remote sensing

MDPI

Article

Quasi-Dense Matching for Oblique Stereo Images through Semantic Segmentation and Local Feature Enhancement

Guobiao Yao [1,2], Jin Zhang [1], Fengqi Zhu [3], Jianya Gong [2,*], Fengxiang Jin [4], Qingqing Fu [1] and Xiaofang Ren [1]

[1] School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan 250101, China; 13565@sdjzu.edu.cn (G.Y.); 2021160106@stu.sdjzu.edu.cn (J.Z.); 13622@sdjzu.edu.cn (Q.F.); 11812@sdjzu.edu.cn (X.R.)
[2] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430070, China
[3] Shandong Provincial Institute of Land Surveying and Mapping, Jinan 250101, China; zhufengqi@shandong.cn
[4] College of Geomatics, Shandong University of Science and Technology, Qingdao 266590, China; fxjin@sdjzu.edu.cn
* Correspondence: gongjy@whu.edu.cn; Tel.: +86-027-6877-8003

Abstract: This paper proposes a quasi-dense feature matching algorithm that combines image semantic segmentation and local feature enhancement networks to address the problem of the poor matching of image features because of complex distortions, considerable occlusions, and a lack of texture on large oblique stereo images. First, a small amount of typical complex scene data are used to train the VGG16-UNet, followed by completing the semantic segmentation of multiplanar scenes across large oblique images. Subsequently, the prediction results of the segmentation are subjected to local adaptive optimization to obtain high-precision semantic segmentation results for each planar scene. Afterward, the LoFTR (Local Feature Matching with Transformers) strategy is used for scene matching, enabling enhanced matching for regions with poor local texture in the corresponding planes. The proposed method was tested on low-altitude large baseline stereo images of complex scenes and compared with five classical matching methods. Results reveal that the proposed method exhibits considerable advantages in terms of the number of correct matches, correct rate of matches, matching accuracy, and spatial distribution of corresponding points. Moreover, it is well-suitable for quasi-dense matching tasks of large baseline stereo images in complex scenes with considerable viewpoint variations.

Keywords: oblique stereo images; deep learning; semantic segmentation; weak texture feature matching; quasi-dense matching

Citation: Yao, G.; Zhang, J.; Zhu, F.; Gong, J.; Jin, F.; Fu, Q.; Ren, X. Quasi-Dense Matching for Oblique Stereo Images through Semantic Segmentation and Local Feature Enhancement. *Remote Sens.* **2024**, *16*, 632. https://doi.org/10.3390/rs16040632

Academic Editor: Shuying Li

Received: 27 October 2023
Revised: 18 January 2024
Accepted: 25 January 2024
Published: 8 February 2024

1. Introduction

In recent years, obtaining high-resolution multiview images of ground scenes has become increasingly easier with the development of ground mobile wide-baseline photography, UAV oblique photography, and other technologies [1]. However, under large viewpoint conditions, substantial changes in the main optical axis can lead to substantial distortions or masking in the scale, orientation, surface brightness, and neighborhood information of the same spatial target in stereo images. In addition, the existence of a large number of weak texture areas poses a great challenge for dense image matching and automated image processing [2–4].

Classical image matching methods can be divided into two categories: grayscale and feature matching. Grayscale-based matching algorithms use the grayscale information of the image to determine the similarity of the image for matching. Common grayscale algorithms include normalized cross-correlation (NCC) [5], mean absolute difference (MAD) [6], and least square matching (LSM) [7]. These grayscale matching algorithms have high accuracy but require high computation and are sensitive to noise. Feature-based matching methods first detect features in the image, subsequently extract feature descriptors, and

finally determine matching features based on the Euclidean distance of the descriptors. Scale-invariant feature matching methods, represented by SIFT, exhibit good scale invariance but are difficult to adapt to considerable changes in viewpoint [8]. Reference [9] optimized SIFT feature points using the NCC method, improving the matching accuracy. Reference [10] constructed a feature extraction method combining filter decomposition and phase consistency rules, and employed a Gaussian mixture model to determine matching points. Reference [11] proposed an affine-invariant oblique image matching method that estimates the initial affine transformation based on image orientation parameters, corrects the image based on the affine transformation, and finally performs SIFT matching on the corrected image. Reference [12] simulated the full-range viewpoint change of the image and performed SIFT feature matching. This method exhibits good affine invariance; however, obtaining matching features in weak texture areas is challenging.

With the rapid development of computer software and hardware, deep learning methods based on convolutional neural networks (CNNs) have opened up a new way for realizing image matching. Deep learning matching is a data-driven image matching method that can autonomously learn the deep-level representation of object features from a large amount of image data. Currently, deep learning matching is classified into dense and sparse matching. The former achieves pixel-by-pixel dense correspondence in overlapping areas by predicting the disparity map of stereo images and the latter is oriented toward feature extraction, description, and matching for staged training and optimization with high matching reliability, such as the classical L2-Net [13]. HardNet [14,15] enhances the differentiation between descriptors by constraining the distance between nonsynonymous descriptors through a loss function based on L2-Net. AffNet proposed in reference [16] uses multiscale Hessian to detect feature point locations, followed by HardNet and its loss function to estimate the affine neighborhood. R2D2 achieved improvements in network architecture, training strategy, and visualization methods as well as improved the computational efficiency and robustness through separable convolutional layers [17]. Inspired by SuperGlue [18], reference [19] introduced the position encoding and attention mechanism using the Transformer network to construct a model called LoFTR, which has texture enhancement capabilities. This method considerably improved the matching performance in weak texture areas; however, adapting to changes in the viewpoint of the images is challenging. Reference [20] proposed a performance baseline for deep feature matching called DFM. It adopts a two-stage approach, where the initial transformation is performed using feature information containing rich deep semantic information. Then, through hierarchical matching from deep to shallow and coarse to fine levels, the final matching pairs are obtained. Similarly inspired by SuperGlue [18], the GlueStick uses a depth map neural network to unify the descriptors of points and lines into one framework, and employs the information between points to glue the lines from the matching images, improving the joint matching efficiency of the model. This indicates that the complementary performance of using two features in a single framework greatly improves performance [21]. Furthermore, reference [22] proposed an end-to-end deep learning network and its weighted average loss function for wide-baseline image matching with high inclination angles. This approach allows nonmatching similarity descriptors to participate in training through weighting, improving the discriminability of nonmatching descriptors and matching performance of matching descriptors. However, adapting to images with multiple planar scenes and oblique perspectives is difficult. VGG16 is a classic deep CNN model comprising 16 convolutional and three fully connected layers with powerful feature extraction capabilities [23]. UNet is a deep learning model for semantic segmentation tasks that comprises symmetric encoder and decoder parts and can achieve pixel-level image segmentation [24]. Reference [25] proposed an integrated VGG16-UNet, which has demonstrated some reliability in image classification and segmentation tasks and provided a feasible method for image segmentation and matching in multiplanar complex scenes.

In summary, for oblique stereo images with complex scenes and geometric distortions, it is difficult to achieve more reliable dense matching results using both classical feature

matching algorithms and deep learning matching strategies. Deep learning segmentation models and texture-enhanced convolutional networks are expected to be the breakthrough in solving such image matching problems. Therefore, this paper proposes a reliable quasi-dense feature matching algorithm that combines image semantic segmentation and local feature enhancement network, which integrates the VGG16-UNet multiplanar semantic segmentation and LoFTR local feature enhancement network. The proposed algorithm first segments and extracts the corresponding planar scene and then applies the weak texture enhancement strategy in the planar scene to obtain quasi-dense feature matching. The effectiveness of the proposed method is verified using actual stereo images of complex scenes with large viewpoints.

2. Materials and Methods

For stereo images of multiplanar scenes with large viewpoints, we first train the VGG16-UNet model using typical segmented data of oblique multiplanar scenes, achieving preliminary segmentation of complex scenes into individual planes. Subsequently, we employ a neighborhood search-based adaptive thresholding strategy to optimize the segmented local regions. Afterward, we use affine-invariant feature matching to recognize corresponding planes and apply the LoFTR method with local feature transformation to extract weak texture features for each identified plane. Finally, we fuse the results of local plane matching and obtain a semi-dense matching result. Figure 1 shows the technical approach of the proposed algorithm.

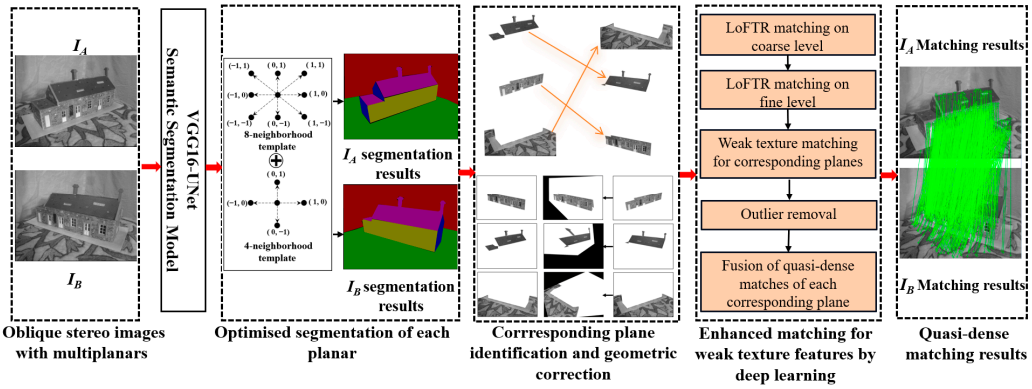

Figure 1. Technical approach of the proposed algorithm.

2.1. Automatic Semantic Segmentation Strategy

2.1.1. Multiplanar Semantic Segmentation Model

VGG16-UNet is a deep CNN model based on the fusion of VGG16 and Unet models. It combines the powerful feature extraction capability of VGG16 and pixel-level semantic segmentation capability of Unet. To cope with the quasi-dense matching task of complex scenes, we propose to apply VGG16-UNet to the semantic segmentation of multiplanar scenes. Figure 2 shows the model structure and the design of each parameter. In the encoding stage, the first 13 convolutional layers of VGG are used as the feature extraction network, and a 3×3 size convolutional kernel is used to compress the input image from 512×512 pixels to $32 \times 32 \times 512$ pixels after four down-sampling steps to achieve the feature extraction from multiplanar scenes. In the corresponding decoding part, up-sampling and feature fusion are used to complete the segmentation of each plane, and the decoder restores the final output layer size to 512×512 pixels through continuous up-sampling and convolutional stacking, and subsequently outputs the segmentation map.

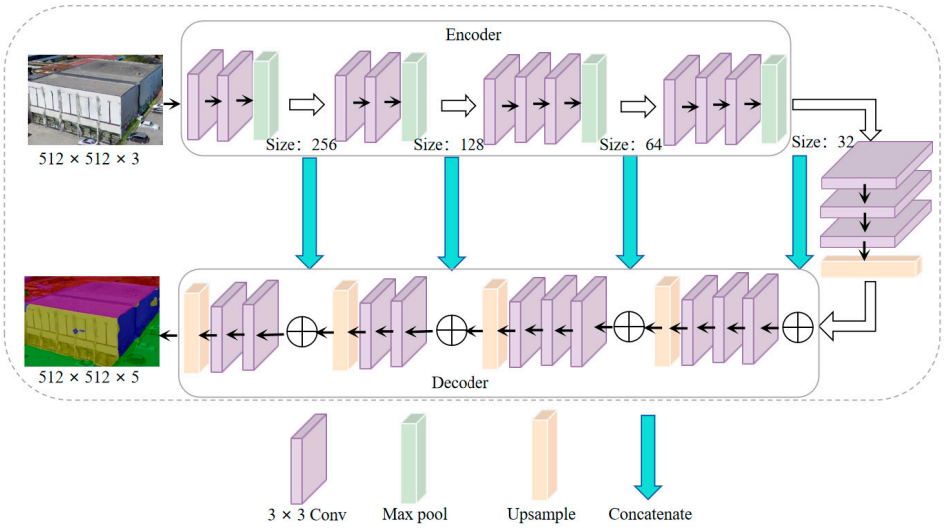

Figure 2. VGG16-UNet network architecture diagram.

2.1.2. Training Data

Extensive testing has revealed that VGG16-UNet has strong feature extraction capabilities and good transfer learning performance. Therefore, to fully train the VGG16-UNet model, we carefully selected 80 typical building image data of various types. These data are taken from low-altitude oblique views, and due to the presence of occlusion factors, the buildings in the images show one top and two side views, with a paucity of texture on the scene surface (Figure 3). These data are manually labeled into five sections: building top (pink), building facade (yellow or purple), ground (green), and background (dark red), corresponding to the 80 labeled images. Figure 3 shows an example of the training data.

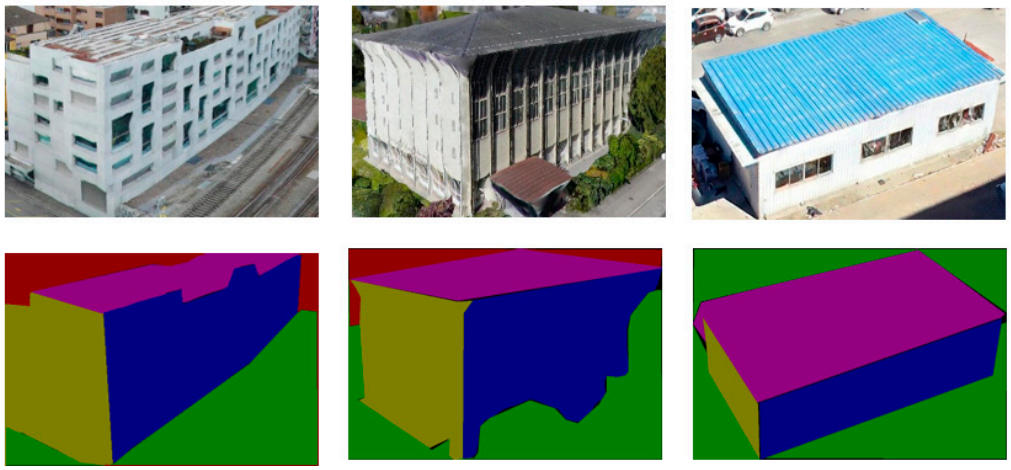

Figure 3. Example of training dataset for image segmentation used in this paper.

2.1.3. Image Segmentation and Adaptive Optimization

The trained VGG16-UNet automatically predicts a set of pixels located in the same plane, extracts the mask map of each plane, and completes the initial segmentation of the local plane. However, some mask maps may contain holes, and the use of segmentation results at this point will inevitably affect subsequent matching results. Therefore, we propose an adaptive optimization method. Figure 4 shows a schematic of adaptive optimization, which mainly includes discrete region removal outside the main plane region and the filling of the hole region in the main plane region. Removal of the discrete region eliminates

segmentation noise outside the main plane, whereas filling the hole eliminates the noise inside the main plane.

$$r = \frac{1}{2max(R)},\ R = \cup_{i=1}^{n} S_i, \tag{1}$$

where R represents the area of the connected region, n denotes the number of iterations, and S_i represents the area of the region obtained by expanding in the i-th iteration. The maximum value of R corresponds to the area of the main plane region.

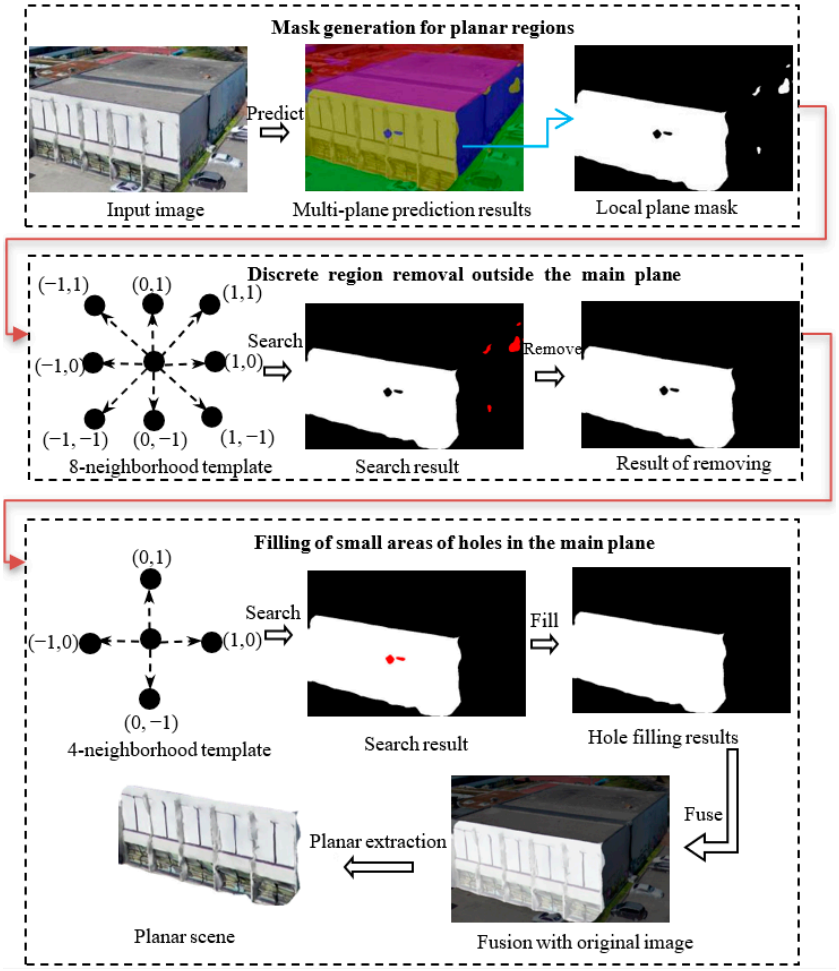

Figure 4. Adaptive optimization for local plane segmentation.

Next, to obtain the optimized results inside and outside the main region, the locally connected regions are color-inverted based on the adaptive threshold r. Considering the optimization accuracy and efficiency, an eight-neighborhood template is used to retrieve small discrete areas, whereas a four-neighborhood template is used to fill small hole areas. When segmenting, the local plane segmentation can be achieved by performing a Bitwise-AND operation between the mask map and the original image. This operation results in an image content containing only the mask region. Figure 5 shows the effect of each plane segmentation optimization. It shows that the proposed strategy achieves adaptive optimization of the planar scene by correcting the noise in the internal and external body regions, ensuring the reliability of segmentation and laying the foundation for subsequent quasi-dense matching of the planar scenes.

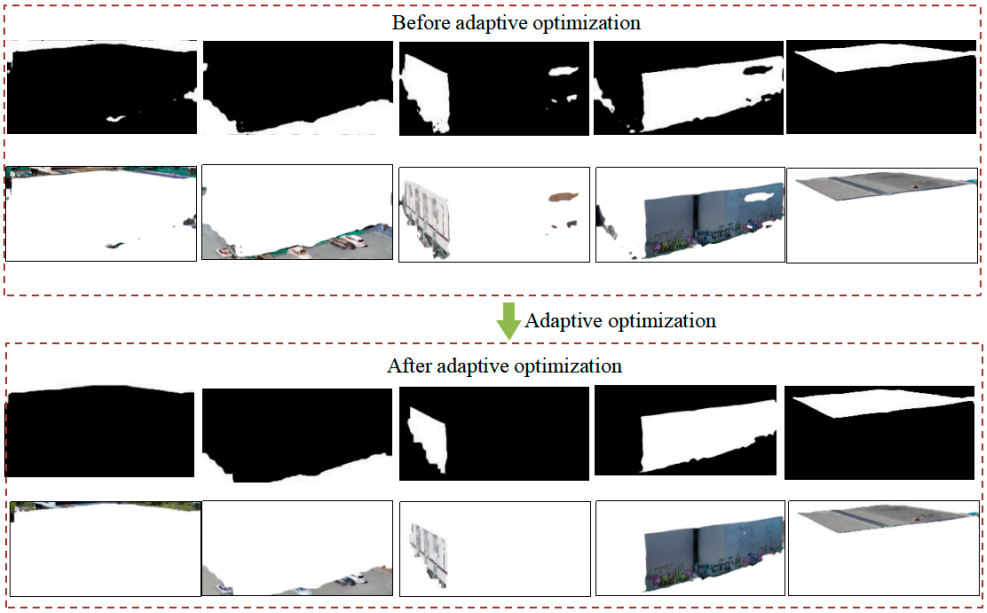

Figure 5. Comparison before and after plane segmentation optimization.

2.2. Quasi-Dense Matching Method

2.2.1. Automatic Identification of Corresponding Planes

Before performing the enhancement matching of weak texture features, it is necessary to first pair and recognize the corresponding plane scenes in the left and right images for obtaining corresponding planes. The affine-invariant feature matching algorithm described in reference [22] can robustly extract corresponding features from plane scenes with large viewpoint variations. Therefore, in this section, we employ this algorithm to automatically recognize corresponding planes. The process can be briefly described as follows: extract any plane from the left image, match it with each plane in the right image, identify the corresponding plane with the most corresponding features, and $>m$ (matching points, set to 8) is identified as the corresponding plane. Similarly, we iterate through all the planes in the left image, complete the feature matching with each plane of the right image and discriminate, and finally obtain each corresponding plane pair. Figure 6 shows the automatic identification process.

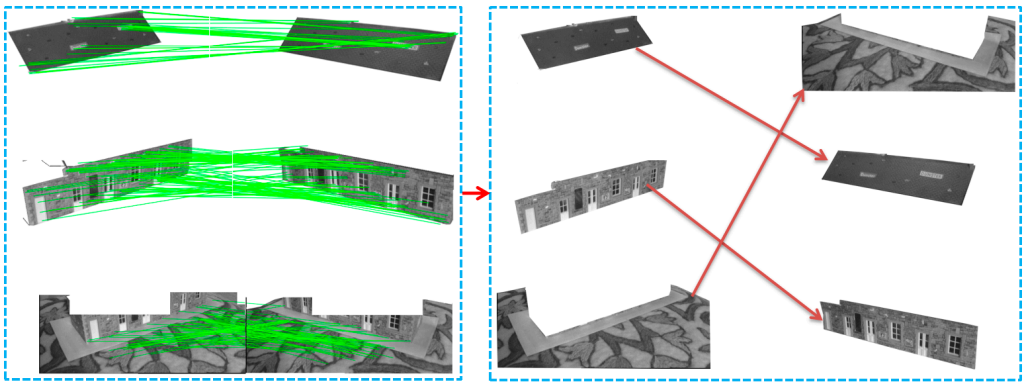

Figure 6. Process of automatic identification of corresponding planes.

2.2.2. LoFTR-Based Weak Texture Feature Enhancement Matching

The LoFTR strategy proposed in reference [19] can effectively enhance feature distinctiveness in weak texture regions; however, it struggles to adapt to affine deformations between images. Therefore, in this section, we first estimate the perspective transformation

matrix based on the obtained corresponding planes and their corresponding features to minimize the geometric deformations between corresponding planes. Subsequently, we apply the LoFTR algorithm to extract weak texture features from the corresponding planes. Figure 7 shows the specific matching process, which primarily comprises the five key steps outlined below.

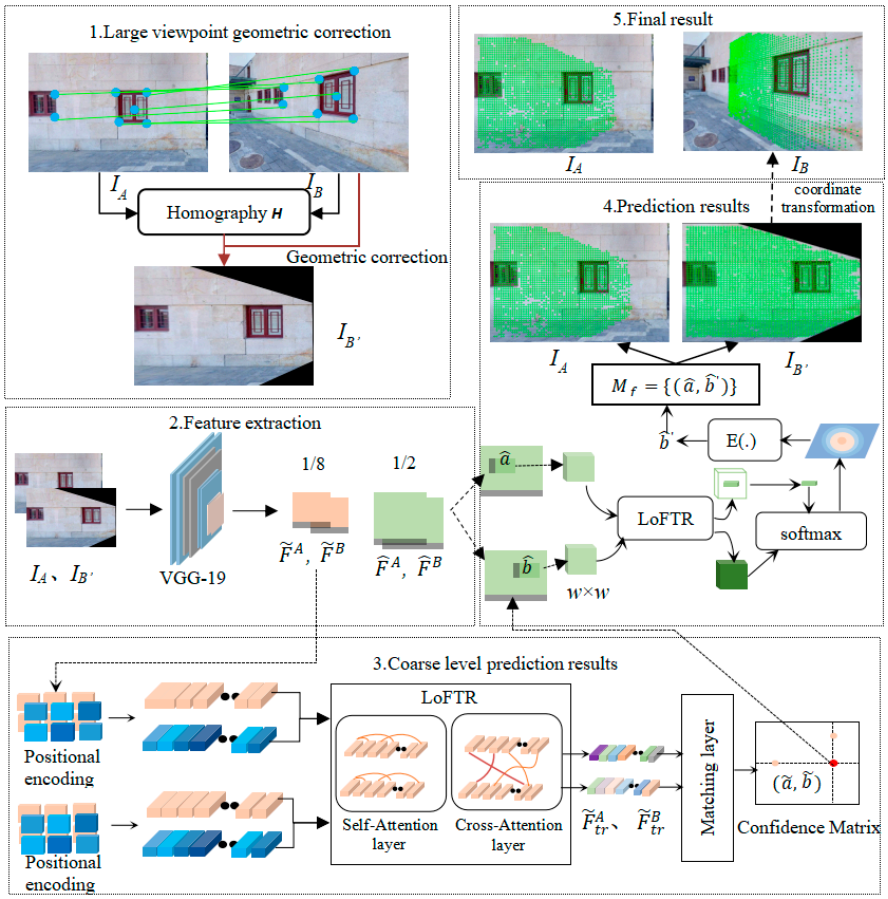

Figure 7. Weak texture feature enhancement matching for large viewpoint transformation.

(1) Large viewpoint correction: First, using the I_A and I_B matching points obtained in the previous plane recognition process, we estimate the projection transformation matrix H based on Equation (2) and the random sample consensus (RANSAC) algorithm as follows:

$$\begin{cases} x' = \frac{h_{11}x+h_{12}y+h_{13}}{h_{31}x+h_{32}y+h_{33}} \\ y' = \frac{h_{21}x+h_{22}y+h_{23}}{h_{31}x+h_{32}y+h_{33}} \end{cases}, \tag{2}$$

where (x, y) and (x', y') represent the feature matching points in I_A and I_B, respectively, and $h_{11}, h_{12} \ldots, h_{33}$ represent the nine projection transformation parameters in H. Subsequently, according to Equation (3), the right image is corrected through projective transformation as follows:

$$\begin{bmatrix} x'' \\ y'' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11}h_{12}h_{13} \\ h_{21}h_{22}h_{23} \\ h_{31}h_{32}h_{33} \end{bmatrix} \begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix}, \tag{3}$$

where (x', y') and (x'', y'') represent the pixel coordinates of I_B and I_B' before and after the correction of the projective deformation in the right image, respectively. After correcting the right image through projective deformation, the affine distortion of the corresponding region is considerably improved, and the geometric consistency

of the plane tends to be better than before. Thus, the LoFTR strategy is introduced for matching.

(2) Feature extraction: For the image pair I_A and $I_B{}'$, feature extraction is first performed using VGG CNN, resulting in 1/8 coarse feature and 1/2 fine feature maps for both images.

(3) Generating coarse-level feature prediction results: The coarse extracted feature maps $\tilde{F}^{A}$ and $\tilde{F}^{B}$ are flattened into one-dimensional vectors, and position encoding is added to each vector. These vectors with position encoding are then inputted into the LoFTR module, which comprises N (N = 4) self-attention and cross-attention layers. The LoFTR module utilizes a self-attention mechanism to capture the correlations between different positions within the image, learning the importance of local features and enhancing the discriminative ability of the convolutional model for different texture features. After processing through this module, two enhanced texture feature maps with higher discriminability, labeled as $\tilde{F}^{A}_{tr}$ and $\tilde{F}^{B}_{tr}$, are outputted. Subsequently, the similarity between these two feature maps is calculated to perform the matching of corresponding features.

(4) Outputting the prediction results: For any coarse-level matching prediction $\left(\tilde{a}, \tilde{b}\right) \in Mc$, local corresponding windows of size $w \times w$ (w = 5) are cropped from the fine feature maps $\hat{a}$,$\hat{b}$. Second, a smaller LoFTR module then transforms the cropped features within each window, yielding two transformed local feature maps, $\hat{F}^{A}_{tr}(\hat{a})$ and $\hat{F}^{B}_{tr}\left(\hat{b}\right)$, centered at $\hat{a}$ and $\hat{b}$, respectively. Third, we correlate the center vector of $\hat{F}^{A}_{tr}(\hat{a})$ with all vectors in $\hat{F}^{B}_{tr}\left(\hat{b}\right)$ and thus produce a heatmap that represents the matching probability of each pixel in the neighborhood of $\hat{a}$ with $\hat{b}$, and the location $\hat{b}'$ is obtained by calculating the expectation of the probability distribution. Finally, all coarse-level matches are refined within the local windows of the fine level, resulting in the fine-level matching predictions M_f for I_A and $I_B{}'$.

(5) Outputting the final result: Finally, the coordinates of the fine-level matching points on $I_B{}'$ are normalized to the original coordinate system of the right image I_B using Equation (3), representing the final result of weak texture feature-enhanced matching.

3. Results

3.1. Experimental Environment

In the experiment, we used RTX2080ti GPU, 9-9900K processor, 64 GB RAM, and Ubuntu18.04 operating system. The software platform is PyCharm (v 2023.3.2). The training dataset of weak texture feature based on LoFTR is adopted from the open-source MegaDepth dataset. During the training process, the Transformer loop count N is set to four, the LoFTR module feature transformation count N_f is set to one, and the window size w for extracting patches from the fine-level feature map is set to five. The threshold θ_C for coarse-level matching prediction is set to 0.2. The training is completed after 30 iterations using the gradient descent algorithm.

3.2. Evaluation Metrics

(1) Number of correct matching points, k_{ε_0}: Fifteen pairs of uniformly distributed corresponding points are manually selected from the stereo images. The fundamental matrix F_0 is estimated using the least-squares method and considered as the ground truth. Using the well-known fundamental matrix F_0, the error of any matching point is calculated using Equation (4). A threshold ε_0 (set to 3.0) is set and imposed for

the error. If the error was less than ε_0, the pair of points is a correct pair of matching points and is included in the count of correct matching points, k_{ε_0}:

$$\varepsilon_i = \sqrt{\left(x_i{}'^T F_0 x_i\right)^2 / ((F_0 x_i)_1^2 + (F_0 x_i)_2^2}. \quad (4)$$

(2) Match correct rate, α: This is defined by $\alpha = k_{\varepsilon_0}/k$, where k denotes the total number of matching points.

(3) Matching root-mean-squared error (RMSE) ε_{RMSE} (pixel). This is calculated using Equation (5):

$$\varepsilon_{\mathrm{RMSE}} = \sqrt{\frac{1}{k}\sum_{i=1}^{k} \varepsilon_i{}^2}, \quad (5)$$

where k represents the total number of matches and ε_i is calculated using Equation (4).

(4) Matching spatial distribution quality, $\hat{D}$: References [26,27] generated Delaunay triangulation based on the matching points. They evaluated the spatial distribution quality of the matching points by considering the areas and shapes of each triangle, as well as the global and local distribution of the matching points. This is calculated using Equation (6):

$$\left.\begin{array}{c} \mathrm{D} = \mathrm{D_A} \times \mathrm{D_S} = \sqrt{\frac{\sum_{i=1}^{n}\left(\left(\mathrm{A_i}/\overline{\mathrm{A}}\right)-1\right)}{n-1}} \times \sqrt{\frac{\sum_{i=1}^{n}(\mathrm{S_i}-1)}{n-1}},\ \overline{\mathrm{A}} = \frac{\sum_{i=1}^{n}\mathrm{A_i}}{n},\ \mathrm{S_i} = \frac{3\max(\mathrm{J_i})}{\pi} \\ \hat{D} = \frac{\mathrm{D}}{D_G},\ D_G = \left(\sum_{i=1}^{n} A_i\right)/A_I \end{array}\right\}, \quad (6)$$

where n represents the total number of generated triangles; $\mathrm{A_i}$ and $\max(\mathrm{J_i})$ represent the area and maximum arc of the i-th triangle, respectively; $\overline{\mathrm{A}}$ represents the average area of the triangles; $\mathrm{D_A}$ represents the uniformity of the areas of each triangle; and $\mathrm{D_S}$ represents the uniformity of the internal angles of the triangles. The lower the D value, the higher the geometric uniformity of the local triangles. $\mathrm{A_i}$ represents the area of the image and $\mathrm{D_G}$ represents the coverage of matching points in the global image. A higher $\mathrm{D_G}$ value indicates a wider spatial distribution of matching points in the image. Therefore, this model can fully reflect the quality of the matching point spatial distribution, and the quality of the matching point spatial distribution increases with decreasing $\hat{D}$.

3.3. Experimental Methods and Data

To fully validate the advantages of our proposed method, we used six methods for comparative testing. (1) DFM: This method achieves high accuracy by performing coarse-to-fine matching of images at different hierarchical levels of features. (2) AffNet: This method uses an affine-invariant estimation network to learn affine parameters. It enhances the distinctiveness between descriptors using the HardNet loss function, making it suitable for scenes with viewpoint changes. (3) SuperGlue: This method constructs an image information aggregation model based on attention mechanisms. The loss function of the model is established using graph neural networks. (4) LoFTR: This method combines position encoding and attention mechanisms in the Transformer, generating a model suitable for weak texture matching. (5) GlueStick: A GNN architecture is designed to be able to combine the contextual information of all features to improve the accuracy of the matching. (6) Our proposed method. To objectively evaluate these six methods, the RANSAC algorithm is used to remove outliers, and the inlier coordinates for each method are outputted. As shown in Figure 8, six groups of low-altitude large viewpoint building scene images (a–f) are selected as the test data.

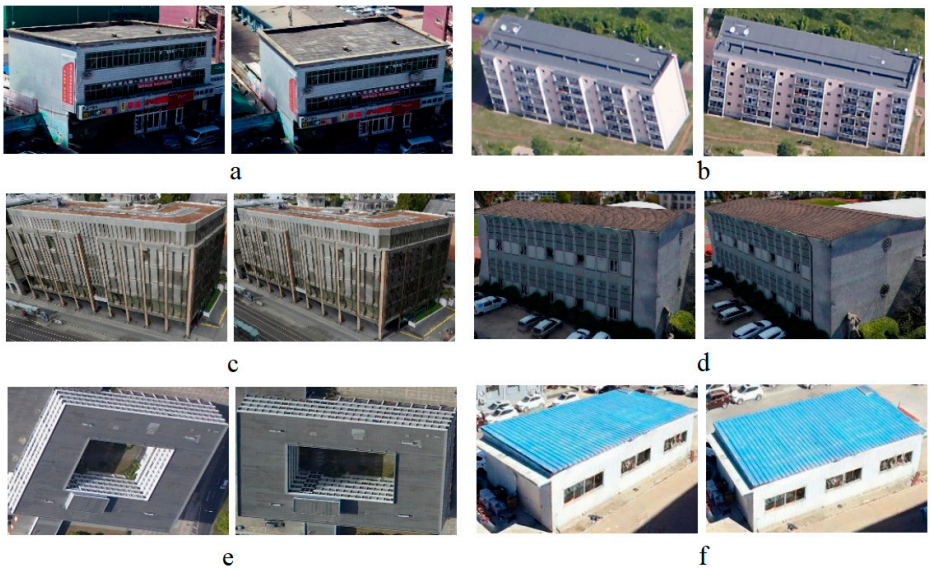

Figure 8. Test images. (**a**–**f**) six groups of large oblique stereo images with building scene.

3.4. Experimental Results and Analysis

Figures 9–14 show the matching results of six groups of data based on DFM, AffNet, SuperGlue, LoFTR, GlueStick, and our proposed method, respectively. Table 1 presents the quantitative experimental results of the six methods. Here, k_{ε_0} and α represent the number of correctly matched points and the correct rate of matching, respectively. ε_{RMSE} represents the RMSE of matching, and $\hat{D}$ represents the quality of spatial distribution of matching points. The optimal test results of each group of data in the table are represented in bold.

Table 1. The contrast of test results using six methods. The best values are highlighted in bold.

Test Data	Evaluation Metrics	Ours	DFM	AffNet	SuperGlue	LoFTR	GlueStick
(a)	k_{ε_0}/(Pair)	**2751**	1199	832	618	1695	336
	α/(%)	**0.80**	0.61	0.41	0.61	0.58	0.58
	ε_{RMSE}/(Pixel)	1.20	0.36	**0.35**	1.83	0.65	0.65
	$\hat{D}$	56.9	59.2	87.2	64.9	59.2	**33.7**
(b)	k_{ε_0}/(Pair)	**898**	31	82	537	520	259
	α/(%)	0.49	**0.53**	0.14	0.52	0.22	0.40
	ε_{RMSE}/(Pixel)	0.35	**0.18**	0.37	0.39	0.38	0.81
	$\hat{D}$	32.6	37.5	**17.13**	54.33	40.9	39.8
(c)	k_{ε_0}/(Pair)	**2751**	602	1100	618	1695	393
	α/(%)	**0.68**	0.34	0.33	0.61	0.58	0.47
	ε_{RMSE}/(Pixel)	0.99	**0.35**	0.37	1.24	2.07	0.71
	$\hat{D}$	45.5	27.7	45.9	**23.3**	46.5	32.6
(d)	k_{ε_0}/(Pair)	**2254**	237	296	330	1291	241
	α/(%)	**0.82**	0.40	0.20	0.50	0.60	0.52
	ε_{RMSE}/(Pixel)	0.86	**0.35**	0.36	2.9	0.57	0.67
	$\hat{D}$	41.1	32.0	28.2	26.8	43.3	**24.4**
(e)	k_{ε_0}/(Pair)	**1530**	56	125	196	1015	179
	α/(%)	**0.64**	0.43	0.22	0.24	0.46	0.48
	ε_{RMSE}/(Pixel)	0.99	**0.32**	0.36	1.88	1.06	0.69
	$\hat{D}$	40.1	24.1	**18.8**	23.5	43.3	23.0
(f)	k_{ε_0}/(Pair)	**2059**	915	974	273	1034	226
	α/(%)	**0.69**	0.46	0.42	0.39	0.47	0.44
	ε_{RMSE}/(Pixel)	0.49	0.37	**0.35**	1.88	1.46	0.75
	$\hat{D}$	**27.5**	83.3	28.2	32.5	43.3	34.6

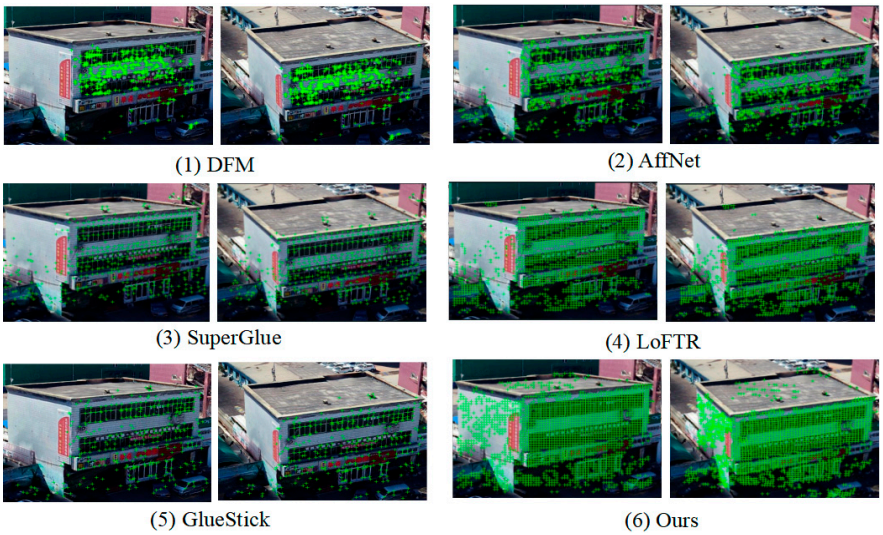

Figure 9. Test results of group images (a).

Figure 10. Test results of group images (b).

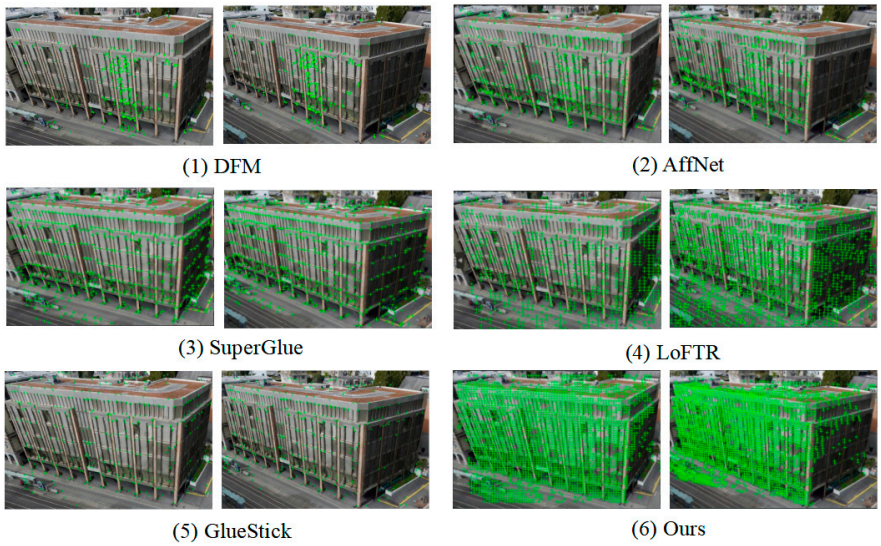

Figure 11. Test results of group images (c).

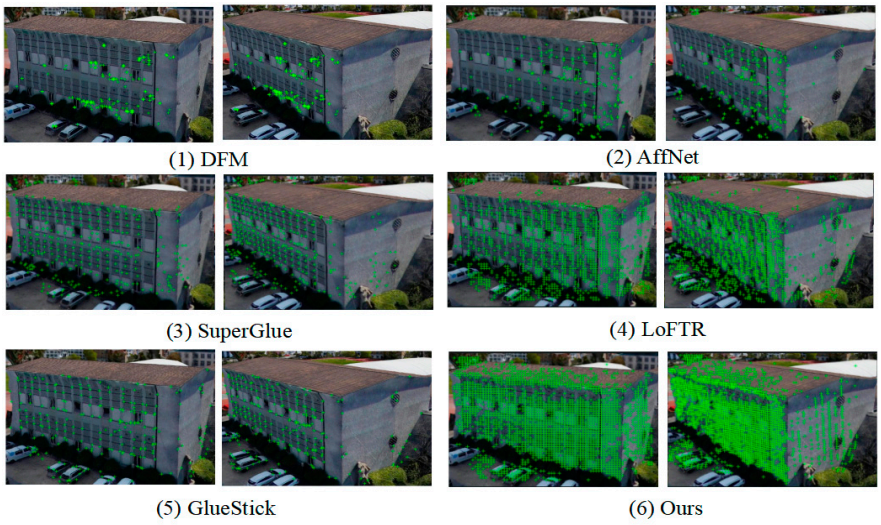

Figure 12. Test results of group images (d).

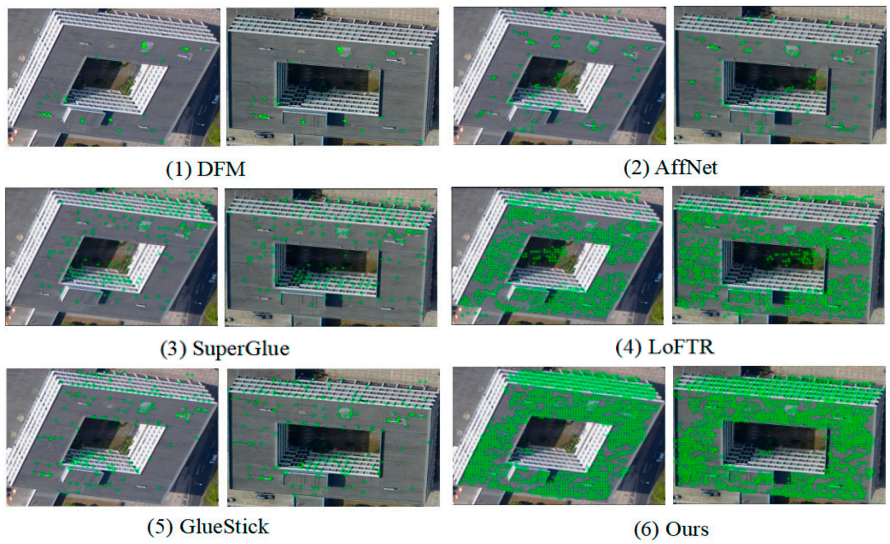

Figure 13. Test results of group images (e).

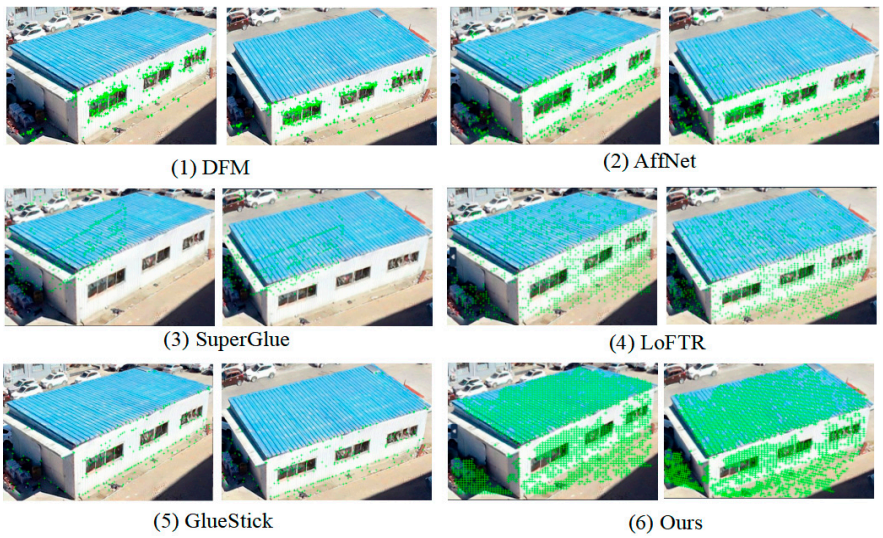

Figure 14. Test results of group images (f).

4. Discussion

(1) The proposed method has significant advantages in terms of the number of correctly matched points. Table 1 presents the quantitative experimental results for six groups of large viewpoint stereo images in architectural scenes that show the highest number of correctly matched points obtained using the proposed method. As shown in Figures 9–14, our proposed method can achieve accurate and dense matching results in each group of images, especially for matching a large number of corresponding points on the top and facades of buildings, which provides sufficient tie points for image orientation and three-dimensional (3D) reconstruction. The reasons are twofold. First, the multiplane segmentation and corresponding plane matching method proposed in this paper can transform the matching of complex 3D scenes into simple plane scene matching. Second, the LoFTR texture enhancement strategy introduced in this paper effectively improves the problem of weak texture on the top and facades of buildings, leading to accurate and dense matching results.

(2) According to the above experimental results, DFM has advantages in accuracy, but its effect on affine changes is poor. Compared with DFM, SuperGlue is more capable of handling large viewpoint affine transformations and single-texture regions; however, the number of matching points is much less than that obtained using our method. The LoFTR algorithm, which is based on the SuperGlue method, uses Transformer positional encoding and attention mechanisms to significantly enhance the texture features of building facades. GlueStick has not improved or even decreased in quantity compared to SuperGlue, but has improved in spatial distribution quality and matching accuracy. However, obtaining a sufficient number of matching points due to the influence of image distortion is challenging.

(3) Our method also demonstrates some advantages in terms of matching accuracy and precision. Table 1 shows that our method achieves high matching correctness rates for most of the test data (a, b, d–f), and sub-pixel matching precision for test data (b–f). The reasons behind this are as follows. First, our method performs individual matching for each planar scene and utilizes strict homography geometric transformations for distortion correction and constrained matching, effectively ensuring matching correctness and precision. Second, during the quasi-dense matching process, the proposed method first conducts coarse-level matching prediction and then refines the matches at a finer level, ensuring the accurate positioning of matching points.

(4) The proposed method exhibits good spatial distribution quality for the matching points. Figures 9–14 show that the distribution area of the matching points of our method in image space has significantly improved. Table 1 demonstrates that our method outperforms DFM and LoFTR algorithms in terms of the spatial distribution quality of matching points. Our method has good spatial distribution quality for matching points.

5. Conclusions

In this study, we propose a matching algorithm that combines image semantic segmentation and local feature enhancement networks for stereo images in complex scenes with significant viewpoint changes. The proposed algorithm first employs an automatic semantic segmentation method to extract the planes of different scenes. The LoFTR strategy is then used to enhance the weak texture features of each local plane, enabling accurate and dense feature matching. The experimental results demonstrate that the proposed method has advantages in terms of the number of correctly matched points, matching accuracy, matching precision, and spatial distribution quality of matched points. It is suitable for the dense matching of wide-baseline oblique stereo images. In future work, we plan to integrate a line feature matching algorithm to achieve more complementary feature matching along building structure edges. This can be applied to the fine-scale 3D reconstruction of urban building scenes.

Author Contributions: Conceptualization, J.G. and G.Y.; methodology, G.Y. and F.Z.; software, G.Y. and J.Z.; data curation, J.Z. and G.Y.; validation, G.Y., J.Z. and F.J.; formal analysis, G.Y. and J.G.; writing—original draft preparation, G.Y. and Q.F.; writing—review and editing, J.G., F.J. and X.R.; supervision, J.G. and F.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China with Project No. 42171435, the Shandong Provincial Natural Science Foundation with Project No. ZR2021MD006, the China Postdoctoral Science Foundation with Project No. 2023M732686, and the Undergraduate Education and Teaching Reform Foundation of Shandong Province with Project No. Z2021014. This work was also funded by the high quality graduate course of Shandong Province with Project No. SDYKC2022151.

Data Availability Statement: Data are available upon request due to restrictions.

Acknowledgments: The authors would like to thank Jiaming Sun and Zhuxuan Wu for providing their key algorithms.

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

References

1. Ge, Y.; Guo, B.; Zha, P.; Jiang, S.; Jiang, Z.; Li, D. 3D Reconstruction of Ancient Buildings Using UAV Images and Neural Radiation Field with Depth Supervision. *Remote Sens.* **2024**, *16*, 473. [CrossRef]
2. Yao, G.B.; Yilmaz, A.; Meng, F.; Zhang, L. Review of wide-baseline stereo image matching based on deep learning. *Remote Sens.* **2021**, *13*, 3247. [CrossRef]
3. Ji, S.; Luo, C.; Liu, J. A Review of Dense Stereo Image Matching Methods Based on Deep Learning. *Geomat. Inf. Sci. Wuhan Univ.* **2021**, *46*, 193–202. [CrossRef]
4. Liu, J.; Ji, S.P. Deep learning based dense matching for aerial remote sensing images. *Acta Geod. Cartogr. Sin.* **2019**, *48*, 1141–1150. [CrossRef]
5. Luo, S.D.; Chen, H.B. Stereo matching algorithm of adaptive window based on region growing. *J. Cent. South Univ. Technol.* **2005**, *36*, 1042–1047.
6. Fritz, C.O.; Morris, P.E.; Richler, J.J. Effect size estimates: Current use, calculations, and interpretation. *Exp. Psychol. Gen.* **2012**, *141*, 2–18. [CrossRef]
7. Yang, H.; Zhang, S.; Zhang, Q. Least Squares Matching Methods for Wide Base-line Stereo Images Based on SIFT Features. *Acta Geod. Cartogr. Sin.* **2010**, *39*, 187–194. [CrossRef]
8. David, G.L. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
9. Yang, H.; Zhang, S.; Wang, L. Robust and precise registration of oblique images based on scale-invariant feature transformation algorithm. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 783–787. [CrossRef]
10. Zhang, Q.; Wang, Y.; Wang, L. Registration of images with affine geometric distortion based on maximally stable extremal regions and phase congruency. *Image Vis. Comput.* **2015**, *36*, 23–39. [CrossRef]
11. Xiao, X.W.; Guo, B.X.; Li, D.R.; Zhao, X.A. Quick and affine invariance matching method for oblique images. *Acta Geod. Cartogr. Sin.* **2015**, *44*, 414–442. [CrossRef]
12. Morel, J.-M.; Yu, G. Asift: A new framework for fully affine invariant image comparison. *SIAM J. Imaging Sci.* **2009**, *2*, 438–469. [CrossRef]
13. Tian, Y.R.; Fan, B.; Wu, F.C. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 661–669. [CrossRef]
14. Mishchuk, A.; Mishkin, D.; Radenovic, F. Working hard to know your neighbor's margins: Local descriptor learning loss. *Adv. Neural Inf. Process. Syst.* **2017**, *1*, 4826–4837. [CrossRef]
15. Zhang, C.; Yao, G.; Man, X.; Huang, P.; Zhang, L.; Ai, H. Affine invariant feature matching of oblique images based on multi-branch network. *Acta Geod. Cartogr. Sin.* **2021**, *50*, 641–651. [CrossRef]
16. Mishkin, D.; Radenovic, F.; Matas, J. Repeatability is not enough: Learning affine regions via discriminability. In Proceedings of the 2018 Computer Vision, Munich, Germany, 8–14 September 2018; pp. 287–304. [CrossRef]
17. Revaud, J.; Weinzaepfel, P.; De Souza, C.; Pion, N.; Csurka, G.; Cabon, Y.; Humenberger, M. R2D2: Repeatable and reliable detector and descriptor. *arXiv* **2019**, arXiv:1906.06195.
18. Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperGlue: Learning feature matching with graph neural networks. In Proceedings of the IEEE 2020 Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020. [CrossRef]
19. Sun, J.; Shen, Z.; Wang, Y.; Bao, H.; Zhou, X. LoFTR: Detector-free local feature matching with transformers. *arXiv* **2021**, arXiv:2104.00680.
20. Efe, U.; Ince, K.; Alatan, A. DFM: A Performance Baseline for Deep Feature Matching. *arXiv* **2021**, arXiv:2106.07791.
21. Pautrat, R.; Suárez, I.; Yu, Y.; Pollefeys, M.; Larsson, V. Gluestick: Robust image matching by sticking points and lines together. *arXiv* **2023**, arXiv:2304.02008.

22. Yao, G.B.; Yilmaz, A.; Zhang, L.; Meng, F.; Ai, H.B.; Jin, F.X. Matching large baseline oblique stereo images using an end-to-end convolutional neural network. *Remote Sens.* **2021**, *13*, 274. [CrossRef]
23. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556. [CrossRef]
24. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015. [CrossRef]
25. Wu, Z.; Han, X.; Lin, Y.L.; Uzunbas, M.G.; Goldstein, T.; Lim, S.N.; Davis, L.S. Dcan: Dual channel-wise alignment networks for unsupervised scene adaptation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 518–534. [CrossRef]
26. Zhu, Q.; Wu, B.; Xu, Z.X. Seed point selection method for triangle constrained image matching propagation. *IEEE Geosci. Remote Sens. Lett.* **2006**, *3*, 207–211. [CrossRef]
27. Yao, G.; Zhang, J.; Gong, J.; Jin, F. Automatic Production of Deep Learning Benchmark Dataset for Affine-Invariant Feature Matching. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 33. [CrossRef]