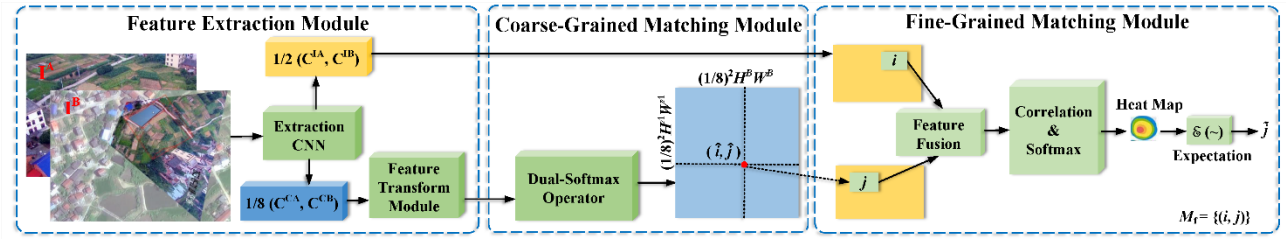

Figure 2. Dual Feature Image Matching Mechanism.

effectively display the results of surveillance videos within a GIS, it is essential to establish a mapping relationship between the pixel coordinate system used in TVS images and the projected Cartesian coordinate system of the RS orthoimages. Currently, commonly used georeferencing fall into two categories: homogenous transformation methods and Line of Sight (LOS) intersection methods with Digital Elevation Models (DEM) (Okolie and Smit, 2022). The former method relies on the accuracy, quantity, and spatial distribution of selected feature points and typically requires significant user interaction to ensure mapping accuracy. The latter method determines the geographic spatial coordinates of image pixels by calculating the intersection points between the viewpoint line and the DEM. However, this method is less frequently applied in practice due to the difficulty in obtaining camera parameters and DEM data. In both 2D and 3D GIS, the generation of the mapping matrix is the final outcome of this process. In 2D GIS, the mapping process is bidirectional, whereas in 3D GIS, it represents a unidirectional mapping from 3D GIS data to 2D surveillance videos. Conversely, 2D surveillance videos cannot be directly mapped in geographic space. Therefore, a direct conversion between GIS digital orthophoto map coordinates (X, Y) and surveillance video pixel coordinates (u, v) is not feasible.

To address this issue, this paper proposes a cross-mapping model between the 2D GIS image and TVS image under height constraints, as illustrated in the following equation:

$$\begin{bmatrix} d_x \\ d_y \\ d_z \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(-\theta_x) & -\sin(-\theta_x) \\ 0 & sin(-\theta_x) & cos(-\theta_x) \end{bmatrix} \begin{bmatrix} cos(-\theta_y) & 0 & sin(-\theta_y) \\ 0 & 1 & 0 \\ -sin(-\theta_y) & 0 & cos(-\theta_y) \end{bmatrix} \begin{bmatrix} cos(-\theta_z) & -sin(-\theta_z) & 0 \\ sin(-\theta_z) & cos(-\theta_z) & 0 \\ 0 & 0 & 1 \end{bmatrix} \left(\begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} - \begin{bmatrix} c_x \\ c_y \\ c_z \end{bmatrix} \right) \quad (1)$$

Here, $d_{x,y,z}$ represent the spatial position of a point after projection, $a_{x,y,z}$ signify the position in 3D to be projected, $c_{x,y,z}$ is the position of the camara, with $\theta_{x,y,z}$ as the rotation angles of the camera. Through Equation (1), we can align points in different coordinate systems between TVS images and RS imagery, achieving geographic co-registration. Building upon this, we perform feature point matching between the TVS images and the RS orthoimages to obtain a sufficient number of corresponding point samples, thereby constructing a standard image sample library for TVS images at specific angles, aided by RS imagery.

2.2 Dual Feature Image Matching Mechanism

To achieve omnidirectional positioning of the target location or area at arbitrary angles, we first need to perform feature detection and extraction on both the standard images of the TVS and the target TVS images to be matched. Subsequently, we apply feature matching algorithms to match these features, ultimately establishing a feature mapping relationship to facilitate rapid target matching. Feature matching techniques can extract distinctive feature points from a given pair of images, which typically exhibit invariance to rotation, scale, and illumination changes, enabling accurate detection under varying positions and angles. Existing feature matching methods are roughly divided into two categories based on whether a feature detector is used: feature detection-based matching methods and non-feature detection-based matching methods. The former is characterized by strong robustness and high interpretability, but it comes with higher computational complexity; the latter can leverage global information and is highly adaptive, making it suitable for complex scenes, but it demands substantial computational resources and has poorer interpretability.

Due to the lack of texture in the terrain of complex suburban areas at multiple focal lengths, existing methods struggle to extract key features. Building upon the work of Sun et al. (Sun et al. 2021), we propose a dual feature image matching mechanism that transitions from coarse to fine granularity. This is a non-feature detection-based matching method that first establishes dense pixel-wise matching between the target TVS images and the standard TVS images, followed by refinement of the matching accuracy. As illustrated in Figure 2, the overall dual feature image matching mechanism comprises a feature extraction module, a coarse granularity matching module, and a fine granularity matching module (from left to right).

2.2.1Feature Extraction Module: Assuming that the target TVS image to be matched and the calibrated reference image sample from the RS imagery are represented as image pairs, we refer to them as I^A and I^B, respectively. We first employ a convolutional feature extraction layer to extract multi-level, rich feature descriptors from the image pair (I^A and I^B). This process yields coarse features (C^{cA} and C^{cB}) at 1/8 the original image dimensions and fine features (C^{fA} and C^{fB}) at 1/2 the original image dimensions. Figure 3 illustrates the details of the convolutional feature extraction layer. It is worth noting that the standard convolutional architecture used in the aforementioned process possesses the inductive bias of unique locality and translation invariance, making it particularly proficient at capturing local feature information from images. Subsequently, we introduce a downsampling operation to reduce the input length of the local feature transformation module, thereby decreasing computational costs.

Figure 3. Convolutional Feature Extraction Layer.

As the processes shown in Figure 2, after feature extraction, we can achieve image comparison at any angle through the steps of dual-feature matching- coarse matching and fine matching. First, through pixel-wise dense matching, a relationship is established

between the target TVS image and the reference image (obtained by overlaying the orthoimage acquired through satellite remote sensing with the map in the orthogonal coordinate system), and then the matching accuracy between the two is further improved. In the feature matching process, referencing the LoFTR algorithm (Sun et al. 2021), we constructed a Transformer-based feature transformation module to extract finer-grained features, as illustrated in Figure 4.

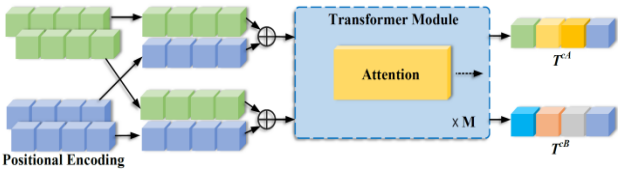

Figure 4. Transformer-based Feature Transformation Module.

As shown in Figure 4, the feature transformation module mainly consists of a position encoding module and a Transformer transformation module. This feature transformation module is capable of converting the coarse features (C^{cA} and C^{cB}) obtained from the feature extraction module into position and context-aware local features, which are subsequently transformed into easily matchable feature representations (T_{tr}^{cA} and T_{tr}^{cB}). The Position Encoding (PE) module unfolds the acquired coarse features into long sequences, assigning positional information to each element of features C^{cA} and C^{cB}. We then add the position encoding information to the image pairs C^{cA} and C^{cB}, ensuring that each element of the coarse features has a unique position, which is beneficial for detecting feature points in areas where features are not prominent.

After the PE by the Encoder module, we feed the feature elements with added positional information into the Transformer transformation module. The core of the Transformer transformation module is the attention mechanism, which primarily consists of query vectors ($\boldsymbol{Q}$), key vectors ($\boldsymbol{K}$), and value vectors ($\boldsymbol{V}$). Similar to information retrieval, the query vector $\boldsymbol{Q}$ computes attention weights based on the dot product with the corresponding key vector $\boldsymbol{K}$ for each value $\boldsymbol{V}$, allowing information to be retrieved from the value vectors $\boldsymbol{V}$. Mathematically, the attention mechanism can be represented as:

$$Attention(\boldsymbol{QKV}) = softmax(\boldsymbol{QK}^T)\boldsymbol{V} \quad (2)$$

As shown in Equation (2), the attention mechanism selects relevant information by calculating the similarity between the $\boldsymbol{Q}$ vector and each $\boldsymbol{K}$ vector. The output vector is the weighted sum of the $\boldsymbol{V}$ vectors based on the similarity scores, ultimately extracting relevant information from the most similar $\boldsymbol{V}$ vectors. Let the lengths of $\boldsymbol{Q}$ and $\boldsymbol{K}$ be represented as N, and their feature dimensions as D. Due to the computational complexity of the dot product operation between $\boldsymbol{Q}$ and $\boldsymbol{K}$ increasing with the length of the input sequences, the computational cost of the aforementioned operations grows quadratically, yielding a complexity of $O(N^2)$. As illustrated in Figure 3, the Transformer transformation module consists of M attention layers. After M attention computations, we obtain the easily matchable feature representations T_{tr}^{cA} and T_{tr}^{cB}.

2.2.2 Coarse Matching Module: This module is designed for the entire image and conducts a global search to find mutually matching areas, establishing pixel-wise dense matching between image pairs I^A and I^B. The module primarily consists of two processes: calculating correlation and normalization, followed by threshold filtering. First, it utilizes the dual-softmax operator as a differentiable matching layer, using Equation 3 to compute the score matrix S between the transformed features:

$$S(i,j) = \frac{1}{\tau} \cdot \langle T^{cA}(i), T^{cB}(j) \rangle \quad (3)$$

Then, applying the dual-softmax operator on the two dimensions of S to obtain the probabilities of nearest neighbor matching. Specifically, the matching probability for coarse matching, denoted as P_C, can be modelled as:

$$P_C(i,j) = softmax\big(S(i,\cdot)\big)_j \cdot softmax\big(S(\cdot,j)\big)_i \quad (4)$$

Finally, based on the confidence matrix P_C, we select matches with confidence levels exceeding the threshold θ_C and further apply the Symmetric Nearest Neighbor (SNT) criterion to filter out irrelevant features. This process can be modelled as the following equation:

$$M_C = \{(i,j) | \forall (i,j) \in \text{SNT}, P_C(i,j) \geq \theta_C\} \quad (5)$$

2.2.3 Fine Matching Module: After establishing coarse matches, we use the fine matching module to refine the aforementioned feature matches to the fine features at 1/2 the original image dimensions.

For each coarse match $(\tilde{\imath}, \tilde{\jmath})$, we first locate its position (i,j) on the fine feature maps C^{fA} and C^{fB}, and then crop two local windows to size W*W. The cropped features within each window will be transformed L times through the Transformer transformation module, producing two local feature maps centered at i and j, referred to as $T_{tr}^{fA}(i)$ and $T_{tr}^{fB}(j)$, respectively. Next, we associate the center vector of $T_{tr}^{fA}(i)$ with all vectors in $T_{tr}^{fB}(j)$, generating a heatmap. This heatmap represents the matching probability of each pixel in the neighborhood of j with respect to i. By calculating the expected value over the probability distribution, we obtain the final position on I^B with sub-pixel accuracy. Finally, we collect all matches{(i,j)} to produce the final fine matching set $M_f = \{(i,j)\}$.

This dual-feature matching mechanism is capable of finding a large number of corresponding point pairs even in areas with sparse texture, significantly enhancing matching accuracy and laying the groundwork for subsequent geographic localization.

2.3 Homography Transformation and Positional Information Conversion

After obtaining corresponding points between image pairs through feature matching, the DFMC utilizes homography transformation to establish the mapping relationship between real-time monitoring images and standard images. Homography transformation, which incorporates perspective projection, can handle objects at both close and far distances, taking into account the camera's position and orientation, thereby preserving the properties of lines and parallel lines in the images. The essence of homography transformation is calculated through the homography matrix $\boldsymbol{H}$. Here, $\boldsymbol{H}$ can be defined as:

Figure 5. Localization Results for Xingfu Suburban Plain Areas at Focal Lengths of 1x, 4x, 7x, and 10x.

$$\boldsymbol{H} = \begin{bmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ H_{31} & H_{32} & 1 \end{bmatrix} \tag{6}$$

where the homography matrix $\boldsymbol{H}$ contains 8 unknowns; therefore, four pairs of non-collinear corresponding points are sufficient to solve for all the unknowns in the homography transformation matrix.

Assuming the coordinates of the corresponding points in the two images are represented as $\boldsymbol{P} = [x, y, 1]^T$ and $\boldsymbol{Q} = [u, v, 1]^T$ where $\boldsymbol{P}$ is a point in the reference image and $\boldsymbol{Q}$ is a point in the target TVS image frame to be matched. Given the homography transformation matrix **H**, the relationship between $\boldsymbol{P}$ and $\boldsymbol{Q}$ can be expressed as:

$$\boldsymbol{P} = \boldsymbol{H} \cdot \boldsymbol{Q} \tag{7}$$

$$[x, y, 1]^T = \boldsymbol{H} \cdot [u, v, 1]^T \tag{8}$$

or:

$$\boldsymbol{x} = H_{11}u + H_{12}v + H_{13} \tag{9}$$

$$\boldsymbol{y} = H_{21}u + H_{22}v + H_{23} \tag{10}$$

The above equations allow for the calculation of the true geographic coordinates of any point in the target TVS image frame. Moreover, this mapping relationship is reversible. For the region of the orthoimage covered by the current TVS image frame, the corresponding point in the surveillance image can be found for any point in the orthophoto through the inverse process.

3. Results

3.1 Experimental Datasets

This study utilizes remote sensing image data provided by the 'YunYao-1' series of satellites, which employs the CGCS2000 coordinate system. The tower-mounted video surveillance images are selected from the Hunan Province Tower Sentinel System. To train the DFMC method for feature extraction from TVS and RS reference images, the MegaDepth dataset (Li and Snavely, 2018) is used as the training dataset, with the training method referencing the LoFTR (Sun et al. 2021). This dataset contains millions of real-world images obtained from the internet, covering a diverse range of landscapes and rich structural scenes. To test and validate the method's effectiveness in complex suburban environments, we utilize real-time screenshots captured under various lighting and weather conditions using tower-mounted video equipment. The test data includes TVS data from the Xingfu suburban area, near the Yiyang city center, and the Changputang suburban area, adjacent to Fenghuang City, both located in Hunan Province. The camera view in Xingfu primarily covers farmland, characterized by flat terrain and few obstructions. In contrast, Changputang are mountainous areas with numerous hills, villages, and forests, featuring significant terrain undulation and many obstructions.

Additionally, due to the large coverage area of the surveillance cameras, it is necessary to adjust the camera parameters (D for rotation angle, T for azimuth angle, and Z for the multiple of focal length) to ensure accurate monitoring of farmlands that are distant from the monitoring range, thereby enlarging the monitoring area for subsequent experiments. Accordingly, the dimensions of the altered surveillance images must be adjusted based on specific equations. To enhance the accuracy of image matching, we sort the matching points according to confidence levels, selecting only the top ten pairs of corresponding points with the highest confidence for the homography transformation. Given that the camera's field of view is 360°, we capture a surveillance image every 60° (referred to as standard images), and implement point modeling based on pseudo-circular scene images, completing the selection of control points for each image relative to the RS images. After the control point pairs are selected, we use a mathematical model of cross-mapping between the two-dimensional geographic information system under height constraints and the TVS image to align points across different coordinate systems.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-173-2025 | © Author(s) 2025. CC BY 4.0 License.

Figure 6. Localization Results for Changputang Village in Suburban Hilly Areas at Focal Lengths of 1x, 1x, 3x and 7x.

3.2 Evaluation Index

Due to the unique characteristics of suburban areas, we primarily employ visualization methods to assess the localization effectiveness. Visualization provides an intuitive representation of complex data in a graphical format, aiding in the analysis of regional features and making it easier to identify patterns, trends, and anomalies, while also improving the accuracy of localization. Additionally, Root Mean Square Error (RMSE) is intuitive as it is expressed in the same units as the original data, making it easy to interpret and compare, while also being sensitive to large errors, effectively highlighting potential issues with the model's performance. Thus, we use the RMSE to measure the error between the observed values and the actual measurements, as shown in the following equation:

$$RMSE = \sqrt{\frac{\sum(x-x')^2+\sum(y-y')^2}{M}} \quad (11)$$

Here, (x, y) are the geographic coordinates calculated for the selected M points, and (x', y') are the real geographic coordinates.

Video Check Frames	RMSE (meter)	Video Check Frames	RMSE (meter)
Frame 1	1.3972	Frame 4	0.8494
Frame 2	1.5406	Frame 5	1.6075
Frame 3	0.9137	Frame 6	1.5642
Average		1.3121	

Table 1. RMSE Localization Errors in Suburban Plain Areas

3.3 Experimental Localization Effects

3.3.1 Localization Effect in Suburban Plain Areas: Figure 5 illustrates the localization results of the proposed DFMC for the TVS images and RS imagery of Xingfu suburban plain areas at different focal lengths (Z = 1, 4, 7, 10). As seen in Figure 5, the areas highlighted by the red lines in both the orthoimage and the surveillance video images correspond to the same area, with a high degree of boundary overlap. Additionally, Table 1 presents the localization error values for six randomly selected video image frames in Xingfu suburban area, revealing a maximum localization error of 1.6075 and an average error of 1.3121, less than 1.5 meters. This indicates that the matching performance between the RS orthoimage and the TVS images is satisfactory, suggesting that the proposed DFMC method is suitable for the geographic localization needs of tower surveillance videos in suburban plain areas.

3.3.2 Localization Effect in Suburban Hilly Areas: Figure 6 illustrates the localization results of the proposed DFMC for the TVS images and RS imagery of Changputang suburban hilly areas at different focal lengths (Z = 1, 1, 3, 7). As shown in Figure 6, the boundaries of the areas highlighted by the red lines in both the orthoimage and the surveillance video images have a high degree of overlap, indicating good localization performance. Additionally, Table 2 presents the localization error values for six randomly selected video check image frames in Changputang suburban hilly area, revealing a maximum localization error of 3.3014 and an average error of 2.9881, less than 3 meters. The geographic localization requirements for TVS in suburban hilly areas are also satisfied by the proposed DFMC method.

Video Check Frames	RMSE (meter)	Video Check Frames	RMSE (meter)
Frame 1	3.0158	Frame 4	3.3014
Frame 2	2.9488	Frame 5	3.1254
Frame 3	2.6458	Frame 6	2.8914
Average		2.9881	

Table 2. RMSE Localization Errors in Suburban Hilly Areas

Compared to the suburban plain areas, the overlap has decreased, but the matching performance between the orthoimage and the ground TVS images remains satisfactory. In actual production processes, plain areas typically require higher positioning accuracy to support activities such as agriculture and urban planning. In contrast, hilly regions have relatively relaxed accuracy requirements due to complex terrain, lower application demands, potential signal obstruction by mountains and other obstacles, and variable climatic factors. Therefore, the proposed DFMC method is suitable for the geographic localization needs of TVS in complex suburban areas.

4. Discussion

The proposed DFMC demonstrates significantly higher localization accuracy at short focal lengths, particularly when the target area is close to the camera. This enhanced accuracy is primarily attributed to improved feature detection, reduced distortion effects, and higher image quality. However, the localization accuracy of the DFMC method decreases when the target is at a greater distance, due to challenges such as reduced image resolution, increased environmental interference, and limited feature visibility. To address these issues, future research should consider implementing a partitioning approach, which involves dividing the target area into smaller sections for localized processing, alongside multi-scale data fusion and dynamic focal length adjustment to enhance feature recognition and matching capabilities. By adopting these strategies, it is possible to significantly improve localization accuracy at varying distances, thereby extending the applicability of the DFMC method in real-world scenarios.

Moreover, to enhance localization accuracy, we can improve the accuracy of geographic registration by integrating high-resolution DEMs. High-resolution DEMs not only provide elevation information of the Earth's surface but also facilitate more accurate localization of complex topographical features in three-dimensional space and help identify terrain occlusion issues. By analyzing elevation changes, we can determine which areas may be occluded, thereby optimizing monitoring angles and line-of-sight selections to avoid data loss or mismatches caused by occlusion. Additionally, high-resolution DEMs can detail topographical features such as valleys, ridges, and rivers. These features can serve as important reference points in map matching, enhancing the correlation between different data sources and further improving the localization accuracy of the DFMC method. Additionally, future research could explore the integration of multi-source data, such as LiDAR data and point cloud data, leveraging the advantages of aerial, terrestrial, and satellite remote sensing. This approach aims to create a comprehensive three-dimensional model better suited for complex environments, thereby further improving the accuracy of map matching.

Finally, the proposed DFMC method currently cannot achieve geographical localization from TVS in real time, as this process requires approximately 2 seconds. This delay may hinder the application of the method in scenarios that demand immediate responses, such as security monitoring and emergency response. To address this limitation, we will explore parallel computing techniques and hardware acceleration strategies, such as utilizing Graphics Processing Units (GPUs) or Field-Programmable Gate Arrays (FPGAs). By leveraging these technologies, it is possible to significantly enhance the processing speed of the DFMC algorithm, enabling it to handle multiple video streams simultaneously and reduce the time required for localization.

5. Conclusions

This paper proposes a dual-feature matching-based collaborative remote sensing precision localization method tailored for complex suburban areas. The method initially establishes a mapping relationship between the pixel coordinate system of TVS images and the projected Cartesian coordinate system of RS orthoimages to accomplish georeferencing. Subsequently, the DFMC method conducts feature detection and dual feature image matching for both the TVS image and the aligned standard image to enable image matching from any angle. Finally, homography transformation is utilized to project the coordinates from the TVS image frames onto the RS reference images, facilitating rapid coordinate conversion and positioning.

As a result, the proposed DFMC method effectively addresses issues in the existing collaborative accurate localization of high-resolution satellite remote sensing imagery and ground tower-mounted video surveillance images, such as inaccurate image feature matching and weak anti-interference capabilities, making it particularly suitable for application in complex suburban areas. Using tower surveillance videos from Hunan Province and orthoimages with a resolution of 0.5 meters from YunYao, the results indicate that our proposed DFMC achieves an average localization accuracy better than 1.5 meters in flat suburban areas and better than 3 meters in complex hilly suburban areas.

6. Acknowledgements

This research was supported in part by the National Key Research and Development Program of China with grant number 2023YFB3906102, in part by the Key R & D projects in Yunnan Province with grant number 202403ZC380001, and the Fundamental Research Fund Program of LIESMARS with grant number 4201-420100071.

References

Casagli, N.; Intrieri, E.; Tofani, V.; Gigli, G.; Raspini, F. Landslide Detection, Monitoring and Prediction with Remote-Sensing Techniques. *Nature Reviews Earth & Environment* **2023**, *4*, 51–64, doi:10.1038/s43017-022-00373-x.

Haight, J.D.; Hall, S.J.; Fidino, M.; Adalsteinsson, S.A.; Ahlers, A.A.; Angstmann, J.; Anthonysamy, W.J.B.; Biro, E.; Collins, M.K.; Dugelby, B.; et al. Urbanization, Climate and Species Traits Shape Mammal Communities from Local to Continental Scales. *Nature Ecology & Evolution* **2023**, *7*, 1654–1666, doi:10.1038/s41559-023-02166-x.

Li, Z.; Snavely, N. MegaDepth: Learning Single-View Depth Prediction From Internet Photos. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 2018.

Milosavljević, A.; Rančić, D.; Dimitrijević, A.; Predić, B.; Mihajlović, V. Integration of GIS and Video Surveillance. *International Journal of Geographical Information Science* **2016**, *30*, 2089–2107, doi:10.1080/13658816.2016.1161197.

Milosavljević, A.; Rančić, D.; Dimitrijević, A.; Predić, B.; Mihajlović, V. A Method for Estimating Surveillance Video Georeferences. *ISPRS International Journal of Geo-Information* **2017**, *6*, doi:10.3390/ijgi6070211.

Okolie, C.J.; Smit, J.L. A Systematic Review and Meta-Analysis of Digital Elevation Model (DEM) Fusion: Pre-Processing,

Methods and Applications. *ISPRS Journal of Photogrammetry and Remote Sensing* **2022**, *188*, 1–29, doi:https://doi.org/10.1016/j.isprsjprs.2022.03.016.

Ouyang, X.; Xu, J.; Li, J.; Wei, X.; Li, Y. Land Space Optimization of Urban-Agriculture-Ecological Functions in the Changsha-Zhuzhou-Xiangtan Urban Agglomeration, China. *Land Use Policy* **2022**, *117*, 106112, doi:https://doi.org/10.1016/j.landusepol.2022.106112.

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 International Journal of Geo-Information* **2020**, *9*, doi:10.3390/ijgi9070448.

Cao, S.; Jin, X.; Yang X.; Sun, R.; Liu, J.; Bo, H.; Xu, W.; Zhou, Y. Coupled MOP and GeoSOS-FLUS models research on optimization of land use structure and layout in Jintan district. *Journal of Natural Resources* **2019**, 34, 1171, doi:https://doi.org/10.31497/zrzyxb.20190604.

Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-Free Local Feature Matching with Transformers. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; **2021**; pp. 8922–8931.

Wang, L.; Zhang, S.; Tang, L.; Lu, Y.; Liu, Y.; Liu, Y. Optimizing Distribution of Urban Land on the Basis of Urban Land Use Intensity at Prefectural City Scale in Mainland China. *Land Use Policy* **2022**, *115*, 106037, doi:https://doi.org/10.1016/j.landusepol.2022.106037.

Wu, Z.; Zhang, C.; Gu, X.; Duporge, I.; Hughey, L.F.; Stabach, J.A.; Skidmore, A.K.; Hopcraft, J.G.C.; Lee, S.J.; Atkinson, P.M.; et al. Deep Learning Enables Satellite-Based Monitoring of Large Populations of Terrestrial Mammals across Heterogeneous Landscape. *Nature Communications* **2023**, *14*, 3072, doi:10.1038/s41467-023-38901-y.

Yang, J.; Gong, P.; Fu, R.; Zhang, M.; Chen, J.; Liang, S.; Xu, B.; Shi, J.; Dickinson, R. The Role of Satellite Remote Sensing in Climate Change Studies. *Nature Climate Change* **2013**, *3*, 875–883, doi:10.1038/nclimate1908.

Zhang, C.; Su, Y.; Yang, G.; Chen, D.; Yang, R. Spatial-Temporal Characteristics of Cultivated Land Use Efficiency in Major Function-Oriented Zones: A Case Study of Zhejiang Province, China. *Land* **2020**, *9*, doi:10.3390/land9040114.