

Article

Joint Deep Learning and Information Propagation for Fast 3D City Modeling

Yang Dong, Jiaxuan Song, Dazhao Fan *, Song Ji and Rong Lei

Institute of Surveying and Mapping, Information Engineering University, Zhengzhou 450001, China

* Correspondence: fdzcehui@163.com

Abstract: In the field of geoinformation science, multiview, image-based 3D city modeling has developed rapidly, and image depth estimation is an important step in it. To address the problems of the poor adaptability of training models of existing neural network methods and the long reconstruction time of traditional geometric methods, we propose a general depth estimation method for fast 3D city modeling that combines prior knowledge and information propagation. First, the original image is downsampled and input into the neural network to predict the initial depth value. Then, depth plane fitting and joint optimization are combined with the superpixel information and the superpixel optimized depth value is upsampled to the original resolution. Finally, the depth information propagation is checked pixel-by-pixel to obtain the final depth estimate. Experiments were conducted using multiple image datasets taken from actual indoor and outdoor scenes. Our method was compared and analyzed with a variety of existing widely used methods. The experimental results show that our method maintains high reconstruction accuracy and a fast reconstruction speed, and it achieves better performance. This paper offers a framework to integrate neural networks and traditional geometric methods, which provide a new approach for obtaining geographic information and fast 3D city modeling.

Keywords: 3D city modeling; multiview image; deep learning; information propagation; depth estimation; superpixel



Citation: Dong, Y.; Song, J.; Fan, D.; Ji, S.; Lei, R. Joint Deep Learning and Information Propagation for Fast 3D City Modeling. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 150. <https://doi.org/10.3390/ijgi12040150>

Academic Editor: Wolfgang Kainz

Received: 19 December 2022

Revised: 13 March 2023

Accepted: 30 March 2023

Published: 2 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Fast 3D city modeling is a fundamental problem in the field of geoinformation science and has a wide range of applications in spatial data management, geospatial artificial intelligence, and advanced geospatial applications. The traditional 3D modeling method for multiview images restores the dense point-to-point correspondence through cost calculation, cost aggregation, disparity calculation, and disparity refinement (such as SGM [1–5] and PatchMatch [6–10]). These methods can solve the problem of 3D modeling in specific scenes through hand-crafted similarity metrics. However, there are still certain limitations in regional scenes, such as weak texture and specular reflection. The accuracy and completeness of the modeling still need to be improved [11]. In recent years, with the successful application of deep convolutional neural network technology in multiple fields, a variety of deep-learning-based methods have also appeared in the area of 3D modeling of multiview images (such as SurfaceNet [12] and MVSNet [13]). Methods based on deep learning can introduce prior knowledge to better solve problems that cannot be solved by traditional methods, such as weak texture and specular reflection. In recent years, tests on multiple public datasets have shown that deep-learning-based methods have gradually surpassed traditional methods and can better deal with the problem of 3D modeling of multiview images in extreme situations [14]. However, methods based on deep learning require a large amount of accurate prior data for training, the modeling accuracy is poor in some untrained scenes, and the applicable scope of the training model shows certain limitations [15].

Traditional 3D modeling algorithms for multiview images can be divided into point-cloud-based methods, voxel-based methods, and depth-map-based methods [16–18]. The

point-cloud-based method [16] directly optimizes the 3D point cloud and obtains a dense 3D point cloud through a seed point stepwise propagation strategy, such as PMVS and its derivative improved algorithm [19,20]. However, this method requires sequential propagation based on seed points. It is difficult to parallelize, consumes too much memory, and its running time is too long [21]. The voxel-based method [17] divides the 3D space into regular grids and then estimates whether the voxels are part of the scene surface one by one [22–25]. The final accuracy of this method depends on the division density of the regular grid. It requires substantial memory consumption, which also limits its final accuracy [26]. The depth-map-based method [18] processes only one reference image and matches multiple images at a time, which can integrate the existing binocular stereo-matching algorithms well and is relatively flexible [27–32]. At the same time, the depth-map-based method can be easily fused and converted into point clouds or voxel reconstructions. Recently, PatchMatch-based 3D modeling methods [33–37] have shown great promise in depth estimation [38]. In addition, since the depth estimation of these methods can be implemented in parallel on GPUs, these methods can greatly reduce the computational cost of 3D modeling. Among them, Xu et al. (2022) designed a multi-scale geometric consistency guided and planar prior assisted multi-view stereo method (ACMMP), which achieves state-of-the-art performance [38]. However, most of these methods are sensitive to textureless areas, reflective surfaces, and repetitive patterns.

To overcome the above difficulties, some researchers have conducted research on multiview image reconstruction algorithms based on deep convolutional neural networks [12,14,39–42]. Hartmann et al. (2017) proposed replacing the similarity cost calculation in traditional algorithms with learning the similarity between multiple facets to improve the final matching accuracy [39]. Ji et al. (2017) suggested using SurfaceNet to map 2D image luminosity information to a 3D voxel space, constructing the 3D cost by selecting multiple pairs of view voxels, and using a 3D convolutional network to infer surface voxels [12]. Kar et al. (2017) proposed LSM, which constructs the 3D cost through microcamera projection and a recursive voxel fusion network and uses a 3D convolutional network to determine whether voxels belong to the scene surface [40]. Yao et al. (2018) proposed MVSNet, which constructs the 3D cost on the reference camera frustum through differentiable homography, uses a 3D convolutional network for depth inference, and calculates the depth residual with the reference image to optimize the depth map output [13]. To further optimize the memory consumption and calculation time of MVSNet, Gu et al. (2020) proposed CasMVSNet, which uses the cascaded cost volume to perform depth map inference from coarse to fine [41]. Fangjinhua et al. (2021) proposed PatchMatchNet [42] and Xu et al. (2022) proposed PVSNet [14], which introduced visibility estimation into their networks and helped these methods obtain reliable reconstruction results on datasets with wide baselines. However, most of these methods cannot achieve scalable high-resolution depth map estimation; the modeling accuracy is poor in some untrained scenes [43].

This paper aims to combine the advantages of deep learning methods and traditional methods to compensate for the limitations of a single deep learning training model and the limitations of a single traditional method. This paper uses the depth estimation result obtained by the deep learning method as the initial value, combined with superpixel segmentation to fit the depth plane, eliminates excessive depth inference errors, optimizes the deep linear surface, and then performs depth propagation pixel-by-pixel. By optimizing the deep nonlinear surface, a relatively complete and accurate depth estimation result for 3D modeling can finally be obtained. Using the deep convolutional neural network learning method can yield a better initial value of depth estimation for non-Lambertian regions and weak texture regions. By combining superpixel information for depth plane propagation and pixel-by-pixel iterative depth propagation, it is possible to optimize the depth estimation results of deep learning from coarse to fine and to compensate for the limitations of deep learning methods. Experiments with multiple real datasets show that the method in this paper can effectively combine the advantages of deep learning methods and traditional methods to obtain complementary advantages. It can ensure the accuracy of modeling while improving the completeness, achieving fast 3D city modeling.

The remainder of this paper is organized as follows: Section 2 introduces the ideas and overall process of our method. Section 3 uses multiview image datasets for experimental verification and compares our method with current mainstream algorithms. Section 4 discusses the performance of our method in detail. Finally, Section 5 summarizes this paper and future work.

2. Materials and Methods

2.1. Acquisition of the Initial Depth Information Based on Existing Neural Network

This paper uses an existing neural network, CasMVSNet, to quickly infer the initial depth map of the reference image, which uses cascaded cost volume calculations, can save memory in the calculation process, and can process relatively high-resolution images on a dedicated graphics server; however, the resolution of images that can be processed is still very limited, which limits the application range of methods based on convolutional neural networks. Therefore, to be able to further process real high-resolution images, we use downsampled multiview images in the neural network for the initial depth estimation, and then in subsequent processing, the inferred depth map is upsampled to the original resolution for optimization processing. Since the neural network method provides only the initial depth and the image depth is smoothed in subsequent processing, image resolution reduction processing within a certain scale range will not affect the final depth inference quality. Therefore, we can process high-resolution images with low memory consumption, avoiding the memory limitations of traditional deep convolutional neural network methods and greatly increasing the value in practical application processing.

2.2. Optimization of Superpixel Depth Information Propagation

Depth inference based on a convolutional neural network has good estimation accuracy in a priori scenes. Although the approximate depth can be inferred in nontraining scenes, its accuracy is severely degraded, as shown in Figure 1b, and the surface exhibits nonlinear fluctuations. A smooth surface in a scene often presents a certain local color consistency in the original image. Therefore, the original image color information can be combined with the neural network to infer the optimization of the initial depth. A superpixel is a collection of local image pixels with color consistency, and the essence of the convolutional neural network method is to learn a priori multiview color consistency for deep inference, so the depth inference based on multiview color consistency should have a certain linear continuity in the superpixel range.

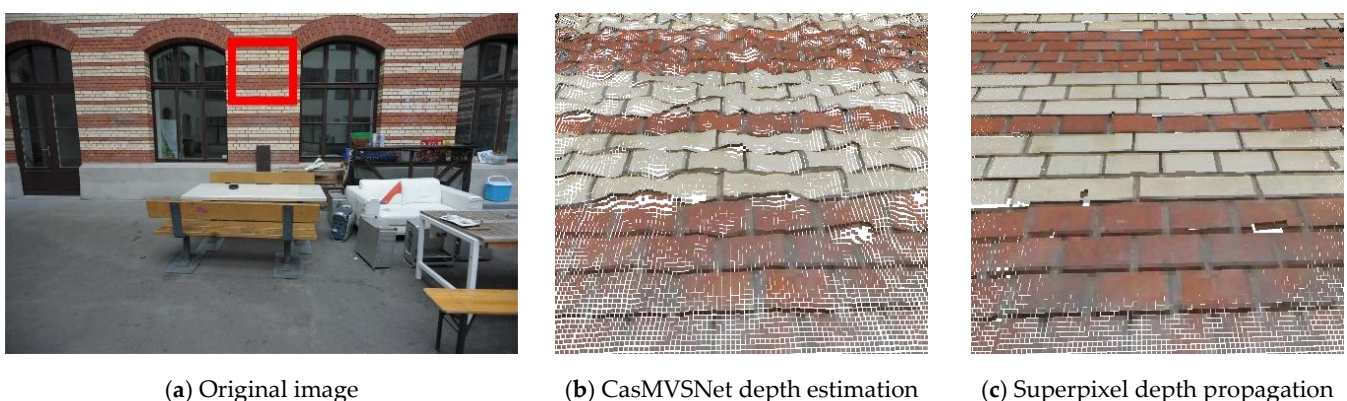


Figure 1. Superpixel information propagation.

Using superpixels as the processing unit to optimize the depth map can effectively filter out noise, repair inference loopholes, and improve the accuracy and completeness of the depth map. Taking the superpixels extracted from the reference image as the processing unit, we perform plane fitting filtering inside the superpixels and propagate the plane

fitting parameters to adjacent superpixels, which can iteratively complete the loophole completion and accuracy optimization of the initial depth map, as shown in Figure 1c.

For each superpixel set $\mathbf{S}$ in the reference image $\mathbf{X}_1$, the corresponding depth $d(p)$ is fitted using a linear plane

$$d(p) = A_s p_x + B_s p_y + C_s \quad (1)$$

where (p_x, p_y) is the image coordinate of the image point p ; and $\zeta_s = (A_s, B_s, C_s)$ is the linear plane fitting parameter, which can be determined by the RANSAC method [44]. We solve the plane fitting parameters of all superpixels in the reference image $\mathbf{X}_1$ and calculate the overall cost of the superpixels using the color consistency between multiview images

$$L_s^{\zeta_s} = \frac{1}{(N-1) \times n_s} \times \sum_{p \in \mathbf{S}} \sum_{i=2}^N \|c_{p_1} - c_{p_i}\|_2 \quad (2)$$

where p_1 is an image point in the superpixel $\mathbf{S}$ of the reference image; p_i is the corresponding image point of another view, which can be calculated through geometric relationship projection between the images; c_{p_i} is the color value of the corresponding viewing pixel; N is the number of multiview images; n_s is the number of pixels in the superpixel set $\mathbf{S}$; and $\|\cdot\|_2$ is the L_2 distance.

Using a similar idea to that of PatchMatch, the propagation of the superpixel plane fitting parameters is achieved. For the current superpixel $\mathbf{S}$ and its neighborhood superpixel $\mathbf{S}'$, we calculate the overall cost $L_s^{\zeta_s}$ and $L_{s'}^{\zeta_{s'}}$ of the current fitting parameter and the neighborhood plane fitting parameter on the current superpixel and take the parameter with the smallest overall cost as the depth plane of the current superpixel fitting parameters. In this way, iterative forward–backward cross-propagation can effectively eliminate the initial depth values with excessive errors and perform depth optimization based on superpixels. Because the prior superpixel segmentation information may have errors, after completing the optimization of superpixel information propagation in the neighborhood, superpixel segmentation optimization can be integrated to achieve overall neighborhood superpixel depth smoothing optimization and further optimize the accuracy of the overall depth map.

2.3. Linear Plane Optimization of Depth Information Combined with Superpixel Optimization

The initial superpixel segmentation may have edge misjudgments, which can be addressed by combining superpixel segmentation information and image depth information to obtain better superpixel segmentation edge- and image-depth inference results [45]. Therefore, in the depth estimation of multiview images, considering the resegmentation optimization of superpixels, the superpixel cost $L_s^{\zeta_s}$ of the reference image can be optimized to obtain the overall cost of the joint color $L_{col}(\cdot)$, position $L_{pos}(\cdot)$, depth $L_{depth}(\cdot)$, and boundary $L_{bou}(\cdot)$, as shown in Equation (3)

$$L(\mathbf{S}, \zeta, \mathbf{f}, \mathbf{o}, \mathbf{d}, \mathbf{X}_1) = \sum_{p \in \mathbf{S}} \left[L_{col}(p, c_{s_p}) + \lambda_{pos} L_{pos}(p, \mu_{s_p}) + \lambda_{depth} L_{depth}(p, \zeta_{s_p}, f_p) \right] + \lambda_{smo} \sum_{\{i,j\} \in N_{seg}} L_{smo}(\zeta_i, \zeta_j, o_{ij}) + \lambda_{bou} \sum_{\{p,q\} \in N_s} L_{bou}(s_p, s_q) \quad (3)$$

where $\mathbf{S}$ is the set of image points within the superpixel; ζ is the superpixel depth plane fitting parameter; $\mathbf{f} = \{f_p \in \{0, 1\} | p \in \mathbf{S}\}$ indicates whether the image point is an outlier in the depth linear plane fitting; $\mathbf{o}$ is the label of the superpixel edge image point—the edge image point labels are classified into abrupt edges, continuous edges, and coplanar edges based on the depth information; $\mathbf{d}$ is the depth value of the image point within the superpixel; $\mathbf{X}_1$ is the reference image; s_p is the superpixel label of the image point p ; μ is the average position of the image point within the superpixel; c is the average color of the image point within the superpixel; N_{seg} is the set of neighboring superpixels; and N_s is the set of neighboring image points of superpixel edge points.

$$L_{col}(p, c_{s_p}) = \|\mathbf{X}_1(p) - c_{s_p}\|_2 \quad (4)$$

is the superpixel color consistency cost, which is used to evaluate the color consistency of image points within the superpixel.

$$L_{pos}(p, \mu_{s_p}) = \|p - \mu_{s_p}\|_2 \quad (5)$$

is the superpixel local position cost, which is used to limit the size and shape of the superpixel.

$$L_{depth}(p, \zeta_{s_p}, f_p) = \begin{cases} |d(p) - \hat{d}(p, \zeta_{s_p})| & \text{if } f_p = 0 \\ \lambda_d & \text{if } f_p = 1 \end{cases} \quad (6)$$

is the superpixel depth consistency cost, which is used to evaluate the consistency of the image point depth within the superpixel.

$$L_{smo}(\zeta_i, \zeta_j, o_{ij}) = \begin{cases} \lambda_{occ} & \text{if } o_{ij} = occ \\ \frac{1}{|B_{ij}|} \sum_{p \in B_{ij}} |\hat{d}(p, \zeta_i) - \hat{d}(p, \zeta_j)| + \lambda_{hinge} & \text{if } o_{ij} = hinge \\ \frac{1}{|S_i \cup S_j|} \sum_{p \in S_i \cup S_j} |\hat{d}(p, \zeta_i) - \hat{d}(p, \zeta_j)| & \text{if } o_{ij} = planar \end{cases} \quad (7)$$

is the smoothing cost of neighboring superpixels, which is used to incentivize superpixels of similar depth to merge and optimize the superpixel segmentation results: $o_{ij} = occ$ denotes a mutated edge between neighboring superpixels; $o_{ij} = hinge$ denotes a continuous edge between neighboring superpixels; $o_{ij} = planar$ denotes a coplanar edge between neighboring superpixels; $\lambda_{occ} > \lambda_{hinge} > 0$ denotes that the cost of a mutated edge is the largest, the cost of a continuous edge is the second largest, and the cost of a coplanar edge is the smallest; and B_{ij} is the set of edge pixels between adjacent superpixels i, j .

$$L_{bou}(s_p, s_q) = \begin{cases} 0 & \text{if } s_p = s_q \\ 1 & \text{otherwise} \end{cases} \quad (8)$$

is the superpixel edge consistency cost for limiting the superpixel edge shape, incentivizing it to converge to simple shapes, and avoiding complex image oversegmentation.

To determine the superpixel integration cost $L(\mathbf{S}, \zeta, \mathbf{f}, \mathbf{o}, \mathbf{d}, \mathbf{X}_1)$, a multivariate joint solution can be used for step-by-step iterative optimization [45]. First, the optimized depth map and segmented superpixels after the propagation of superpixel information are used to initialize the acquisition of $\mathbf{S}, \mu, c, \zeta, \mathbf{o}, \mathbf{f}$. The joint solution is determined for the superpixel set $\mathbf{S}$, the outlier set $\mathbf{f}$, the average position μ , and the average color c . Then, the edge set $\mathbf{o}$ and the depth plane fitting parameter ζ are solved step by step, and finally, the overall optimization is completed by circular iteration.

2.4. Deep Optimization Based on Pixel-by-Pixel Information Propagation

Depth map optimization based on superpixel information propagation can effectively filter out depths with excessive errors and effectively fill an occluded region. However, due to the linear plane fitting inside the superpixel, it does not cope well with nonlinear surfaces such as spherical surfaces, and the imprecision of superpixel segmentation can introduce further errors in depth map optimization. Therefore, a pixel-by-pixel PatchMatch matching check can be used for depth maps that have been superpixel planar-fitted. For the related PatchMatch method, we can use the method involved in COLMAP [46], or the optimized ACMMP series method [38]. Additionally, since the depth map has been optimized and most depths have only small estimation errors, the set of selected candidate depths and normals can be simplified in these methods. Using the simplified cost and candidate set, the depth map can be optimized again in only a few iterations to obtain the final depth map estimate.

2.5. Multiview Image Depth Estimation Method Implementation

The general flow of the depth estimation algorithm designed in this paper for fast 3D city modeling is shown in Figure 2. Firstly, for the input multiview image, downsampling is performed, and the initial depth map is inferred using a deep convolutional neural network, while the initial superpixel extraction of the reference image is performed; then, based on the superpixel extraction results, neighborhood iterative propagation of the depth fitting plane is performed to filter out large estimation errors in the depth map; subsequently, the results of superpixel extraction and depth map superpixel propagation are combined to perform superpixel segmentation; finally, the depth map is upsampled, and the depth information is propagated pixel by pixel with the original multiview image to optimize the nonlinear surface of the depth map to obtain the final accurate depth estimation.

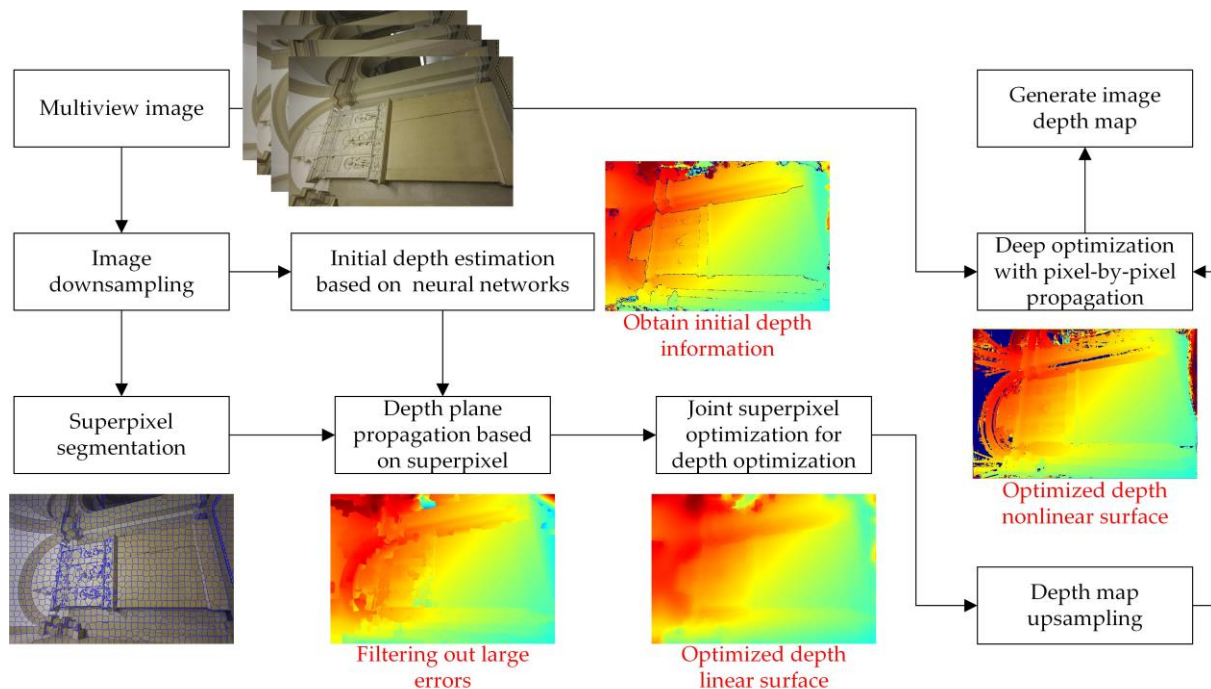


Figure 2. Multiview image depth estimation method.

3. Results

3.1. Setup

The ETH3D multiview image dataset [43,47,48] is used for experimental validation. The ETH3D multiview image dataset contains multiview images of several indoor and outdoor scenes; each image in the dataset has known positional parameters, and each scene has been scanned by high-precision LiDAR to obtain real scene point cloud data. Thirteen high-resolution scene image datasets with real point-cloud data published in ETH3D are used for the experiments. In the experiments, due to the limitation of the convolutional neural network, the image size in the ETH3D high-resolution dataset is uniformly cropped to a multiple of 32 (6144×4096).

In the experiment, we use the CasMVSNet neural network for initial depth estimation, using the COLMAP and ACMMP methods for pixel-by-pixel optimization. This resulted in two experimental methods, which we called NPMVS-COLMAP and NPMVS-ACMMP. Multiple methods are designed for comparison in the experiments. First, for the NPMVS-COLMAP method, a comparison experiment based on depth map estimation is conducted with the CasMVSNet method and COLMAP method without the geometric constraints involved in the paper, which is used to analyze the effectiveness of our method. Then, the depth map estimated by NPMVS-COLMAP and NPMVS-ACMMP is fused and reconstructed as a 3D scene point cloud and compared with CasMVSNet [41], COLMAP

without geometric constraints, COLMAP with geometric constraints [46], ACMMP [38], PLC [49], HPMVS [50], MVE [51], and OpenMVS [52]. In the experiments, the downsampling magnification in our method is set to eight times; the superpixel extraction algorithm uses the GPU-accelerated simple linear iterative clustering algorithm (gSLIC) [53,54], and the number of superpixels extracted is set to 1000; the RANSAC method is used in the superpixel depth information plane parameter fitting, and the number of superpixel-based plane parameter propagation iterations is set to 2, 10 for joint superpixel-image depth optimization. The convolutional neural network parameters in CasMVSNet are adopted directly from the network parameters provided in reference [41], which are obtained by using the DTU training dataset [55] for training. NPMVS–COLMAP, CasMVSNet, and COLMAP depth map fusion without geometric constraints are fused using the COLMAP method, where the minimum number of fused pixels corresponding to the 3D points is set to two and the other fusion parameters are used as default parameters. NPMVS–ACMMP and ACMMP depth map fusion with geometric constraints are fused using the ACMMP method. In the CasMVSNet method, to contrast it with our method, the depth map is inferred by downsampling the original image by the same multiple and then upsampling to the original resolution for 3D point-cloud fusion. The COLMAP, PLC, HPMVS, MVE, and OpenMVS methods with geometric constraints all set the maximum image resolution processed to the original resolution, and all other parameters are set to the default parameters provided in the reference.

The main comparison metrics used in the depth map comparison experiments are the mean error (ME), root-mean-square error (RMSE), a90 error, and time consumption [48]. The main comparison metrics used in the 3D modeling comparison experiments are reconstruction completeness, accuracy, F1 score, and overall reconstruction time [48]. The reconstruction completeness r characterizes the integrity of the reconstructed scene within the distance threshold t ; the reconstruction accuracy p characterizes the accuracy of the reconstructed scene within the distance threshold t ; and the F1 score is the combined calculated value of the reconstruction accuracy and reconstruction completeness, defined as $2 \cdot (p \cdot r) / (p + r)$. The experimental platform configuration is an Intel Core i9-9900K CPU with an NVIDIA RTX 2080 GPU and 64 GB memory.

3.2. Results

3.2.1. Experimental Results of Depth Estimation

Considering that the CasMVSNet neural network cannot handle large-sized images, the images are downsampled four times to form an image dataset that the neural network can process for a fair comparison. Using the downsampled image set, the accuracy and time consumption statistics of the depth maps generated by NPMVS–COLMAP, CasMVSNet, and COLMAP are compared, as shown in Figure 3. Figure 3a–c shows the mean error, root-mean-square error, and a90 error, respectively; the corresponding error intervals of all experimental images are counted, the horizontal axis indicates the size of the error interval, the vertical axis indicates the number of images within the corresponding error interval, yellow indicates our method, green indicates the CasMVSNet method, and purple indicates the COLMAP method. Figure 3d shows the average time consumption of each method; the average time consumption of the image depth map within each dataset is counted, the horizontal axis shows the name of each dataset, the vertical axis is the average time consumption of depth map estimation for the corresponding dataset, the gray squares indicate our method, the red dots indicate the CasMVSNet method, and the blue triangles indicate the COLMAP method.

From the experimental results, it can be seen that the accuracy of the depth map estimated by our method is better than that of both the COLMAP and CasMVSNet methods, and the time consumption is less than that of the CasMVSNet method and better than the COLMAP method. Overall, our method can effectively improve the depth estimation accuracy of the original CasMVSNet method and expand the applicability of the deep neural network method while effectively reducing the overall processing time of the

original COLMAP method and improving the completeness of the depth estimation of the traditional method. The depth estimation based on our method effectively combines the advantages of the neural network method and the traditional geometric method, and it effectively improves the completeness and computational speed of depth estimation while guaranteeing accuracy, which integrally improves the comprehensive performance of the algorithm.

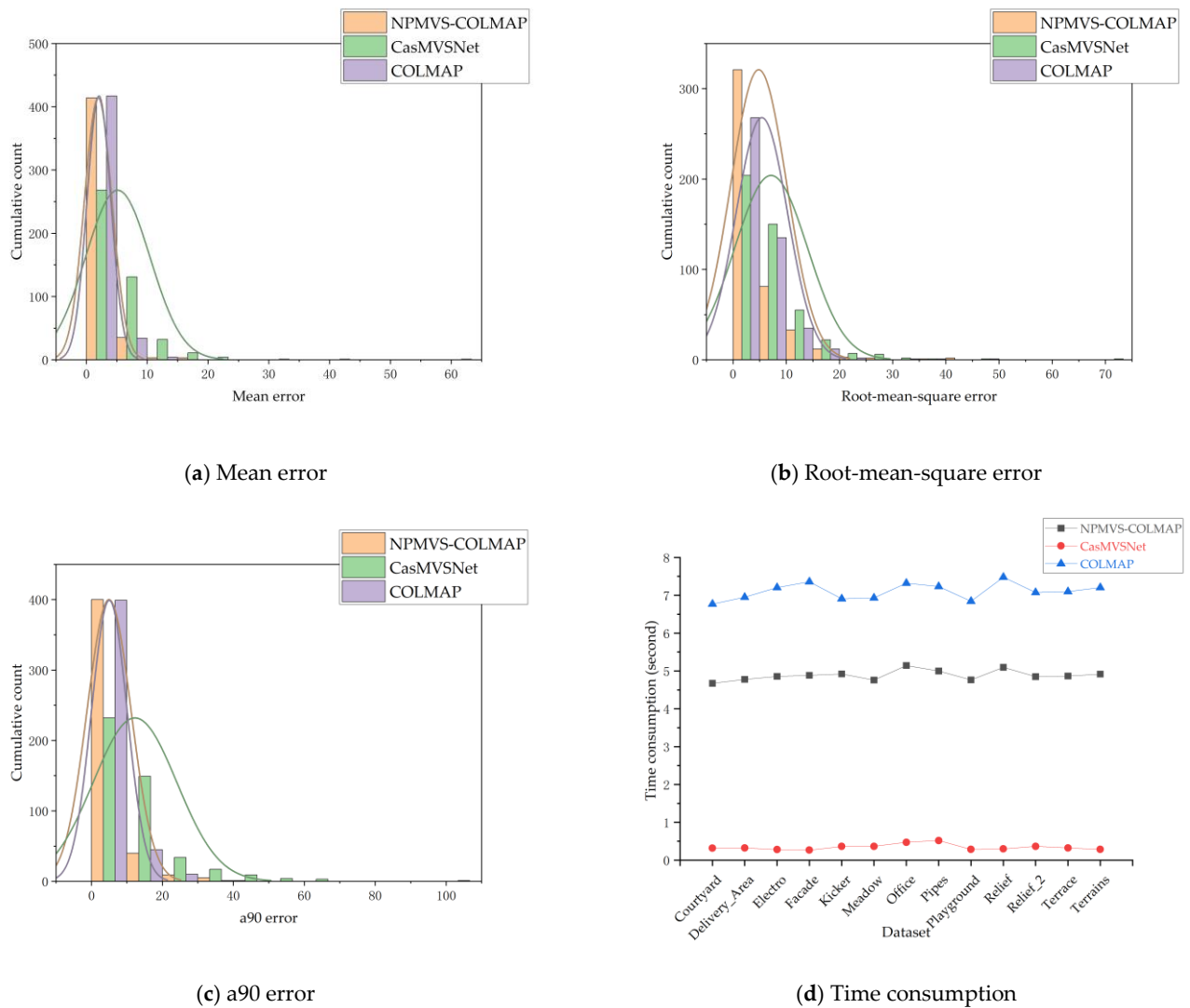


Figure 3. Experimental results of the depth map comparison (lower is better).

3.2.2. Experimental Results for 3D Modeling

The results of the comparison experiments are shown in Tables 1–4 and Figure 4. Tables 1–4 are statistics of the reconstruction time consumption, accuracy, completeness, and F1 values under each dataset at the distance threshold $t = 0.1$. Figure 4 shows the visualization corresponding to Tables 1 and 4, where the horizontal axis is the time consumption and the vertical axis is the F1 value; the smaller the time consumption is, the better, and the larger the F1 value is, the better, so the closer the position of a point in the figure is to the upper left corner, the better.

Table 1. Comparison of time (s) (lower is better).

Method	Avg.	Indoor					Outdoor							
		Deli.	Kick.	Offi.	Pipes	Relief	Relief.	Terra.	Cour.	Elec.	Facade	Mead.	Play.	Terrace
CasMVSNet	1402.73	1365.37	846.29	633.50	304.26	1677.70	1521.94	1996.85	1290.27	1111.33	5401.17	292.76	1046.92	747.20
COLMAP	3026.71	3446.42	2074.49	1655.43	962.18	3117.86	2956.17	3774.24	2972.35	3402.33	9414.43	951.66	2728.46	1891.15
COLMAP-GEO	7698.31	8052.55	5132.84	4516.79	2555.34	9294.34	9157.88	11,965.90	6555.20	8116.39	19,570.60	2431.57	6371.15	6357.46
ACMMP	2059.67	2790.42	2098.65	1275.53	536.13	2189.36	2042.35	3240.24	2713.84	3094.78	-	756.23	2696.27	1282.25
PLC	14,119.26	14,841.40	11,282.70	9123.19	4942.87	14,851.40	12,981.40	19,964.30	12,741.80	15,277.70	33,320.80	5194.15	18,613.90	10,414.80
HPMVS	2146.69	918.91	300.09	157.98	87.38	8457.83	6242.12	4455.22	1284.50	702.88	2475.84	127.21	218.60	2478.46
MVE	9641.91	10,037.30	3211.70	919.13	837.05	14,792.30	14,060.00	7257.59	13,933.70	6519.10	39,942.40	719.75	5251.61	7863.16
OpenMVS	7010.35	8782.24	4898.89	3246.76	1960.22	6820.72	7151.68	9224.69	9196.07	8209.61	18,647.80	2107.62	6225.12	4663.11
NPMVS–COLMAP (ours)	2635.92	2768.04	2139.06	1384.57	722.16	2577.81	2546.97	3027.74	2669.81	2798.57	8840.57	847.28	2353.23	1591.15
NPMVS–ACMMP (ours)	521.18	660.51	451.13	356.19	203.21	480.40	484.50	635.52	555.80	662.46	1145.96	239.10	548.03	352.48

Table 2. Comparison of accuracy (higher is better).

Method	Avg.	Indoor					Outdoor							
		Deli.	Kick.	Offi.	Pipes	Relief	Relief.	Terra.	Cour.	Elec.	Facade	Mead.	Play.	Terrace
CasMVSNet	76.63%	81.80%	76.59%	87.73%	85.17%	86.59%	80.67%	75.11%	83.41%	91.54%	73.77%	55.80%	56.74%	61.33%
COLMAP	95.58%	96.50%	93.60%	94.27%	96.22%	97.65%	96.78%	94.77%	97.46%	97.62%	95.30%	92.25%	94.35%	95.80%
COLMAP-GEO	98.79%	99.38%	99.42%	99.55%	99.53%	99.61%	99.53%	98.88%	99.06%	99.62%	98.54%	93.28%	98.66%	99.22%
ACMMP	97.72%	98.75%	92.96%	94.64%	98.88%	98.43%	98.34%	98.35%	99.55%	99.01%	-	96.56%	98.27%	98.93%
PLC	98.02%	99.40%	97.40%	95.87%	97.99%	99.61%	99.55%	98.87%	98.93%	99.73%	98.34%	90.20%	98.83%	99.58%
HPMVS	96.22%	95.08%	96.34%	98.73%	95.42%	98.59%	98.86%	96.52%	94.63%	97.66%	95.65%	88.27%	97.86%	97.29%
MVE	94.61%	86.23%	94.49%	96.81%	95.87%	98.10%	97.50%	93.58%	96.86%	94.37%	96.26%	92.04%	96.76%	91.03%
OpenMVS	97.45%	98.20%	97.35%	97.59%	99.11%	98.27%	98.16%	97.17%	97.85%	98.27%	95.73%	96.37%	96.28%	96.44%
NPMVS–COLMAP (ours)	93.99%	96.83%	82.17%	86.14%	94.63%	97.70%	97.04%	94.89%	97.88%	97.76%	95.28%	91.43%	93.84%	96.34%
NPMVS–ACMMP (ours)	98.27%	99.20%	97.21%	98.43%	99.11%	99.36%	99.31%	98.59%	98.64%	99.44%	97.76%	92.05%	98.90%	99.54%

Table 3. Comparison of completeness (higher is better).

Method	Avg.	Indoor					Outdoor							
		Deli.	Kick.	Offi.	Pipes	Relief	Relief.	Terra.	Cour.	Elec.	Facade	Mead.	Play.	Terrace
CasMVSNet	52.07%	66.41%	55.11%	37.23%	28.05%	75.69%	71.41%	80.94%	48.12%	53.37%	48.28%	13.78%	24.80%	73.76%
COLMAP	76.32%	87.22%	70.91%	52.05%	52.13%	83.05%	81.97%	90.99%	83.90%	83.96%	78.27%	56.82%	77.12%	93.76%
COLMAP-GEO	64.35%	79.84%	45.05%	39.52%	42.12%	73.09%	74.74%	80.07%	73.46%	75.21%	70.03%	39.18%	58.21%	86.07%
ACMMP	94.12%	98.19%	94.19%	95.65%	90.17%	94.06%	93.52%	97.54%	97.93%	96.94%	-	80.42%	93.09%	97.70%
PLC	73.34%	85.76%	77.13%	64.21%	59.78%	75.78%	52.44%	88.43%	82.49%	78.86%	78.65%	53.24%	69.11%	87.52%
HPMVS	28.01%	31.88%	21.95%	10.31%	15.08%	32.95%	30.71%	21.16%	52.63%	21.80%	40.75%	28.20%	20.08%	36.61%
MVE	31.84%	37.11%	19.90%	14.49%	6.37%	31.41%	38.86%	43.08%	51.76%	21.25%	50.20%	14.45%	33.02%	52.05%
OpenMVS	69.50%	85.92%	47.40%	44.41%	49.68%	77.02%	77.44%	85.42%	84.74%	82.37%	69.27%	41.25%	64.76%	93.86%
NPMVS–COLMAP (ours)	86.06%	92.44%	92.98%	89.13%	73.57%	85.23%	86.37%	94.39%	92.20%	86.02%	82.61%	67.53%	79.67%	96.62%
NPMVS–ACMMP (ours)	86.19%	95.35%	83.41%	74.41%	80.48%	88.45%	87.16%	97.43%	94.58%	91.43%	84.74%	71.12%	78.13%	93.84%

Table 4. Comparison of F1 score (higher is better).

Method	Avg.	Indoor					Outdoor							
		Deli.	Kick.	Offi.	Pipes	Relief	Relief.	Terra.	Cour.	Elec.	Facade	Mead.	Play.	Terrace
CasMVSNet	59.75%	73.31%	64.10%	52.27%	42.20%	80.78%	75.76%	77.92%	61.03%	67.43%	58.37%	22.10%	34.51%	66.97%
COLMAP	84.21%	91.62%	80.69%	67.07%	67.62%	89.76%	88.77%	92.84%	90.17%	90.28%	85.95%	70.33%	84.87%	94.77%
COLMAP-GEO	76.69%	88.54%	62.00%	56.58%	59.19%	84.31%	85.37%	88.48%	84.36%	85.71%	81.87%	55.18%	73.22%	92.18%
ACMMP	95.82%	98.47%	93.57%	95.14%	94.32%	96.19%	95.87%	97.94%	98.73%	97.97%	-	87.75%	95.61%	98.31%
PLC	83.41%	92.08%	86.09%	76.91%	74.25%	86.08%	68.69%	93.36%	89.96%	88.07%	87.40%	66.96%	81.34%	93.17%
HPMVS	42.22%	47.75%	35.76%	18.67%	26.05%	49.39%	46.86%	34.72%	67.64%	35.65%	57.15%	42.74%	33.33%	53.20%
MVE	45.59%	51.89%	32.88%	25.21%	11.95%	47.58%	55.57%	59.00%	67.47%	34.69%	65.99%	24.97%	49.24%	66.23%
OpenMVS	79.82%	91.65%	63.76%	61.05%	66.18%	86.36%	86.58%	90.92%	90.83%	89.62%	80.38%	57.78%	77.44%	95.14%
NPMVS–COLMAP (ours)	89.58%	94.58%	87.24%	87.61%	82.78%	91.04%	91.39%	94.64%	94.96%	91.51%	88.49%	77.69%	86.17%	96.48%
NPMVS–ACMMP (ours)	91.68%	97.23%	89.78%	84.75%	88.83%	93.59%	92.84%	98.01%	96.57%	95.27%	90.79%	80.24%	87.30%	96.60%

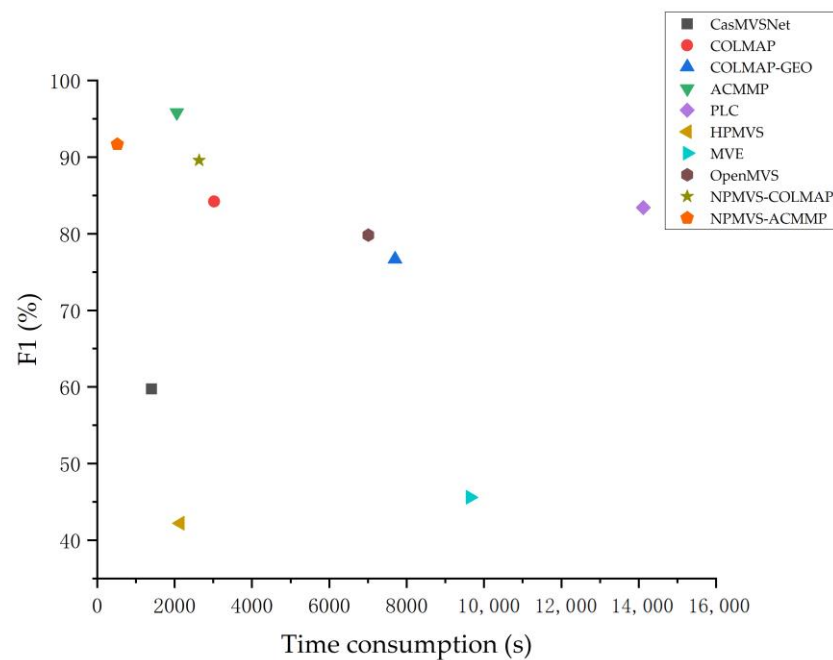


Figure 4. Scatter plot of time consumption and F1 value.

From the experimental results, it can be seen that the method in this paper can maintain the optimal level in processing high-resolution image data, outperforming the comparison algorithms in terms of time consumption and accuracy while maintaining a high level of completeness. The experimental results also illustrate the effectiveness and necessity of the reduced-resolution processing step in our method, which enables the algorithm to perform effective and dense reconstruction of high-resolution images on a common configuration platform. Overall, the experimental results on a variety of different scenes show that our method can effectively perform 3D dense reconstruction, which can be used for fast 3D city modeling.

4. Discussion

Depth map estimation accuracy and time. The method in this paper uses the neural network method to infer the depth as the initial value, optimizes the depth based on superpixel information propagation, and finally optimizes the depth based on pixel-by-pixel information propagation to perform depth inference from coarse to fine. In the experiments, the training data of the original neural network method are the DTU dataset, and the model obtained with this training is used for depth inference on the ETH3D dataset; it is shown that the accuracy of the obtained depth information is poor, which further illustrates the limitations of the deep learning method.

When the predicted scene content exceeds the a priori content of the deep learning network training, the inferred depth results of the deep learning method are often poor, which greatly limits the application of the deep learning method. In this paper, we use a method similar to PatchMatch information propagation to optimize the depth from superpixels to pixel-by-pixel propagation based on the inferred results of deep learning methods. The method in this paper uses the deep learning method as the initial value for optimization, so the depth inference accuracy is better than that of the learning method and the speed is inferior to that of the learning method. Additionally, compared with the traditional PatchMatch propagation method, which uses random values as the initial depth values, this paper has better initial propagation values and performs coarse-to-fine propagation, so the number of iterations is less than that of the traditional PatchMatch propagation method, and the accuracy and time of the results are better than those of the traditional method. The experimental results in Figure 3 also verify that the accuracy

of the depth inference results of our method is better than that of the deep learning method (CasMVSNet) and the traditional PatchMatch method (COLMAP). From Figure 3, it can be seen that the deep learning method has a faster prediction speed and is very advantageous in terms of overall time consumption. Although the method in this paper combines the deep learning method and the traditional method, it can still outperform the traditional geometric method in terms of time consumption for depth inference. Therefore, considering the accuracy and time of depth map estimation, our method combines the advantages of both deep learning methods and traditional geometric methods and shows better performance.

Three-dimensional modeling time. From the experimental results in Table 1, it can be seen that the modeling time of our method (NPMVS-ACMMP) is significantly shorter than that of the other comparison methods. The NPMVS-ACMMP method can shorten the modeling time by about 60–80% compared with the ACMMP method, and the NPMVS-COLMAP method can shorten the modeling time by about 10–20% compared with the COLMAP method. Moreover, the modeling time of the NPMVS-ACMMP method is faster than the CasMVSNet method, and the maximum time can be shortened by about 78%. This is because CasMVSNet reconstruction uses the COLMAP fusion method, and NPMVS-ACMMP reconstruction uses the ACMMP fusion method, so it is faster.

Three-dimensional modeling accuracy. From the experimental results in Table 2, it can be seen that the modeling accuracy of the CasMVSNet method is significantly inferior to that of the other comparison methods. Our method (NPMVS-ACMMP) and the COLMAP-GEO method always maintain a high level of accuracy. Our method as well as the COLMAP, ACMMP, and PLC methods are based on the iterative information propagation of PatchMatch, using the color photometric similarity between multiple images for pixel-by-pixel depth determination. This result verifies the high accuracy of the traditional pixel-by-pixel geometric depth estimation method and meets the expected results of the experiment.

Three-dimensional modeling completeness. From the experimental results in Table 3, it can be seen that the modeling completeness of ACMMP and our method is always significantly better than that of the other methods. The CasMVSNet method has poor extrapolation adaptability for the training model, which leads to poorer scene modeling completeness. The other traditional geometric methods cannot achieve 3D modeling well for non-Lambertian surfaces or weakly textured surfaces and produce some voids; thus, the final modeling completeness is also poor. The depth inference of the neural network in our method can provide relatively good and complete initial values, and depth filtering and information propagation based on superpixels can fill the inferred holes and filter out the noise of non-Lambertian and weakly textured surfaces. Finally, the pixel-by-pixel propagation check can guarantee the modeling accuracy, so our algorithm can achieve complete 3D modeling. The experimental results illustrate the necessity of the superpixel-based linear fitting and information propagation step in our method. This step can be used as a generalized module, which can be easily added to the existing modeling process and provides an idea for improving and optimizing other 3D modeling algorithms.

A comprehensive evaluation of 3D modeling. From the experimental results in Table 4, it can be seen that the F1 value of ACMMP and our method is significantly higher than those of other methods in the experimental dataset of multiple indoor and outdoor scenes. The NPMVS-ACMMP method has the shortest time consumption, and the NPMVS-COLMAP method is only inferior to the CasMVSNet, PMVS, and HPMVS methods. The PMVS and HPMVS are both direct point-cloud-based modeling methods without depth map fusion, which achieve faster overall 3D modeling. However, the final modeling accuracy and completeness of these three methods are poor, for which there are significant differences from our method. Meanwhile, our method adopts better initial depth values, as well as various coarse-to-fine strategies such as downscaled resolution processing and joint superpixel and pixel-by-pixel propagation, which reduce the number of iterations and greatly save running time compared with the same types of traditional geometric methods. Therefore, our method maintains a high accuracy and completeness

level and significantly outperforms in terms of time consumption. Therefore, considering the F1 value and time consumption, it can be seen from Figure 4, which characterizes the comprehensive performance, that the comprehensive performance of our method is better than those methods.

In this study, the downsampling process enabled the neural network to process high-resolution images while increasing the robustness of the subsequent processing and saving processing time. The method uses the neural network for initial depth value prediction in order to greatly reduce the number of subsequent optimization iterations and save overall time while providing better initial values for non-Lambertian surfaces and improving the completeness of the modeling results. The method also uses superpixels for depth plane fitting and joint propagation optimization, which can address loopholes and filter out noise to improve the completeness of the modeling results. The method carries out depth information propagation optimization based on pixel-by-pixel propagation, which better ensures modeling accuracy. The method combines superpixels with pixel-by-pixel propagation and downsampling with upsampling to form a depth inference process from coarse to fine overall, which ensures accuracy; at the same time, it saves overall processing time, improves the modeling completeness, and finally achieves a better modeling result.

5. Conclusions

The goal of this study was to realize the fusion of deep learning methods and traditional methods for fast 3D city modeling. This fusion paves the way for new opportunities to address the problems of poor adaptability of the training models of existing deep learning depth map estimation methods and the general lack of completeness of the reconstruction results of traditional geometric depth map estimation methods. In this study, a new depth map estimation and fast 3D modeling method is proposed by combining the deep learning method and the traditional geometric method. First, the original image is downsampled and input to the neural network to obtain the initial value of depth, and the depth plane fitting and joint optimization are performed using the superpixel information. Then, the optimized depth map is upsampled, and the final depth optimization is performed using pixel-by-pixel depth information propagation. Experimental analysis using various indoor and outdoor scenes showed that the running time and comprehensive performance of the final 3D modeling results of our method are better than those comparison methods.

The framework proposed in this paper is of great significance and provides ideas for related research. The joint superpixel information and downsampling processing method designed in this study can also be directly introduced to other existing method processes to improve performance. The deep learning method and traditional method used in this paper also could be replaced with other methods. Additionally, although the method in this paper achieves good completeness, it still lacks results for some complex non-Lambertian surface reconstructions, which is the next direction for improvement.

Author Contributions: Conceptualization and methodology—Yang Dong, Jiaxuan Song, and Song Ji; software—Yang Dong and Jiaxuan Song; validation and analysis—Jiaxuan Song and Rong Lei; writing—original draft preparation, Yang Dong; writing—review and editing, Dazhao Fan, Rong Lei, and Song Ji; supervision—Dazhao Fan; funding acquisition—Dazhao Fan and Song Ji. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China Project, grant number 41971427; the Program of Song Shan Laboratory (Included in the management of Major Science and Technology Program of Henan Province), grant number 221100211000-4; the High-Resolution Remote Sensing, Surveying, and Mapping Application Demonstration System (Phase II), grant number 42-Y30B04-9001-19/21.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

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

References

- Hirschmuller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 328–341. [\[CrossRef\]](#) [\[PubMed\]](#)
- Tatar, N.; Arefi, H.; Hahn, M. High-Resolution Satellite Stereo Matching by Object-Based Semiglobal Matching and Iterative Guided Edge-Preserving Filter. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1841–1845. [\[CrossRef\]](#)
- Lee, Y.; Kim, H. A High-Throughput Depth Estimation Processor for Accurate Semiglobal Stereo Matching Using Pipelined Inter-Pixel Aggregation. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 411–422. [\[CrossRef\]](#)
- Khan, M.A.U.; Nazir, D.; Pagani, A.; Mokayed, H.; Liwicki, M.; Stricker, D.; Afzal, M.Z. A Comprehensive Survey of Depth Completion Approaches. *Sensors* **2022**, *22*, 6969. [\[CrossRef\]](#) [\[PubMed\]](#)
- Liao, Y.H.; Zhang, S. Semi-Global Matching Assisted Absolute Phase Unwrapping. *Sensors* **2022**, *30*, 411. [\[CrossRef\]](#) [\[PubMed\]](#)
- Barnes, C.; Shechtman, E.; Finkelstein, A.; Goldman, D.B. PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. *ACM Trans. Graph.* **2009**, *28*, 24. [\[CrossRef\]](#)
- Bleyer, M.; Rhemann, C.; Rother, C. PatchMatch Stereo-Stereo Matching with Slanted Support Windows. In Proceedings of the British Machine Vision Conference, Dundee, UK, 29 August – 2 September 2011; pp. 1–11. [\[CrossRef\]](#)
- Shen, S. Accurate Multiple View 3D Reconstruction Using Patch-Based Stereo for Large-Scale Scenes. *IEEE Trans. Image Process.* **2013**, *22*, 1901–1914. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zheng, E.; Dunn, E.; Jojic, V.; Frahm, J.M. PatchMatch Based Joint View Selection and Depthmap Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1510–1517. [\[CrossRef\]](#)
- Xu, Q.; Tao, W. Multi-View Stereo with Asymmetric Checkerboard Propagation and Multi-Hypothesis Joint View Selection. *arXiv* **2018**, arXiv:1805.07920. [\[CrossRef\]](#)
- Yang, X.; Jiang, G. A Practical 3D Reconstruction Method for Weak Texture Scenes. *Remote Sens.* **2021**, *13*, 3103. [\[CrossRef\]](#)
- Ji, M.; Gall, J.; Zheng, H.; Liu, Y.; Fang, L. SurfaceNet: An End-to-End 3D Neural Network for Multiview Stereopsis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2326–2334. [\[CrossRef\]](#)
- Yao, Y.; Luo, Z.; Li, S.; Fang, T.; Quan, L. MVSNNet: Depth Inference for Unstructured Multi-view Stereo. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 1–17. [\[CrossRef\]](#)
- Xu, Q.; Su, W.; Qi, Y.; Tao, W.; Pollefeys, M. Learning Inverse Depth Regression for Pixelwise Visibility-Aware Multi-View Stereo Networks. *Int. J. Comput. Vis.* **2022**, *130*, 2040–2059. [\[CrossRef\]](#)
- Dong, Y. *Spatio-Temporally Coherent 4D Reconstruction from Multiple View Video*; Information Engineering University: Zhengzhou, China, 2020. [\[CrossRef\]](#)
- Furukawa, Y.; Ponce, J. Accurate, Dense, and Robust Multiview Stereopsis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1362–1376. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hernandez, C.; Vogiatzis, G.; Cipolla, R. Probabilistic Visibility for Multi-View Stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8. [\[CrossRef\]](#)
- Campbell, N.D.F.; Vogiatzis, G.; Hernández, C.; Cipolla, R. Using Multiple Hypotheses to Improve Depth-Maps for Multi-View Stereo. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 766–779. [\[CrossRef\]](#)
- Lhuillier, M.; Quan, L. A Quasi-Dense Approach to Surface Reconstruction from Uncalibrated Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 418–433. [\[CrossRef\]](#) [\[PubMed\]](#)
- Habbecke, M.; Kobbelt, L. A Surface-Growing Approach to Multi-View Stereo Reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8. [\[CrossRef\]](#)
- Zhu, Z. *Research on Dense 3D Reconstruction from Unordered Images*; National University of Defense Technology: Changsha, China, 2015; Available online: <https://cdmd.cnki.com.cn/Article/CDMD-90002-1017834269.htm> (accessed on 19 November 2022).
- Curless, B.; Levoy, M. A Volumetric Method for Building Complex Models from Range Images. In Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 4–9 August 1996; pp. 303–312. [\[CrossRef\]](#)
- Zach, C.; Pock, T.; Bischof, H. A Globally Optimal Algorithm for Robust TV-L1 Range Image Integration. In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007; pp. 1–8. [\[CrossRef\]](#)
- Vogiatzis, G.; Torr, P.H.; Cipolla, R. Multi-View Stereo via Volumetric Graph-Cuts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 21–23 September 2005; pp. 391–398. [\[CrossRef\]](#)
- Sinha, S.N.; Mordohai, P.; Pollefeys, M. Multi-View Stereo via Graph Cuts on the Dual of an Adaptive Tetrahedral Mesh. In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007; pp. 1–8. [\[CrossRef\]](#)
- Furukawa, Y.; Curless, B.; Seitz, S.M.; Szeliski, R. Reconstructing Building Interiors from Images. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009; pp. 80–87. [\[CrossRef\]](#)
- Goesele, M.; Curless, B.; Seitz, S.M. Multi-View Stereo Revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 2402–2409. [\[CrossRef\]](#)
- Strecha, C.; Fransens, R.; Van Gool, L. Combined Depth and Outlier Estimation in Multi-View Stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 2394–2401. [\[CrossRef\]](#)
- Vogiatzis, G.; Esteban, C.H.; Torr, P.H.S.; Cipolla, R. Multiview Stereo via Volumetric Graph-Cuts and Occlusion Robust Photo-Consistency. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2241–2246. [\[CrossRef\]](#) [\[PubMed\]](#)
- Goesele, M.; Snively, N.; Curless, B.; Hoppe, H.; Seitz, S.M. Multi-View Stereo for Community Photo Collections. In Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007; pp. 1–8. [\[CrossRef\]](#)

31. Hu, X.; Mordohai, P. Least Commitment, Viewpoint-Based, Multi-View Stereo. In Proceedings of the International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission, Zurich, Switzerland, 13–15 October 2012; pp. 531–538. [\[CrossRef\]](#)
32. Bailer, C.; Finckh, M.; Lensch, H.P. Scale Robust Multi View Stereo. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 398–411. [\[CrossRef\]](#)
33. Galliani, S.; Lasinger, K.; Schindler, K. Massively Parallel Multiview Stereopsis by Surface Normal Diffusion. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 873–881. [\[CrossRef\]](#)
34. Schonberger, J.L.; Zheng, E.; Frahm, J.-M.; Pollefeys, M. Pixelwise View Selection for Unstructured Multi-View Stereo. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 501–518. [\[CrossRef\]](#)
35. Wei, J.; Resch, B.; Lensch, H. Multi-View Depth Map Estimation With Cross-View Consistency. In Proceedings of the British Machine Vision Conference, Nottingham, UK, 1–5 September 2014; pp. 1–13. [\[CrossRef\]](#)
36. Romanoni, A.; Matteucci, M. TAPA-MVS: Textureless-Aware PatchMatch Multi-View Stereo. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10412–10421. [\[CrossRef\]](#)
37. Xu, Z.; Liu, Y.; Shi, X.; Wang, Y.; Zheng, Y. MARMVS: Matching Ambiguity Reduced Multiple View Stereo for Efficient Large Scale Scene Reconstruction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5980–5989. [\[CrossRef\]](#)
38. Xu, Q.; Kong, W.; Tao, W.; Pollefeys, M. Multi-Scale Geometric Consistency Guided and Planar Prior Assisted Multi-View Stereo. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4945–4963. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Hartmann, W.; Galliani, S.; Havlena, M.; Van Gool, L.; Schindler, K. Learned Multi-patch Similarity. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1595–1603. [\[CrossRef\]](#)
40. Kar, A.; Häne, C.; Malik, J. Learning a Multi-View Stereo Machine. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 364–375. [\[CrossRef\]](#)
41. Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; Tan, P. Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2492–2501. [\[CrossRef\]](#)
42. Wang, F.; Galliani, S.; Vogel, C.; Speciale, P.; Pollefeys, M. PatchmatchNet: Learned Multi-View Patchmatch Stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, virtual, 19–25 June 2021; pp. 14189–14198. [\[CrossRef\]](#)
43. Thomas, S.; Johannes, L.S.; Silvano, G.; Torsten, S.; Konrad, S.; Marc, P.; Andreas, G. ETH3D Benchmark. Available online: <https://www.eth3d.net> (accessed on 19 November 2022).
44. Dong, Y.; Fan, D.; Ji, S.; Lei, R. The Purification Method of Matching Points Based on Principal Component Analysis. *Acta Geod. Cartogr. Sin.* **2017**, *46*, 228–236. [\[CrossRef\]](#)
45. Yamaguchi, K.; McAllester, D.; Urtasun, R. Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 756–771. [\[CrossRef\]](#)
46. Schönberger, J.L.; Frahm, J. Structure-from-Motion Revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113. [\[CrossRef\]](#)
47. Schöps, T.; Sattler, T.; Pollefeys, M. BAD SLAM: Bundle Adjusted Direct RGB-D SLAM. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 134–144. [\[CrossRef\]](#)
48. Schöps, T.; Schönberger, J.L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; Geiger, A. A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2538–2547. [\[CrossRef\]](#)
49. Liao, J.; Fu, Y.; Yan, Q.; Xiao, C. Pyramid Multi-View Stereo with Local Consistency. *Comput. Graph. Forum* **2019**, *38*, 335–346. [\[CrossRef\]](#)
50. Locher, A.; Perdoch, M.; Van Gool, L. Progressive Prioritized Multi-view Stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3244–3252. [\[CrossRef\]](#)
51. Fuhrmann, S.; Langguth, F.; Goesele, M. MVE—A Multi-View Reconstruction Environment. In Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage, Darmstadt, Germany, 6–8 October 2014; pp. 11–18. [\[CrossRef\]](#)
52. Cernea, D. OpenMVS: Multi-View Stereo Reconstruction Library. Available online: <https://cdsacave.github.io/openMVS> (accessed on 19 November 2022).
53. Ren, C.; Prisacariu, V.; Reid, I. gSLICr: SLIC superpixels at over 250Hz. *arXiv* **2015**, arXiv:1509.04232. [\[CrossRef\]](#)
54. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Aanaes, H.; Jensen, R.R.; Vogiatzis, G.; Tola, E.; Dahl, A.B. Large-Scale Data for Multiple-View Stereopsis. *Int. J. Comput. Vis.* **2016**, *120*, 153–168. [\[CrossRef\]](#)

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