Integration of Depth Normal Consistency and Depth Map Refinement for MVS Reconstruction

Lifang Yang, Zhengyao Bai(✉), and Huijie Liu

School of Information Science and Engineering, Yunnan University, Kunming 650500, China
baizhy@ynu.edu.cn

Abstract. To address the problem of incomplete Multi-view Stereo (MVS) reconstruction, the initial depth and loss function of the depth residual iterative network are investigated, and a new multi-view stereo reconstruction network integrating depth normal consistency and depth map thinning is presented. Firstly, downsampling the input image to create an image pyramid and extracting a feature map from the image pyramid; Then, constructing a cost volume from the 2D feature map, adding the depth normal consistency to the initial cost volume to optimize the depth map. On the DTU data set, the network is tested and compared to traditional reconstruction approaches and MVS networks based on deep learning. The experimental results show that the proposed MVS reconstruction network was produced the better results in completeness and increased the quality of MVS reconstruction.

Keywords: Normal-depth consistency · Feature loss · Cost volume · Depth map refinement · MVS

1 Introduction

MVS (Multi-view Stereo) is a popular topic in computer vision, it has been widely employed in virtual reality, automatic driving, digital libraries, and cultural relics restoration [1]. To calculate the correspondence of high-density 3D point clouds and recover 3D point information, traditional MVS algorithms [2, 3] typically use artificially built rules and indicators. Approaches provide satisfactory accuracy, but reconstruction completeness still needs to be improve. Recently, a deep learning method [4, 5] employs a Deep Neural Network to infer the depth map of each view, this approach can extract identifying features and encode the scene's global and local information, allowing it to learn high brightness or reflection information and provide robust feature matching.

MVS reconstruction based on depth learning has yielded satisfactory results. The MVSNet network introduced by Yao et al. [6] is the most well-known. The most important stage is to create a cost volume based on plane scanning, regularize it with a 3D CNN network, and achieve effective depth reasoning accuracy. However, because of the network's high memory comsumption, it is not applicable to in large-scale scenarios. To address this issue, Yao et al. [7] presented R-MVSNet, a cyclic network that employs

Y. Wang et al. (Eds.): IGTA 2022, CCIS 1611, pp. 118–129, 2022.
https://doi.org/10.1007/978-981-19-5096-4_9

Gate Recurrent Unit (GRU) instead of 3D CNN to regularize the cost volume, reducing storage consumption at the cost of the increased average error of estimated depth and running time. Chen et al. [8] proposed the Point-MVSNet network, which iteratively predicts the depth residue as well as visual brightness using edge convolution of the k closest neighbors of each 3D point. The network accuracy improves, but the running time increases linearly as the number of iteration layers increases. A pyramid residual network has recently been utilized to iteratively infer depth reconstruction of multi-view stereo [9, 10] with promising results. The depth residual network tackles the problem of decreasing operating efficiency as the network deepens. The network performance and speed are excellent, but because it uses the coarsest depth as the residual depth to estimate the next level of depth, the depth map generated at the coarsest level is critical to the final reconstruction. The initial depth discontinuity can lead to a loss in the completeness of the entire network since errors at the coarsest level might spread to the final level and cause details to be lost. This research offers a depth reasoning supervision network to tackle the issues by making the estimated depth map continuous.

2 Related Work

Voxels [11], level sets [12], polygonal meshes [13], and depth maps [14] are commonly used in traditional MVS approaches to represent the three-dimensional geometry of objects or scenes. Due to its great accuracy and excellent performance in many settings, the COLMAP algorithm proposed by Schonberger et al. [15] is representative of classic MVS. However, it runs for a long period and is inefficient. Although the classic 3D reconstruction still remains the main part of the research, more and more researchers begin to focus on the MVS method based on volume and depth. Most objects or sceneries can be modeled using volume representation. The volume-based method separates the entire body into small voxels and then applies a photometric consistency measure to determine if the voxel belongs to the surface, given a set volume of an item or scene. These approaches have limitations in modeling scenes. They do not impose constraints on the geometry of objects. The MVS method based on the depth map, on the other hand, allows for more degrees of freedom in scene modeling.

Deep learning-based algorithms are commonly utilized to tackle stereo matching difficulties and obtain good results in three-dimensional vision challenges. These learning-based approaches, on the other hand, are not well suited to multi-view reconstruction challenges. Kar et al. [16] proposed a learnable method for projecting pixel features upwards into three-dimensional objects and classifying whether a voxel is filled by a surface. These networks, however, are incapable of handling large-scale scenarios because the used volume representation requires a lot of memory. MVSNet proposed by Yao et al. [6], was the first multi-view 3D reconstruction using a depth map. MVSNet estimates multi-view depth based on the cost volume generated by differential homography transformation, which is inspired by the binocular stereo matching estimation approach. MVSNet takes a reference image and several source images as inputs, transform the features of several source images into reference images to construct cost volume, regularizes them with 3D CNN to obtain probability, and uses argmax to select the depth with the highest probability as the depth of points. The key to MVSNet is to build low-cost volume using differentiable transformations. Because the network learns the depth

map of each view using 3D CNN regularization and derives 3D geometry from many views by fusing the estimated depth map, the network storage capacity will expand, making it harder to utilize the remaining information in high-resolution images. Yao et al. [7] proposed the R-MVSNet circular network for large-scale scene reconstruction to address this problem. The cost volume is first built in the same way as MVSNet [6] and then regularized sequentially using GRUs rather than 3D CNN. This approach requires less memory, but it takes longer to execute. Chen et al. [8] presented Point-MVSNet, a framework for predicting depth from coarse to fine on point clouds that allows information from k nearest neighbors to be obtained in 3D space while repeatedly refining the depth map to greatly minimize running time. It works in the same way as Cascade MVSNet [17], but it minimizes the searching range of cost volume and estimate the high- resolution depth of huge scenes with reduced GPU consumption and improved estimation fidelity.

This study differs from the previously discussed network iterative depth map refinement from coarse to fine. First, these methods ignore the impact of the initial depth map's edge discontinuity on the output, but our network includes a depth normal consistency module after the coarsest depth. Depth normal consistency method [18] that has been formalized because the normal of a surface can represent identical properties on the same plane, it can be used as a constraint to better communicate semantic information, which is similar to using the normal as a depth function and applying a hard constraint to it. Second, typical multi-view stereo reconstruction supervises the learning and training model with a pixel-by-pixel loss function, which produces a big mistake when the same shot moves one pixel or uses various resolutions. Inspired by multi-scale loss functions [18, 19], we use feature loss function to multi-view stereo reconstruction to optimize the training of a deep iterative network, thus improving the reconstruction completeness and robustness.

3 Main Methods

This paper focuses on the study of the depth normal consistency and feature loss function. Depth normal consistency ensures that depth estimation matches geometric prediction results and eliminates the problem of a discontinuous edge at the beginning of the depth measurement. The similarity of object features is taken into account by the feature loss function, which improves the accuracy of the final estimated depth. The main modules of this network are the feature pyramid, cost volume pyramid, depth normal consistency, and loss function, as shown in Fig. 1.

3.1 Feature Pyramid

In this paper, the input source image and reference image are down-sampled to different scales, and an L + 1 level image pyramid $\left\{I_i^L\right\}_{i=0}^N$, $i \in \{1, 2, ..., N\}$ is constructed. The undersampled original image $I_i^0 = I_i$ is represented by the lowest layer of the image pyramid. After downsampling, the image's resolution gradually decreases. The smaller the image and the lower the resolution, and vice versa. After the acquisition of the image pyramid, the feature extraction network CNN is used to compute features at each

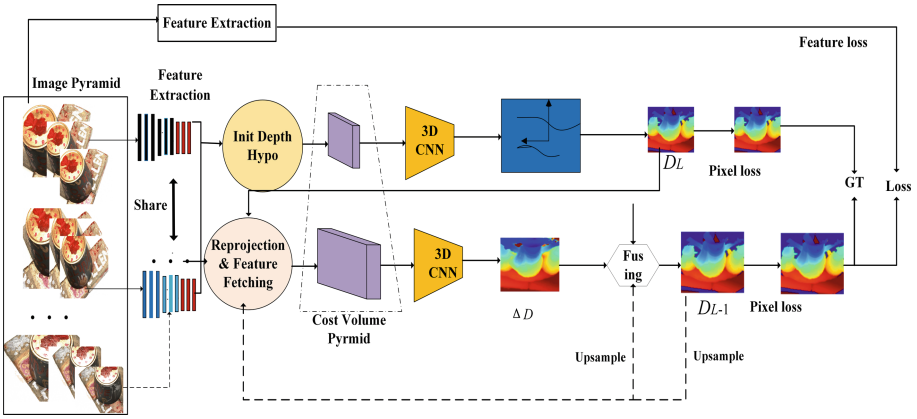

Fig.1. Overall framework of the network in this article

image scale to build the feature pyramid. There are nine convolutional layers in the CNN. LeakyReLU activation layer is inserted after each convolutional layer. The last layer of the feature pyramid is formulized by $\left(f_i^L\right)_{i=0}^N, f_i^L \in R^{H/2^L * W/2^L * F}$, F = 16, W and H are length and width of the feature map. With less computing effort, the feature pyramid combines feature maps with strong low-resolution semantic information and weak high-resolution semantic information.

3.2 Cost Volume Pyramid

The cost volume pyramid is mainly composed of the cost volume of rough depth map reasoning and the cost volume of multi-scale depth residual reasoning. A cost volume for the L-level feature map with the lowest resolution is first established. The cost volume of the reference map is created by uniformly sampling M parallel planes in the depth range, assuming that the depth range measured on the reference image of the scene is $d_{min} - d_{max}$. The sampling depth is $d = d_{min} + (d_{maxx} - d_{\min})/M, m \in \{0, 1, 2, ..., M - 1\}$ represent depth plane, and its normal $\boldsymbol{n_0}$ is the reference camera's main axis. Given the reference image set I_{ref} and the camera parameter $\{\boldsymbol{K}_i, \boldsymbol{R}_i, \boldsymbol{t}_i\}, i = I_{ref} \cup A$, the differentiable homography matric between the first source view and reference view with depth d is defined as

$$\boldsymbol{H_i}(d) = \boldsymbol{K}_i^L \boldsymbol{R}_i \left(\boldsymbol{I} - \frac{(\boldsymbol{t}_0 - \boldsymbol{t}_i)\boldsymbol{n}_0^T}{\boldsymbol{d}} \right) \boldsymbol{R}_0^{-1} \left(\boldsymbol{K}_0^L \right) \tag{1}$$

where $\boldsymbol{K}_i^L$ and $\boldsymbol{K}_0^L$ is $\boldsymbol{K}_i$ and $\boldsymbol{K}_0$ calibration internal parameter matrix at the L layer, and I is the identity matrix, which $\boldsymbol{K}_i, \boldsymbol{R}_i, \boldsymbol{t}_i$ represents the camera's intrinsic characteristics and external items. This study reconstructs the feature map $\left\{f_i^L\right\}_{i=1}^N$ corresponding to the reference view f_0^L using differentiable bilinear interpolation, then produces the cost

volume prediction, given the source view and the L-level feature pyramid $\left\{\overline{f}_{i,d}^{L}\right\}_{i=1}^{N}$. The feature variance of N + 1 views is defined as the cost volume of all pixels at depth d.

$$C_d^L = \frac{1}{N+1}\sum\nolimits_{i=0}^{N}\left\{\overline{f}_{i,d}^{L} - \overline{f}_{d}^{L}\right\} \tag{2}$$

where $\overline{f}_{d}^{L}$ is the depth of the reference image and the mean value of all feature maps is d. A multi-scale 3D CNN network is used to regularize the cost volume, and the probability distribution of depth estimation at different depth samples is obtained to eliminate the influence of non-ideal Lambertian volume. The second stage is the cost volume prediction using multi-scale depth residuals, which will be covered in depth normal consistency Sect. 3.3.

3.3 Depth Normal Consistency

Due to interference factors such as the environment, noise, and mutual occlusion between objects, the depth map of the reference image at the coarsest level is discontinuous, affecting the depth map D_0 of the inferred reference view and resulting in low reconstruction completeness. This research suggests employing depth normal consistency to improve the continuity of the predicted depth map D_{L+1} so that multi-scale 3D convolution gives useful context information for depth residual estimate, based on the orthogonality between normal and local surface tangent, as illustrated in Fig. 2.

First step: To get the normal of each central point, one must first figure out where it is neighbor and how much weight it has. In this work, eight nearby sites are chosen to deduce the normal vector of the central point P_i, forming a set of neighborhood coordinates of the central points to. The central point P_i(P is the camera coordinate system coordinate, and P is the pixel coordinate system coordinate) can be identified if the depth Z_i and camera internal parameter matrix $\boldsymbol{K}$ are known. Because $\overline{P_iP_{ix}}$ and $\overline{P_iP_{iy}}$ orthogonality, the central point normal vector $\boldsymbol{N}_i$ may be computed using a cross product as follows:

$$\overline{\boldsymbol{N}_i} = \overline{P_iP_{ix}} \times \overline{P_iP_{iy}} \tag{3}$$

In order to increase the credibility, the normal vector in this paper is averaged over 8 neighborhoods $\boldsymbol{N}_i = \frac{1}{8}\sum_{i=1}^{8}\overline{\boldsymbol{N}_i}$.

Second step: The final optimized depth map can be produced from the normal depth map and the beginning depth map. Each pixel $\boldsymbol{p}_i(x_i, y_i)$ should be refined to the depth of its neighbor pixel points $P_{neighbor}$. Assume that the camera's internal parameter matrix is K, the depth is Z_i, the camera coordinate system's corresponding point is P, the normal vector $\overline{\boldsymbol{N}_i}\left(\boldsymbol{n}_x, \boldsymbol{n}_y, \boldsymbol{n}_z\right)$ infers the depth of nearby points P_i, and the calculation algorithm is

$$\left(\boldsymbol{K}^{-1}Z_i - \boldsymbol{K}^{-1}D_{neighbor}P_{neighbor}\right)\begin{bmatrix}\boldsymbol{n}_x\\ \boldsymbol{n}_y\\ \boldsymbol{n}_z\end{bmatrix} = 0 \tag{4}$$

Weights are used to make the depth more consistent with geometry due to the discontinuity of normal vectors on some edges or irregular surfaces. The weight $W_i = e^{-\partial_1 |\nabla I_i|}$ are determined by the gradient between P_i and $P_{neighbor}$, with the bigger the gradient, the lower the depth optimization's dependability. Because this study calculates the depth of eight neighborhoods, the weight is determined as $W_i^{'} = W_i / \sum_{i=1}^{8} W_i$. The weighted total of depth in eight distinct directions is the depth $\overline{D}_{neighbor}$ after adding depth normal consistency refinement, and the calculation formula is as follows:

$$\overline{D}_{neighbor} = \sum_{i=1}^{8} W_i^{'} D_{neighbor} \tag{5}$$

This improves the continuity of the initial depth map.

The depth D_0 of the reference image is determined iteratively, starting with the depth estimation of the L + 1 layer to obtain the depth map D_L of the preceding layer, and ending with the depth D_0 of the lowest reference image. Firstly, D_{L+1} is up-sampled upper layer by bicubic interpolation, and sample $\uparrow \widetilde{D}_{neighbor}$ is obtained. Then the cost volume is constructed and the residual depth chart ΔD_L of D_L is obtained by regression method, and the iteration depth of layer L is $D_L = \uparrow \widetilde{D}_{neighbor} + \Delta D_L$. In this fashion, the depth of the following layer is refined iteratively until the final refined depth D_0 is attained.

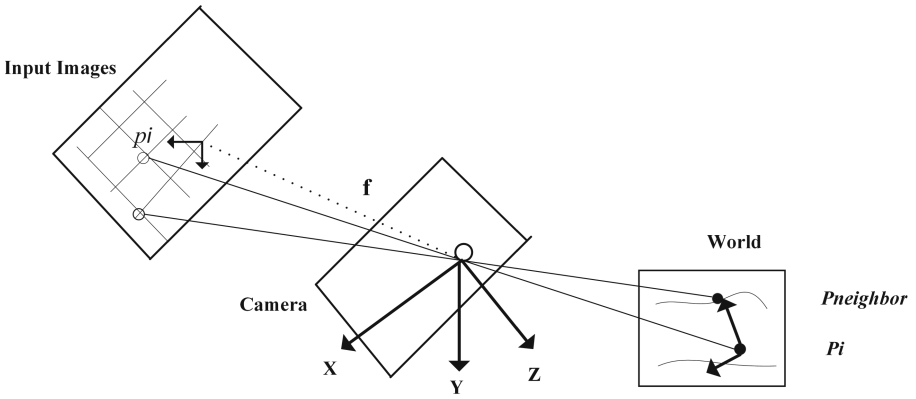

Fig. 2. Normal depth consistency

3.4 Loss Function

The loss function is used to evaluate the difference between predicted value of the model and its actual value, and it is crucial to the model's performance. The loss function is mostly used to restrict the pixel layer information in supervised learning MVS to ensure texture detail matching. Pixel-level constraints, on the other hand, contain several limits, such as illumination and image translation, which will result in pixel alterations.

Supervised learning technique is used to train the loss function in this paper. Because the human visual system perceives a scene through features rather than single pixels, and semantic information can mitigate obstacles in non-ideal areas (such as low texture, etc.) to some extent, this paper uses the weighted sum of pixel loss and feature loss as the loss function, with the following calculation formula.

$$L_{oss} = \beta_1 L_{xsss} + \beta_2 L_{tzss} \tag{6}$$

Feature loss helps stabilize training, improve the network robustness, and improve the reconstruction accuracy.

3.4.1 Pixel Loss

The difference between the predicted image pixel and the actual image pixel is calculated by the pixel loss. The pixel loss in this paper takes into account both the thinned initial depth map and the residual iterative depth map at the same time, and the difference between the true and estimated depths is calculated using the norm. The loss function is defined as the weighted sum of the residual iterative loss for each training sample, and the computation procedure is

$$L_{xsss} = \sum_{i=0}^{L} \sum_{m \in \mathbf{\Omega}} \left\| D_{GT}^{L}(p) - D^{L}(p) \right\| 1 \tag{7}$$

3.4.2 Loss of Features

For two identical images, such as the same image moving one pixel or using different resolutions, feature loss is utilized. Despite the similarity of the images, the pixel loss will produce a big error value, but the feature loss can sense the image from a higher dimension, minimizing the error output. The feature loss is calculated in this research using the features extracted by the pre-trained VGG16 network. Because the VGG16 network's model parameters are enormous, the transfer learning approach is used to transfer the taught weights to this network as the model's initial parameters, avoiding the need to train a significant quantity of data from scratch and thereby enhancing the model's training speed. In this paper, the VGG16 layer 4, 8 and 11 feature map outputs are taken and for each feature, the feature loss is constructed based on the concept of crossed multiple views, using the corresponding pixel $p_i^{'}$ in F_{src}. The feature matching expression from the reference image feature F_{ref} to the source image feature F_{src} is estimated. The loss L_F is calculated as

$$\begin{aligned} F_{src}^{'} &= F_{src}\left(p_i^{'}\right) \\ L_F &= \frac{1}{m} \sum \left(F_{ref} - F_{src}^{'}\right) \times M \end{aligned} \tag{8}$$

The final feature loss is the weighted sum of features of different scales, and the features of the 4th, 8th, and 11th layers of the network are taken as the feature loss of this study,

where M denotes the total number of masks and M indicates the total number of effective points in each mask.

$$L_{tzss} = \partial_1 L_{F4} + \partial_2 L_{F8} + \partial_3 L_{F11} \tag{9}$$

L_{F4} represents the feature loss of layer 4 in the pre-trained network VGG16, $\partial_1, \partial_2, \partial_3$ is the weight coefficient, the value of the weight coefficient can be adjusted to control the degree of influence of the initial depth map and iterative refinement depth map on the network training, in this paper $\partial_1, \partial_2, \partial_3$ is set to 0.1, 0.5, 0.5.

4 Experiment and Result Analysis

4.1 Datasets and Parameters Setting

The DTU dataset [5] contains 124 different scences, each one was taken at 49 or 64 different angles. The image resolution is 1600 × 1200, and there are seven different lighting conditions ranging from orientation to diffusion. The network is implemented on a Linux system, with Pytorch 1.4. 0, Python 3.6, and Python 2.7 as deep learning frameworks, a GPU of NVIDIA RTX 3090Ti, and point cloud visualization using OpenCV. There were 27097 (49 × 7 × 49) images trained and 7546 (49 × 7 × 22) images evaluated in this study. The depth map is also set to 1600 1152, as is the resolution of the training and testing input images. Each reference image is evenly sampled on these virtual planes with a sampling interval of 1 and batch size of 1 on 192 depth virtual planes that are up-sampled from 425 mm to 935 mm. Using the Adam optimizer, iteratively train 28 epochs. The first epoch's learning rate is set to 0.001, while the sixth, twelveth, eighteenth, and twentieth epochs are multiplied by 0.2. Each training uses one reference image and two source images, for a total of three views each iteration.

4.2 Results and Analysis

The trained network is tested with the DTU test set, and the training process is the same as in CVP-MVSNet [9]. The network reconstruction performance is evaluated using three quantitative indicators provided by the DTU data set: calculation accuracy, completeness, and overall. The lower the value for these three indexes, the better the algorithm reconstruction quality.

This work compares classic approaches like Furu [20], Tola [21], Camp [3], Gipuma [22], Colmap [15] with learning-based SurfaceNet [23] MVSNet [6], P-MVSNet [24], Point-MVSNet [8], and CVP-MVSNet [9]. In terms of completeness and integrity, the approach presented in this work outperforms the old algorithm, as demonstrated in Table 1. The overall improved by 11.4% and the completeness by 19.3% when compared to MVSNet [6], while the overall index climbed by 2.8% and the completeness increased by 7.7% when compared to Point-MVSNE [8]. The overall index increased by 0.3%, while the completeness increased by 7.2% as compared to CVP-MVSNE [9]. Figure 3 depicts a portion of the scene depth map and reconstruction representations. The suggested method not only entirely reconstructs the target object, but also has superior reconstruction fidelity than existing methods, according to experimental results.

Table 1. Comparison of test results of different methods

Method	Overall (mm)	Accuracy (mm)	Completeness (mm)
Furu	0.777	0.613	0.941
Tola	0.766	0.342	1.190
Camp	0.695	0.835	0.554
Gipuma	0.578	*0.283*	0.873
Colmap	0.532	0.400	0.664
SurfaceNet	0.745	0.450	1.040
MVSNet	0.462	0.396	0.527
P-MVSNet	0.420	0.406	0.434
Point-MVSNet	0.376	0.342	0.411
CVP-MVSNet	0.351	0.296	0.406
CVP-MVSNet*	0.389	0.426	0.352
Our	*0.348*	0.362	*0.334*

Note: the lettering indicates the optimal value, CVP-MVSNet* is the experimental result of using CVP-MVSNet's method on our equipment.

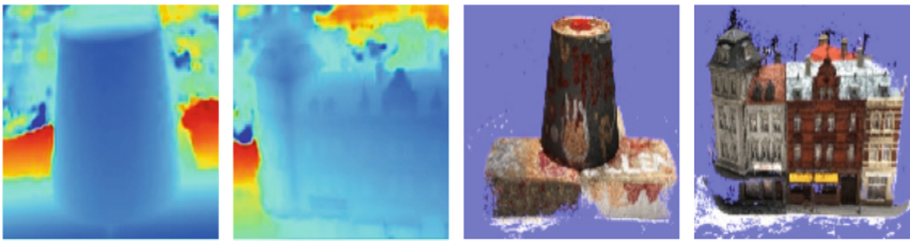

Fig. 3. Example of test results (*From left to right*) depth map of scan1; depth map of scan9; reconstruction effect map of scan1; construction effect map of scan9.

4.3 Ablation Test

Ablation tests and qualitative analysis are carried out for the depth normal consistency module and feature loss module suggested in this study in order to prove the usefulness of this network. The depth normal consistency module and feature loss module are introduced to the system foundation of this article based on CVP-MVSNet [9]. To analyze the advantages of these two modules, four groups of ablation tests were conducted. To assess the quality of the reconstruction, accuracy and completeness indicators are used, while completeness indicators are used to assess its overall performance. Memory use, running time, and model parameters are all kept track of. Table 2 shows the results of the experiment. S stands for the unrefined initial depth module in CVP-MVSNet [9], P for the pixel loss function module, F for the initial depth module corrected by depth normal consistency, and T for the feature loss function module in this article.

For cycling an epoch, the running time is the average of the running times of a single model parameter. The depth normal consistency module is added to the original network CVP-MVDNet [9], and the overall is decline by 1.5%, the completeness is improved by 5.9%, the memory is increased by 398M, and the model parameters are increased by 27063, as shown in Table 2. This is because the normal consistency of depth enhances the quality of the estimated initial depth map, allowing the images at the margins and non-ideal locations to be reconstructed as well, resulting in more complete reconstruction results. The completeness of the original network is improved by 4.4%, the memory is increased by 315M, and the model parameters are increased by 11836 by adding the feature loss module, which is attributed to the fact that the feature loss module retains low-level semantic information such as geometry and texture, which is useful for network supervision and training; by adding the normal depth consistency module and feature loss module to the original network, the completeness is improved by 7.2%, the memory is increased by 315M, and the model parameters are increased in this study, combining two modules yields not only clear and high-quality reconstruction results, but also the ability to reconstruct some edges or small sections, as well as a higher reconstruction completeness and efficiency.

Table 2. Comparison of ablation test results

Method	Overall (mm)	Accuracy (mm)	Completeness (mm)	GPU/M	Time/s	Parameters
S + P	0.351	*0.296*	0.406	*3641*	*0.052*	*55185*
S + P + F	0.366	0.385	0.347	4012	0.053	55864
S + P + T	0.370	0.378	0.362	3956	0.064	56436
S + P + F + T	*0.348*	0.362	*0.334*	4275	0.064	56842

Furthermore, this article compares non-ideal images with CVP-MVSNet [9] to demonstrate the superiority of this method. First, as illustrated in Fig. 4, from the DTU data set, this paper picks 13 scenes with uneven texture distribution for comparative studies. The network reconstruction capabilities of this study is superior than the CVP-MVSNet [9] approach, and there are many points reconstructed at the edge in this paper. Finally, for a comparative experiment, scene 24 with repeating texture is picked from the DTU data set. The points reconstructed via the CVP-MVSNet [9] approach are missing in the lower-left corner of the highest chimney. In comparison to this method, the network reconstruction points in this study are dense, allowing for the restoration of more details and a superior overall reconstruction result. The suggested algorithm's reconstruction in the fine structure is cleaner and has less various points, as evidenced by comparison testing of these scenes.

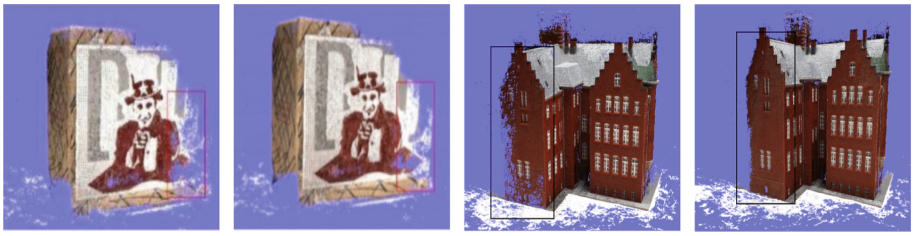

Fig. 4. Comparison of test results of scenes (*From left to right*) The two images on the left is a comparison of scene13 CVP-MVSNet with our results; The two images on the right is a comparison of scene24 CVP-MVSNet with our results.

5 Conclusion

A multi-view stereo reconstruction network with depth normal consistency and depth map refinement is presented based on the depth residual iterative network to alleviate the problem of low reconstruction completeness caused by anomalous and discontinuous initial depth. The normal depth consistency module is used in this paper to improve the quality of the final iterative depth map by refining the initial depth. Simultaneously, the feature loss module is presented to reduce output error ofthe pixel-level loss function and the non-optimal model training owing to image resolution or movement, thereby improving the completeness of multi-view stereo reconstruction. The proposed network has the best completeness according to experimental results on DTU data sets.

The parameters of deep learning neural networks become increasingly complex as the number of layers increases, leading expensive experimental equipments. Future work will focus minimizing running time and memory consumption, and design a lightweight, real-time 3D reconstruction system.

References

1. Liu, J.G.: Three-dimensional reconstruction of multi-view images of movable cultural relics. J. Archeol. **12**, 97–103 (2016)
2. Seitz, S.M., Curless, B., Diebel, J., et al.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp: 519--528. IEEE Press, New York (2006)
3. Campbell, N.D.F., Vogiatzis, G., Hernández, C., Cipolla, R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5302, pp. 766–779. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88682-2_58
4. Merrell, P., Akbarzadeh, A., Wang, L., Mordohai, P.: Real-time visibility-based fusion of depth maps. In: Proceedings of the 2007 IEEE International Conference on Computer Vision, pp: 1--8. IEEE Press, Brazil (2007)
5. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E.: Large scale multi-view stereopsis evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp: 406--413. IEEE Press, Columbus (2014)
6. Yao, Y., Luo, Z., Li, S., Fang, T.: MVSNet: depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision, pp: 767--783. Springer Press, Munich (2018)

7. Yao, Y., Luo, Z., Li, S., et al.: Recurrent MVSNet for high-resolution multi view stereo depth inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp: 5525--5534. IEEE Press, Long Beach (2019)
8. Chen, R., Han, S.F., Xu, J., Su, H.: Point-based multi-view stereo network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp: 1538--1547. IEEE Press, Seou (2019)
9. Yang, J., Mao, W., Alvarez, J. M., Liu, M.: Cost volume pyramid based depth inference for multi-view stereo. In: Proceedings of the IEEE/CVF Conference Computer Vision and Pattern Recognition, pp: 4877--4886. IEEE Press, Seattle (2020)
10. Ye, C.K., Wan, W.G.: Multi-view depth estimation based on feature pyramid network. J. Electr. Measur. Technol. **11**, 91–95 (2020)
11. Kutulakos, K.N., Seitz, S.M.: A theory of shape by space carving. J. Int. J. Comput. Vis. **38**(3), 199–218 (2000)
12. Pons, J.P., Keriven, R., Faugeras, O., Hermosillo, G.: Variational stereovision and 3D scene flow estimation with statistical similarity measures. In: Proceedings Ninth IEEE International Conference on IEEE Computer Vision, p: 597. IEEE Press, Nice (2003)
13. Esteban, C.H., Schmitt, F.: Silhouette and stereo fusion for 3D object modeling. J. Comput. Vis. Image Underst. **96**(3), 367–392 (2004)
14. Kang, S.B., Szeliski, R., Chai, J.: Handling occlusions in dense multi-view stereo. In: Proceeding of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp:1. IEEE Press, Kauai (2001)
15. Schonberger, J.L., Frahm, J.M.: Structure from motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp: 4104--4113. IEEE Press, Las Vegas (2016)
16. Kar, A., Hane, C., Malik, J.: Learning a multi-view stereo machine. J. Adv. Neural Info. Process. Syst. **30**, 365—376 (2017)
17. Gu, X., Fan, Z., Zhu, S., Dai, Z.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp: 2495--2504. IEEE Press, Seattle (2020)
18. Yang, Z., Wang, P., Xu, W.: Unsupervised learning of geometry with edge-aware depth-normal consistency. J. arXiv preprint arXiv. (2017)
19. Johnson, J., Alahi, A., Li, F.F.: Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision, pp:694--711. Springer Press, Amsterdam (2016)
20. Furukawa, Y.,Ponce, J.: Accurate, dense, and robust multi view stereopsis. J. IEEE Trans. Pattern Anal. Mach. Intell.. **32**(8), 1362—1376 (2009)
21. Tola, E., Strecha, C., Fua, P.: Efficient large-scale multi-view stereo for ultra high-resolution image sets. J. Mach. Vis. Appl. **23**(5), 903–920 (2012)
22. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multi-view stereopsis by surface normal diffusion. In: Proceedings of the IEEE International Conference on Computer Vision, pp: 873--881. IEEE Press, Santiago (2015)
23. Ji, M., Gall, J., Zheng, H., Liu, Y.: Surfacenet: An end-to-end 3D neural network for multi-view stereopsis. In: Proceedings of the IEEE International Conference on Computer Vision, pp: 2307--2315. IEEE Press, Venice (2017)
24. Luo, K., Guan, T., Ju, L., Huang, H.: P-MVSNet: learning patch-wise matching confidence aggregation for multi-view stereo. In: Proceeding of the 2019 IEEE/CVF International Conference on Computer Vision, pp: 10452--10461. IEEE Press, Long Beach (2019)