DENSE RECONSTRUCTION FOR TUNNELS BASED ON THE INTEGRATION OF DOUBLE-LINE PARALLEL PHOTOGRAPHY AND DEEP LEARNING

Rongchun Zhang [a, c], Meiru Jing [a], Xuefeng Yi [b, *], Hao Li [b], Guanming Lu [c]

[a] School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, 210023 Nanjing, China - (rongchunzhang,1021173516) @njupt.edu.cn
[b] School of Earth Sciences and Engineering, Hohai University, 211100 Nanjing, China - hhuyxf@sina.com,lihao@hhu.edu.cn
[c] School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, 210023 Nanjing, China - (rongchunzhang, lugm) @njupt.edu.cn

Commission II, WG II/10

KEY WORDS: Tunnels, Dense Reconstruction, Deep Learning, Double-line Parallel Photography

ABSTRACT:

In scenes of tunnels and underground engineering projects, the 3D modeling based on the traditional horizontal baseline photography method is difficult to make a trade-off between the modeling efficiency and image overlap under narrow space and close photography constraint conditions. Parallel photography provides a better alternative. Furthermore, the undulating tunnel surfaces make occlusion unavoidable, and the pixel scales vary evidently in parallel photography, both of which make it difficult to obtain expected models neither by Semi-Global Matching nor by Patch-based Multi-view Stereo techniques. Comparatively, more accurate 3D reconstruction results can be achieved by using learning techniques, which consider the global semantic information such as specular prior and reflective prior to making the matching more robust. Besides, it is convenient for geologists to acquire photos with smartphones or cameras by a single-line parallel photography method at the tunnel sites. But enough image overlap for reconstruction is still difficult for the above photography way. Therefore, this paper proposes a dense reconstruction method for tunnels using deep learning and double-line parallel photography techniques, and respectively makes comparisons with single-line parallel photography and traditional modeling methods. Experimental results show the feasibility and robustness of the proposed method, especially for tunnel surfaces with complicated textures and occlusions. Moreover, both the completeness and accuracy of the tunnel model with double-line photography are better than with single-line photography.

1. INTRODUCTION

With the continuous development of modernization construction, there are more and more projects in water conservancy, transportation and mining, and there are a large number of tunnel projects in these projects. Due to the complexity of engineering geology and rock mass structure, accidents are prone to occur during tunnel construction, resulting in serious economic losses and a large number of casualties. Geological surveying and mapping can provide necessary data support for tunnel construction and reduce the occurrence of accidents to a certain extent. In the process of tunnel construction, the conventional measurement method is manual surveying and mapping, which has the problems of low work efficiency, low precision, discontinuous data, unintuitive pictures and inability to reproduce the original data of the tunnel. Although the measurement method using a 3D laser scanner can obtain high-precision tunnel model data, it lacks texture information and has a low point cloud density, which cannot show the details of the tunnel surrounding rock that engineers are concerned about. Moreover, the equipment is expensive and has high requirements on the working environment, which is not conducive to the rapid acquisition and analysis of results. However, the dense reconstruction of multi-view images obtained by mobile phones or ordinary digital cameras can obtain more details of the tunnel visualization point cloud.

In terms of 3D reconstruction, the traditional Multi-View Stereo (MVS) is to perform dense reconstruction based on images and poses to recover the geometric information of the scene. It has the characteristics of high precision, low hardware requirements and a large measurement range. Although a problem that has been studied in computer vision for many years, it still remains a challenge due to the limitations of weak textures, transparency, reflections, repeated textures, etc. The rise of Convolutional Neural Networks (CNN) has brought hope that data-driven models can solve problems that traditional MVS models cannot. The 3D reconstruction technology based on deep learning considers global semantic information such as mirror prior and reflection prior, which makes the matching more robust and can obtain more accurate 3D reconstruction results. At present, most learning-based MVS methods are based on the construction of 3D cost volume, which is regularized by 3D CNN, and then the depth map is regressed, but 3D CNN is time-consuming and memory-consuming.

Therefore, in this work, we propose a tunnel dense reconstruction method based on the fusion of double-line parallel photography and deep learning. A new approach to engineering photogrammetry. It aims to reconstruct the tunnel model quickly, efficiently and accurately in the case of limited resources. It not only inherits the speed and efficiency of the learning-based method, but also improves the accuracy of the model with the help of the double-line parallel photography method. We used different modeling methods and different data collection methods to evaluate the model. The experimental results show that our method is superior to other methods in terms of completeness and overall quality, showing the stable generalization ability of the

* Corresponding author

network, and the running time and memory occupation are also superior to other methods, and have good applicability in engineering applications. Applying it in engineering construction can significantly improve the engineering measurement accuracy, which is of great significance to the completion of engineering projects.

2. RELATED WORK

2.1 Traditional Sparse Reconstruction and Dense Reconstruction

Structure from motion, that is, given a sparse corresponding set of multiple images and their image features, to estimate the position of 3D points. This solution process usually involves simultaneous estimation of 3D geometric structure and camera pose motion (the schematic diagram is shown in Figure 1). Image data is large, sourced, and potentially unordered. In view of the above characteristics, three Structure from Motion (SfM) strategies have emerged, including incremental, hierarchical and global.

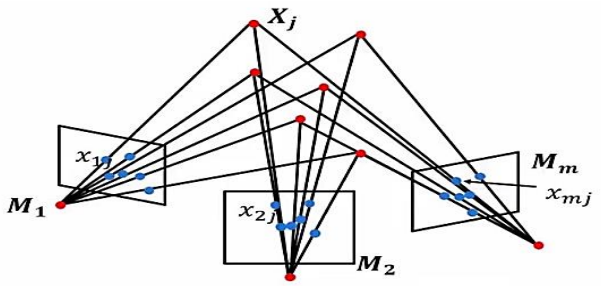

Figure 1. Schematic diagram of Structure from motion.

Since SfM is through feature matching of points, it is impossible to obtain dense point clouds. MVS, on the other hand, matches almost every pixel in the photo and reconstructs the three-dimensional coordinates of almost every pixel to obtain a dense point cloud. The theoretical basis is that there are geometric constraints between multiple views. On the basis of the known camera pose, the same name point on the photo is searched pixel by pixel, and a consistency metric function is established to obtain a dense 3D point cloud on the surface of the target scene. There are four main types: voxel-based, point cloud-based, patch-based, and depth map fusion-based methods.

2.2 Traditional Methods of Tunnel Image Modeling

Image modeling is to use a camera to shoot real-world objects and scenes and process them through computer vision technology to obtain a three-dimensional model of the object. Wang et al. (2018) used alternating-current photography to collect images of tunnel surface information, and used 3D Scene Restoration Structure for 3D reconstruction of tunnels, which has high processing efficiency and high precision. But the accuracy of its model depends on the stability of the measurement system. Xue et al. (2018) used the data obtained from a 3D laser scanner to reconstruct the surface of the tunnel using traditional reconstruction algorithms such as semi-global matching, multi-frequency heterodyne method, etc., the accuracy is higher but the corresponding cost is also higher. Newcombe et al. (2011) proposed the Kinect fusion algorithm, which has the advantages of limited drift and high accuracy, and can reconstruct 3D scenes of complex regions with high precision in real time, but the region is limited. Whelan et al. (2011) proposed the kintinuous algorithm by improving the Kinect fusion algorithm, which enhances the robustness and reduces the overall drift potential, and solves the limitation of not being able to work in a large area, but it has not been successfully applied in complex construction scenarios such as tunnels.

2.3 3D Modeling Based on Deep Learning

David et al. (2014) used deep learning to recover depth maps from a single image, estimate depth from coarse to fine and propose a scale-invariant loss function. Christopher et al. (2016) used voxels for single-view or multi-view 3D reconstruction, establishing a mapping from a 2D image to a 3D voxel model. Fan et al. (2017) used point cloud to do 3D reconstruction of a single RGB image, used a deep network to directly generate a point cloud from a single image, and constructed a Min-of-N loss function to solve the problem of shape diversity after 2D image reconstruction. Improved accuracy of reconstruction. Wang et al. (2018) used triangular meshes for 3D reconstruction of a single RGB image, used graph convolutional networks to represent 3D mesh information, and defined four loss functions to generate better mesh models. Wei et al. (2019) implemented 3D modeling of invisible parts of the scene and proposed a consistency loss for online optimization. Dusmanu et al. (2019) proposed D2-Net to solve the problem of reliable correspondence of pixels in complex scenes. Revaud et al. (2019) proposed R2D2 for keypoint detection and local feature description, which can obtain sparse and reliable keypoints after self-supervised training.

2.4 The Development History of the MVS Algorithm Based on Deep Learning

Yao et al. (2018) proposed MVSNet, a cost volume based on differentiable homography transformation for multi-view depth estimation, which opened up a new field of 3D reconstruction by depth model prediction. Subsequent improvements were made to it, and RMVSNet was proposed, which replaced the 3D convolution with the GRU time series network to reduce the size of the model, and the loss function was also changed to the cross-entropy loss of multi-classification. Although the model became smaller, the accuracy was also reduced. At the same time, Guo et al. (2018) changed the original TensorFlow version to the PyTorch version, which increased the readability of the code and provided great help for subsequent algorithm improvements. Chen et al. (2019) proposed PointMVSNet to convert the predicted depth map into a point cloud and then optimize the depth regression. Yang et al. (2020) proposed CVP-MVSNet, which changed the MVSNet model to a hierarchical one, and used the predicted depth information to reduce the cost volume, but it was slow and computationally expensive. Later, Yu et al. (2020) proposed Fast-MVSNet, which improves the running speed of MVSNet by adopting sparse cost body and Gauss-Newton layer. Yan et al. (2020) proposed D2HC-RMVSNet, which uses LSTM to process the cost volume, and proposes a Dynamic Consistency Checking for post fusion, which significantly reduces memory consumption. Wei et al. (2021) proposed AA-RMVSNet to utilize an adaptive hybrid network with cyclic structure for cost volume regularization, which improves the performance in low-texture regions.

3. METHOD

First, considering that traditional horizontal baseline photography is prone to generate more voids when modeling the surrounding rock surface of the tunnel, this paper adopts the double-line parallel photography method to obtain multi-view images, that is, stereo pairs are photographed in parallel along both sides of the central axis of the tunnel. Secondly, due to the

occlusion in the tunnel scene, it is difficult for traditional modeling techniques to achieve good results, and due to the limitation of engineering resources, 3D laser scanning was not performed in this experiment. Instead, the PatchmatchNet method with high computational efficiency and low memory requirements is introduced (the method flow is shown in Figure 2).

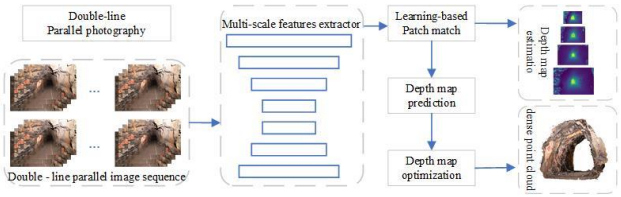

Figure 2. Flowchart of the method in this paper.

3.1 Double-line Parallel Photography

Parallel photography. Due to the topographical constraints of the narrow passage, the width of the lateral site required by conventional photogrammetry was insufficient for observation. Parallel photography can solve this problem. Parallel photography refers to a photography method in which the main optical axis of the camera is consistent with the baseline direction. Compared with the traditional photography method, parallel photography has a smaller intersection angle of the three-dimensional image pair, and can lay the baseline in a special way in the underground cave or in a narrow area, which overcomes the problem that the conventional photography method does not have enough venues to lay the baseline in the above environment. A large number of experiments have proved that the accuracy of parallel photogrammetry is much higher than that of ground stereoscopic photogrammetry in long and narrow areas, and the implementation is more convenient.

Double-line parallel photography. That is, on the basis of parallel photography, the surveyor holds a digital camera and shoots three-dimensional pairs in parallel at intervals along both sides of the central axis of the tunnel. Each image includes the side wall and top of the tunnel. To determine the hole diameter, it is necessary to ensure that the side walls and tops of two adjacent images have a certain overlap area (Figure 3). Different from the binocular lens, double-line parallel photography mainly uses a single camera. The reason is that the professional geological workers working on the field usually carry mobile phones or ordinary monocular cameras with them. Therefore, double-line parallel photography has better practicability.

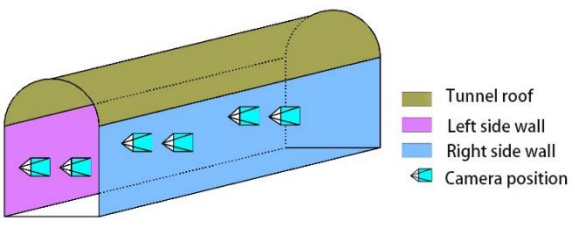

Figure 3. Schematic diagram of the tunnel double-line parallel sequence image.

3.2 PatchmatchNet

This is an end-to-end global matching algorithm that directly computes depth maps without epipolar correction. A cascade structure with learning-based patchmatch as the main body. The model is embedded in the accelerated computing framework in a coarse-to-fine fashion. Multi-layer depth maps are separately predicted by learning-based block matching, and optimized by passing depth residuals through 2D convolution. Randomness is added in the training phase to improve the robustness of the model in terms of visibility estimation and generalization. PatchmatchNet does not rely on 3D cost volume regularization, but embeds it into a learning-based cascade formulation followed by an adaptive learning propagation and evaluation module based on deep features. The structure is simpler and more efficient. Outperforms existing learning-based MVS methods in terms of running time and memory consumption. The performance is at least 2.5 times faster than the existing best MVS scheme, and the memory footprint is 2 times less (as shown in Table 1, source: Wang et al., 2020). Experiments show that this method is very suitable for resource-constrained engineering applications. The emergence of PatchmatchNet breaks the inherent thinking of traditional learning methods. Although it does not completely subvert the previous algorithm process, by introducing the learned patchmatch, it has reached a level comparable to the SOTA method with its simplified network structure, faster operation efficiency and very little memory usage. It includes three modules, namely multi-scale feature extraction using FPN, depth map prediction and depth map optimization based on learning patchmatch, and spatial refinement.

Learning-based algorithm	GPU memory (GB)	Run-time(s)	Overall error(mm)
CasMVSNet	5.5	0.60	0.37
UCS-Net	4.0	0.55	0.38
CVP-MVSNet	5.7	1.20	**0.35**
Fast-MVSNet	4.1	0.52	0.39
R-MVSNet	7.2	1.60	0.42
MVSNet	10.8	1.25	0.57
PatchmatchNet	**1.9**	**0.37**	0.36

Table 1. Error comparison with learning-based MVS methods on GPU memory and runtime (source: Wang et al., 2020).

3.2.1 Learning-based Patchmatch Step

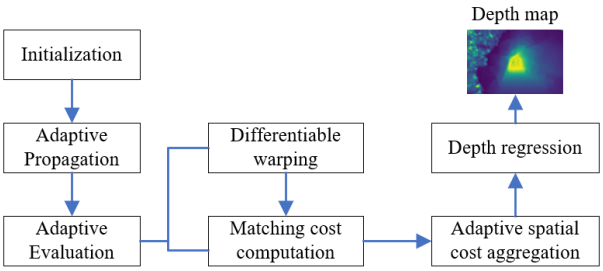

Figure 4. Learning-based patchmatch flowchart.

Based on the traditional patchmatch, it is extended into a learnable, adaptive module. It consists of three main steps (Figure 4): (1) Random initialization: random depth assumptions are made for each pixel based on a pre-defined depth range; (2) Adaptive propagation: hypotheses are collected from pixels on the same surface and propagated to its neighborhood by deformable convolutional network, so that depth hypotheses can be collected more efficiently in both textured and untextured regions. (3) Adaptive evaluation: calculate the matching cost of all hypotheses and select the hypothesis with the least cost as the best scheme to generate depth prediction results. It consists of the following four steps: differentiable warping, matching cost

calculation, adaptive spatial cost aggregation and depth regression. After initialization, iterate between propagation and evaluation until convergence. In order to avoid bias within a certain depth range, a deep residual network is also designed to optimize the depth map.

3.2.2 Confidence

The quality of depth estimation is defined as the probability of real ground depth in a small range around the predicted depth. Since the depth assumption is sampled discretely along the frustum of the camera, so we only need to compute the probability sum of the four nearest depth hypotheses around to assess the quality of the depth estimate. The estimated depth $D(q)$ at pixel q is computed by exploiting an inverse depth regression based on a soft argmin operation in Equation 1 (Figure 5).

$$\mathrm{D(q)} = \left(\sum_{j=0}^{D-1} \frac{1}{d_j} \cdot \mathrm{P(q,}j\mathrm{)}\right)^{-1} \quad (1)$$

where $P\ (q,j)$ is the probability of pixel q under the $j-th$ depth assumption.

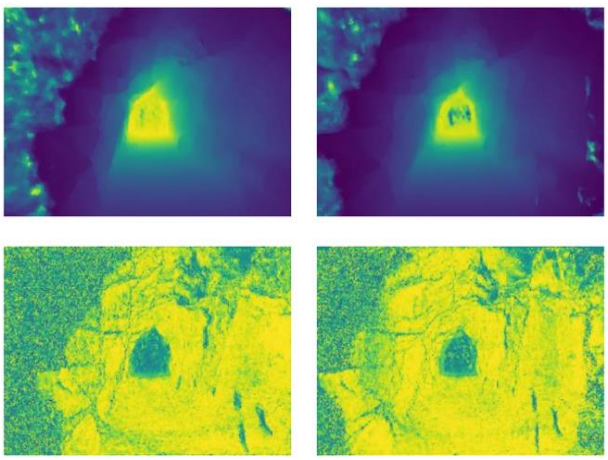

Figure 5. Top: Example of a tunnel depth map, Bottom: Example of a tunnel confidence map.

3.2.3 Loss Function

All depth estimation losses are considered in the adopted loss function, i.e., the total loss function is the sum of the patchmatch iteration losses of all depth map estimation stages in Equation 2.

$$\mathrm{L_{total}}=\sum_{k=1}^{3}\sum_{i=1}^{n_k}\mathrm{L_i^k}+\mathrm{L_{ref}^0} \quad (2)$$

We used smooth L1 loss for L_i^k, the loss of the $i-th$ iteration in stage k (k = 1, 2, 3) of patchmatch and L_{ref}^0, to finally refine the loss of the depth map.

3.2.4 Training Strategy

Many learning-based methods select the best two source views for training, but if the selected source view and reference view have strong correlation, it will affect the weight of model training. The training strategy of PatchmatchNet adopted in this paper is based on Pixelwise Visibility-aware multi-view Stereo Network (PVSNet). For each reference image, four were randomly selected from the ten best source images to train on the DTU dataset. This strategy shortens the training time of the model and also improves the generalization performance of the model. In addition to this, random source images with weak visibility are also trained, which further improves the generalization ability of the model and the robustness of visibility estimation. Therefore, we used the trained model without any fine-tuning to test our data, and the result shows its stable generalization ability.

4. EXPERIMENT

In this paper, experiments were carried out in the exploration tunnel of the Daigusi dam site of the Bailongjiang River Water Diversion Project. In order to verify the performance of this method, we compare it with traditional dense reconstruction algorithms such as motion-based structural pipeline (SfM) and Patch-based Multi-View Stereo (PMVS). In addition, the effects of the two image acquisition methods on the tunnel modeling results are compared. Since we use an ordinary mobile phone or a monocular camera to capture the images, there is no need for calibration and pixel block matching, so the accuracy and stability of the measurement system are guaranteed to a certain extent.

4.1 Experimental Data

The data in this paper come from the exploration tunnel at the Daigusi dam site of the Bailongjiang River Diversion Project. A total of 57 sequence images were shot. The photography lights were used to illuminate the images. The photography lights were lit along the direction of the tunnel shot by the camera and as parallel to the main optical axis as possible to minimize the generation of shadows. 36 images were selected from the sequence images as the main experimental objects of this paper for learning-based double-line parallel photography dense reconstruction (Figure 6). And 19 images were selected as single-line images for comparative analysis. The true value of the data used in this paper comes from the point cloud data extracted from the video collected by the Kinect DK.

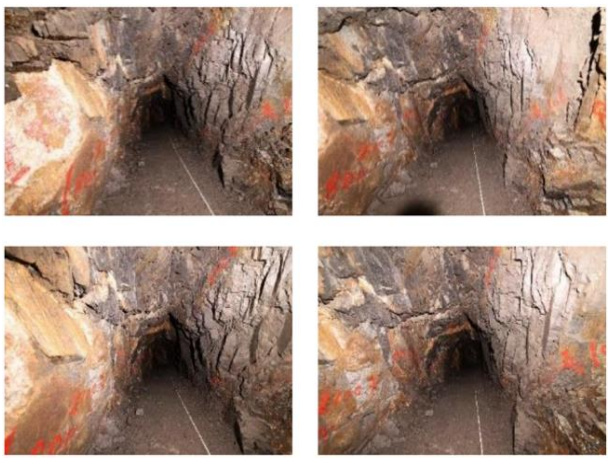

Figure 6. Example image of tunnel double-line parallel photography sequence.

4.2 Comparative Experiment

To verify the quality of the proposed method, we conduct four sets of comparative experiments. They are: (1) The single-line image obtained based on the traditional horizontal baseline photography method is reconstructed by the traditional method, namely SfM + PMVS to reconstruct the 3D model of the tunnel; (2) Double-line images obtained based on double-line parallel photography method are reconstructed using the traditional method, namely SfM + PMVS to reconstruct the 3D model of the tunnel;(3) The single-line image obtained based on the traditional

horizontal baseline photography method adopts the learning-based method, namely PatchmatchNet to reconstruct the 3D model of the tunnel; (4) The double-line image obtained based on the double-line parallel photography method adopts the learning-based method, namely PatchmatchNet to reconstruct the 3D model of the tunnel.

4.3 Run Time Comparison and Result Analysis

We compared the running time of the two image acquisition methods using different reconstruction methods, as shown in Table 2, and observed that the running time of the learning-based method is much less than that of the traditional method, which takes about half an hour, is nearly 6 times that of learning-based methods. The reconstruction results of the traditional method and this method are shown in Figure 7. The single-line and double-line images of the traditional method generate 520,000 and 1.12 million points, respectively, and the single-line and double-line images based on the learning method generate 7.35 million and 21.92 million points, respectively. It can be seen that the deep learning-based method reconstructs a denser and more detailed point cloud, which reflects a high degree of completeness, and the reconstruction effect on the boundary is also better. And the model obtained by double-line photography is more dense, more complete, more accurate and more precise than single-line photography. The side walls and top details can be reconstructed well, which provides convenience for subsequent engineering applications. In contrast, the traditional method can only reconstruct an outline, lose a lot of details, bring inconvenience to subsequent work, and cannot provide good help for practical engineering applications. Although the model generated based on double-line photography is more accurate than single-line photography, it still cannot meet the actual needs of engineering applications (Figure 8-11).

Acquisition	Processing	Run time(min)	Mean distance(m)
Traditional	Traditional	10	0.170972
Double-Line	Traditional	30	0.122022
Traditional	Patchmatchnet	2	0.0652622
Double-Line	Patchmatchnet	5	0.0611644

Table 2. Comparison of four experimental results.

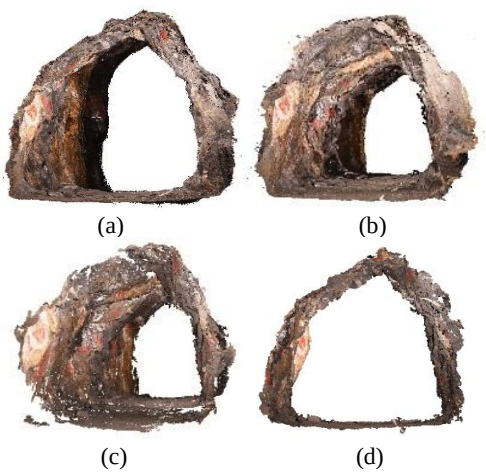

Figure 7. Dense point cloud. (a): single line based on deep learning; (b): double line based on deep learning; (c): double line based on SfM+PMVS. (d): single line based on SfM+PMVS.

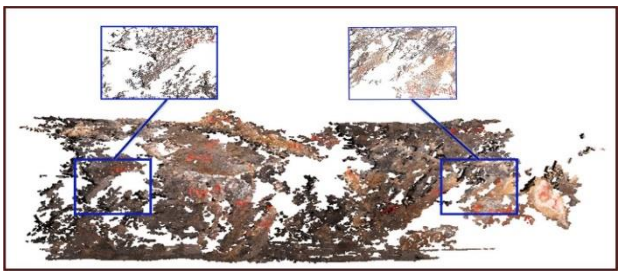

Figure 8. Example of 3D reconstruction of single-line SfM + PMVS (side wall).

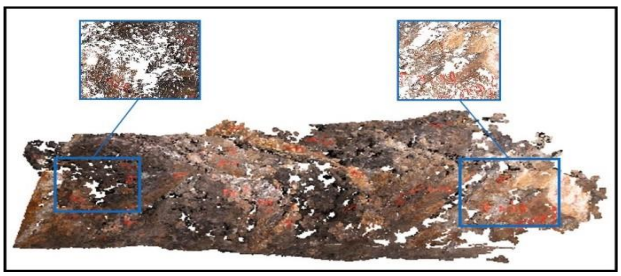

Figure 9. Example of 3D reconstruction of double-line SfM + PMVS (side wall).

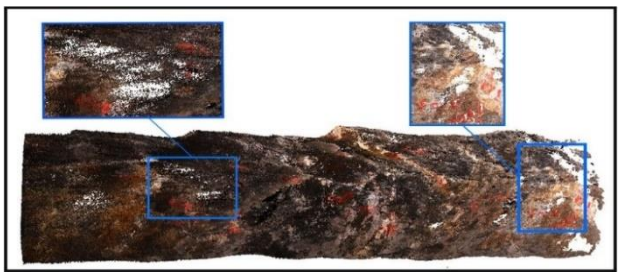

Figure 10. An example of a single-line 3D reconstruction based on deep learning (side wall).

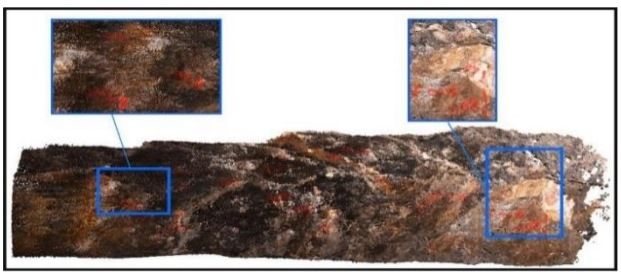

Figure 11. An example of double-line 3D reconstruction based on deep learning (side wall).

4.4 Accuracy Evaluation

The point cloud extracted from the video data collected by Kinect DK is used as the experimental ground truth to evaluate the details and completeness of the method. Figure 12 shows a dense point cloud of 2.56 million points. Compare and analyze the closest point distance to the generated dense point cloud. The results are shown in Figure 13. The point-to-point distance of the single-line and double-line results of this method is less than 0.06m, which are 70.96% and 74.01%, respectively. The point-to-point distance of the single-line and double-line results of the traditional method is less than 0.06m, which are 22.53% and 36.36%, respectively. By comparing with the traditional method, it can be found that the method in this paper has better performance in terms of completeness and details.

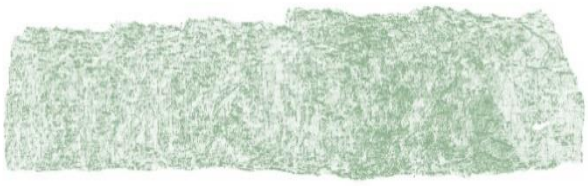

Figure 12. Ground truth, dense point cloud extracted from video data collected by Kinect DK, with a total of 2.56 million points.

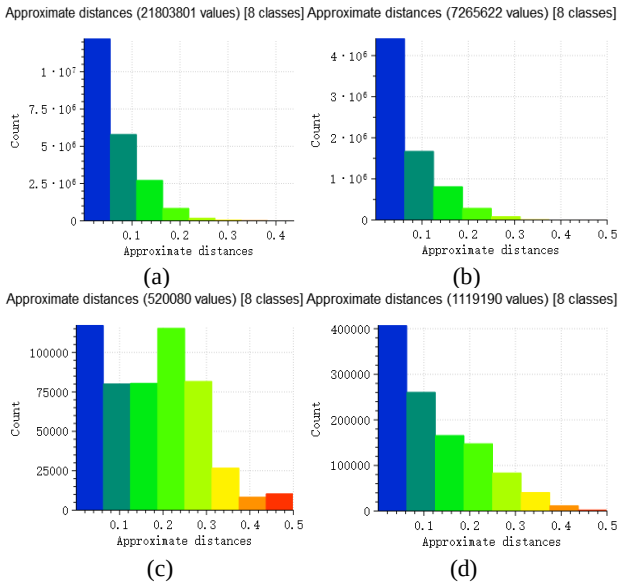

Figure 13. Analysis of point cloud registration accuracy.
(a): Accuracy map of double-line image of deep learning;
(b): Accuracy map of single-line image of deep learning;
(c): Accuracy map of single-line image of traditional method;
(d): Accuracy map of double-line image of traditional method.

5. CONCLUSIONS

In view of the complex texture and occlusion characteristics of tunnels, this paper proposes a dense reconstruction method of tunnels that uses deep learning and double-line parallel photography technology. It solves the problem that underground engineering such as tunnels cannot quickly and efficiently reconstruct models with precise details due to limited resources, and provide technical support for the subsequent practical production and application of underground engineering. And through a series of experiments to verify its performance, the results show that the accuracy, completeness and accuracy are better than the traditional technology. The method acquires images by means of double-line parallel photography and uses improved patchmatch to estimate depth map.

The advantages of this method are: First, the parallel photography reconstruction method based on deep learning can extract features more clearly and accurately, and reduce the influence of complex texture, occlusion and roughness of the tunnel surface. Second, the comparison of the local details of the tunnel shows that the tunnel model generated by the method in this paper is more detailed and complete. In addition, the double-line parallel photography mode better considers the efficiency and accuracy of tunnel reconstruction. Third, the proposed method has better applicability to geological engineering reconstruction. The application on tunnel data shows that PatchmatchNet not only outperforms previous methods in performance, but also improves efficiency and speed by several times. Furthermore, without any fine-tuning, it produces excellent results on tunnel data, showing its stable generalization ability.

The comparison experiment with the traditional method (SfM+PMVS) shows that due to the adaptive depth propagation and evaluation, the number of point clouds reconstructed by the method in this paper increases significantly, and the generated model has better integrity and higher accuracy.

All in all, in the scenes of tunnels and underground engineering, the deep learning dense reconstruction method based on double-line parallel photography has higher precision, higher accuracy and better integrity, and has more practical value in the geological field.

ACKNOWLEDGEMENTS

This research was funded by the National Natural Science Foundation of China (Grants No. 41901401), the Natural Science Foundation of Jiangsu Province (Grants No. BK20190743), and the China Postdoctoral Science Foundation (Grants No. 2021M691653).

REFERENCES

Choy, C. B., Xu, D., Gwak, J., Chen, K., Savarese, S. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, 628-644. Springer, Cham.

Chen, R., Han, S., Xu, J., Su, H. 2019. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1538-1547.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., Sattler, T. 2019. D2-net: A trainable cnn for joint description and detection of local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8092-8101.

Eigen, D., Puhrsch, C., Fergus, R. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, *27*.

Fan, H., Su, H., Guibas, L. J. 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 605-613.

Furukawa, Y., Hernández, *C.*, 2015. Multi-View Stereo: A Tutorial. *Found. Trends Comput. Graph. Vis.*, 9, 1-148.

Luo, K., Guan, T., Ju, L., Huang, H., Luo, Y., 2019. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10452-10461.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Fitzgibbon, A. 2011. KinectFusion: Real-time dense surface mapping and tracking, *10th IEEE International Symposium on Mixed and Augmented Reality*, 127-136.

Revaud, J., Weinzaepfel, P., De Souza, C., Pion, N., Csurka, G., Cabon, Y., Humenberger, M. 2019. R2D2: repeatable and

reliable detector and descriptor. *arXiv preprint arXiv:1906.06195*.

Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M., 2021. PatchmatchNet: Learned Multi-View Patchmatch Stereo. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14194-14203.

Wang, L., Chen, R., Kong, D. 2014. An improved patch based multi-view stereo (PMVS) algorithm. In *3rd International Conference on Computer Science and Service System*, 9-12.

Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y. G. 2018. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, 52-67.

Wei, X., Zhang, Y., Li, Z., Fu, Y., Xue, X., 2020. Deepsfm: Structure from motion via deep bundle adjustment. In *European conference on computer vision*. Springer, Cham.

Wei, Z., Zhu, Q., Min, C., Chen, Y., Wang, G. 2021. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6187-6196.

Wei, Y., Liu, S., Zhao, W., Lu, J. 2019. Conditional single-view shape generation for multi-view stereo reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9651-9660.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 767-783.

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L. 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5525-5534.

Yang, J., Mao, W., Alvarez, J. M., Liu, M. 2020. Cost volume pyramid-based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4877-4886.

Yan, J., Wei, Z., Yi, H., Ding, M., Zhang, R., Chen, Y., ... Tai, Y. W. 2020, August. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European Conference on Computer Vision*, 674-689. Springer, Cham.

Yu, Z., Gao, S. 2020. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1949-1958.