IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 22, 2021, accepted February 8, 2021, date of publication February 11, 2021, date of current version February 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3058522

PA-MVSNet: Sparse-to-Dense Multi-View Stereo With Pyramid Attention

KE ZHANG[1], MENGYU LIU[1], JINLAI ZHANG[2], AND ZHENBIAO DONG[1]
[1]School of Mechanical Engineering, Shanghai Institute of Technology, Shanghai 201418, China
[2]College of Mechanical Engineering, Guangxi University, Nanning 530004, China

Corresponding authors: Mengyu Liu (freelmy7579@163.com), Jinlai Zhang (228434973@qq.com), and Zhenbiao Dong (dzb0312@126.com)

The project was supported by Innovation Project of Guangxi Graduate Education (YCBZ2021019).

ABSTRACT Multi-view based 3D reconstruction aims to obtain 3D structure information of objects in space through two-dimensional images. In this paper, we propose a new multi-view stereo network that can robustly reconstruct the scene. To enhance the feature representation ability of Point-MVSNet, a pyramid attention module is introduced. Specifically, we exploit the attention mechanism for the multi-scale feature pyramid to capture larger receptive fields and richer information. Instead of constructing a feature pyramid as the input, results of the pyramid attention module at different scales are directly used for the next layer. The network eventually generates a high-quality depth estimation for 3D reconstruction from sparse to dense by an iterative refinement schedule. Experiments have been performed to evaluate 3D reconstruction quality by comparison with existing state-of-the-art methods on the DTU dataset. The experimental results indicate our method performs the best in overall quality compared with previous methods, proving the effectiveness of our method. In the end, we use the data collected by mobile devices to implement 3D reconstruction with a combination of traditional and learning-based methods, providing ideas for the 3D reconstruction technology on mobile devices.

INDEX TERMS Multi-view stereo, pyramid attention, point cloud, depth estimate, deep learning.

I. INTRODUCTION

Multi-view stereo (MVS), which intends to reconstruct a complete 3D representation of an object or scene from a series of images taken from known camera viewpoints, has been developed and obtained tremendous success as a division of computer vision for decades [1]. The emergence of a large number of consumption level data acquisition tools promotes the application of 3D reconstruction technology and serves different practical needs. Traditional MVS technique (e.g., structure from motion) that extracts handcrafted features in images and recovers 3D structure has been proved a great success in recent MVS benchmarks [2]–[5]. Even so, due to the limitations of handcrafted features and methods, there are some shortages of texture-less surfaces, computing consumption, and so on.

Recently, the success of the convolutional neural networks in various computer vision tasks has promoted the improvement of MVS methods and stimulated the interest of research on this topic. Eigen *et al.* [6] firstly proposed the application of convolutional neural network for monocular depth estimation, dividing the network into two modules for coarse global prediction and local refinement prediction respectively and defining loss function by scale-invariant error, which provided ideas and guidance for further research on learning-based 3D reconstruction methods. Then some end-to-end neural networks are designed to predict the depth of scenes directly from a sequence of images (e.g., MVSNet [7] and R-MVSNet [8]). Even though the accuracy of these methods has been verified on various datasets, most of them utilize 3D CNNs to predict depth maps or voxel occupancy, leading to excessive memory consumption and limiting the improvement of resolution. Compared with other 3D data formats, the structure of the point cloud is simpler and easier to be processed, so it becomes one of the research emphases of 3D reconstruction. Considering the advantages of the point cloud in contrast to other 3D representations, a point-based multi-view stereo network (Point-MVSNet) is proposed by Chen *et al.* [9], not only processing point cloud directly but also fusing depth and texture information for feature enhancement. However, during 2D-3D information fusion, there exists a few deficiencies such as narrow receptive fields,

The associate editor coordinating the review of this manuscript and approving it for publication was Utku Kose.

insufficient use of global context information, and time consuming. As the attention mechanism becomes popular, researchers try to apply it to solve related computer vision problems [10].

In this paper, we introduce the attention mechanism to extract richer high-level features and long-range feature correspondences from the feature pyramid without too much computation burden. Specifically, we perform convolutions to downsample the input images to get a multi-scale feature pyramid firstly. A scale agnostic attention module [11] which takes full advantage of self-similarities is exploited to capture long-range feature correspondences in the top-down pathway. The improved structure of the feature pyramid is based on the classic feature pyramid network (FPNs) [12] which applies ResNet [13] as a backbone to various tasks. The improved network is tested on the DTU dataset which is standard multi-view stereo benchmarks. Experiments show that compared with previous start-of-the-art MVS reconstruction, ours obtains better results on the overall quality. In addition, we implement experiments on the data collected by mobile devices. A 3D reconstruction pipeline is used to obtain the camera poses as the input. The reconstruction results show better quality than traditional methods.

II. RELATED WORKS

A. TRADITIONAL MULTI-VIEW STEREO RECONSTRUCTION

According to the object models, MVS algorithms can be roughly divided into four categories: voxel-based [14]–[17], deformable polygonal mesh-based [18]–[20], patch-based [21]–[23], and depth map-based approaches [24]–[29]. Recent results on MVS benchmarks have demonstrated that the depth map-based method is the most arcuate and robust of the above. At present, MVS algorithms referring to the classical framework of parallax calculation often perform depth estimation by cost volume of construction and aggregation, coarse estimation of depth maps, and multi-view refinement. An open-source MVS implementation named COLMAP proposed by Schonberger and Frahm [30] offers a wide range of features for the reconstruction of ordered and unordered image collections. Open Multiple View Geometry (OpenMVG) [31] is a well-known open-source library that deals with multi-view solid geometry, providing feature extraction and matching methods and a complete toolchain for structure from motion. While OpenMVG can recover camera poses and a sparse 3D point cloud from an input set of images, there is none addressing the last part of the photogrammetry chain-flow. Then open Multi-View Stereo reconstruction library (OpenMVS) is presented, aiming at filling that gap by providing a complete set of algorithms to recover the full surface of the scene to be reconstructed. However, they suffer from the texture-less region and utilizing cross-scale features.

B. LEARNING-BASED MULTI-VIEW STEREO

Recently, an increasing number of researches on MVS reconstruction applying the convolution neural network has achieved remarkable progress. Ji *et al.* [32] firstly propose an end-to-end learning framework named SurfaceNet for multi-view stereo where both photo-consistency as well as geometric relations of the surface structure can be directly learned. Inspired by Long Short-Term Memory (LSTM) [33], Choy *et al.* [34] propose a 3D recurrent reconstruction neural network (3D-R2N2), which extends the standard LSTM framework to build the mapping of 2D graphics to 3D voxel, completing the 3D reconstruction of single or multiple views (input from multiple views will be treated as a sequence of input to the LSTM). However, there is a problem that to improve the accuracy requires to improve the resolution but the increase of resolution will greatly increase the calculation time. Then MVSNet and DPSNet [35] are proposed which use a differential warping process to construct a cost volume and regress the depth map from the cost volume. Furthermore, R-MVSNet based on the recurrent network utilizes the gate recurrent unit (GRU) rather than 3D-CNN to regularize the cost volume, reducing memory consumption effectively. In addition, Point-MVSNet converts the coarse depth map generated by MVSNet into point cloud and refines the point cloud with the fashion of depth residual prediction between the current iteration and the ground truth, avoiding too much burden in 3D CNN computation.

C. PYRAMID ATTENTION

The feature pyramid is often used for object detection at different scales as a basic component. When recognizing objects of Large size differences, a classic approach is to enhance the multi-scale variation by image pyramids. While this method can generate multi-scale feature representations via feature extraction at every scale and produce feature maps of abundant semantic information, it would greatly increase the time consumption required and make it impractical to train an end-to-end deep neural network in the form of image pyramids. Even though common object detection networks often exploit a single high-level feature map for prediction like Fast R-CNN [36], it remains a problem that low-level feature maps are poor of the semantic information as well as the resolution. Feature pyramid networks (FPNs), solving the above shortcomings, can combine low-resolution feature maps that have rich semantic information with high-resolution feature maps that have poor semantic information under the premise of adding less computation. Pyramid attention networks are the application of attention mechanisms on the feature pyramid. Li *et al.* [37] apply feature pyramid attention (FPA) module for semantic segmentation to learn a better and richer feature representation via performing spatial pyramid attention structure on high-level output. In addition, Ren *et al.* [38] propose a pyramid self-attention module (PSAM) for salient object detection to capture a richer high-level feature and enlarge the receptive field of the model.

D. NON-LOCAL ATTENTION

Non-local means is an effective algorithm widely used in image denoising tasks [39]. Compared with local means, non-

local mean filtering, utilizing self-similarity prior to reduce corruptions, can efficaciously remove the noise while preserving the image edge details. Inspired by the non-local means, non-local operation [40] is presented to capture long-range dependencies of one-dimensional temporal signals, images, and video sequences. The core idea of non-local attention is calculating the correlation between the block of pixels to be computed and rest blocks, highlighting concerned areas, and eliminating noise effectively. Non-local attention has a tremendous performance in other vision tasks, such as object detection [41], image restoration [42], [43], semantic segmentation [44], [45] and person re-identification [46], [47]. For MVS reconstruction, the application of non-local attention has not been found yet.

However, the non-local operation only takes processing information at the same scale into account simply, both ignoring advantages of multi-scale feature fusion and involving mismatches in attention units when performing pixel-wise feature matching as well. In order to solve this problem, scale agnostic attention is designed to capture correspondences between two scales so that non-local attention can be applied to the feature pyramid.

III. METHOD

A. PROBLEM DEFINITION AND GOALS

The current deep-learning based MVS methods have not been integrated with the self-attention mechanism to improve the accuracy of results although it has achieved the state of the art in various benchmarks of object detection tasks. In addition, it is difficult to capture large-scale features and obtain images of high resolution. In this work, we focus on applying self-attention to MVS neural network in order to obtain larger and adaptive receptive fields, higher accuracy, and more complete and precise point cloud. Meanwhile, we explore the application of MVS reconstruction on mobile devices. Random indoor objects are photographed and reconstructed by MVS network, which ensures the completeness and accuracy of the point cloud while achieving more convenient and fast reconstruction.

The improved neural network is introduced in this section. We elaborate on the principle of pyramid attention and scale agnostic attention firstly. Then the architecture of our network is introduced and illustrated. The complete network framework is shown in Figure 1.

B. PYRAMID ATTENTION MODULE

1) FORMAL DEFINITION

Following the non-local mean filtering, the non-local operation is defined as:

$$\mathbf{y}_i = \frac{1}{C(\mathbf{x})} \sum_{\forall j} f\left(\mathbf{x}_i, \mathbf{x}_j\right) g\left(\mathbf{x}_j\right) \tag{1}$$

where i,j are indices of the input $\mathbf{x}$ and output $\mathbf{y}$ respectively. In addition,$\mathbf{x}$ and $\mathbf{y}$ have the same size. The function f computes pair-wise affinity between $\mathbf{x}_i$ and $\mathbf{x}_j$. The feature transformation function g calculates a new representation of $\mathbf{x}_j$. Then the output response is normalized by a scalar function $C(\mathbf{x})$. Symbol $\forall j$ in formula (1) indicates that all positions are considered in the non-local operation.

Pyramid attention is applied to solve the problem of scale constraint. Similar to non-local operation, a function computes affinity between a target feature and regions in pyramid attention. Meanwhile, we sum up the weighted input to obtain a response feature over the multi-scale input. Given a set of scale factor $S = \{1, s_1, s_2, \ldots, s_n\}$, pyramid attention can be obtained by formula (2):

$$\mathbf{y}_i = \frac{1}{C(x)} \sum_{s \in S} \sum_{\forall j} f\left(\mathbf{x}_i, \mathbf{x}_j^{\delta(s)}\right) g\left(\mathbf{x}_j^{\delta(s)}\right) \tag{2}$$

here $\delta(s)$ is the neighborhood of s^2 centered on input $\mathbf{x}_j$. The formula (2) degrades the previous non-local operation as shown in formula (1) under the situation of only one scale factor $s = 1$.

2) SCALE AGNOSTIC ATTENTION

In this section, non-local operation is extended to two scales. Given two scale factors s_1, s_2, it is critical to evaluate the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j^{\delta(s)}$, where $\mathbf{x}_j^{\delta(s)}$ is used to aggregate information to achieve $\mathbf{y}_i$. While usual similarity measurements such as Gaussian or embedded Gaussian have been achieved great results, it is impracticable to apply these methods to features with different dimensions. We reduce the scale of region $\mathbf{x}_j^{\delta(s)}$ in a pixel feature $\mathbf{z}_j$, so that the spatial information of $\mathbf{x}_j^{\delta(s)}$ could be squeezed into a single region descriptor. Additionally, a descriptor map $\mathbf{z} = \left(\frac{H}{s} \times \frac{W}{s}\right)$ is obtained by down-scaling the original input $\mathbf{x}(H \times W)$ for search over the entire feature map. Then scale agnostic attention can be represented by using $\mathbf{x}_i$ and $\mathbf{z}_j$ to describe the correlation between $\mathbf{x}_i$ and $\mathbf{x}_j^{\delta(s)}$, as shown in formula (3):

$$\mathbf{y}_i = \frac{1}{C(\mathbf{x}, \mathbf{z})} \sum_{\forall j} f\left(\mathbf{x}_i, \mathbf{z}_j\right) g\left(\mathbf{z}_j\right) \tag{3}$$

here $C(\mathbf{x}, \mathbf{z})$ is a scalar factor to normalize the response. In this paper, we choose embedded Gaussian, a simple extension of Gaussian function for function f:

$$f\left(\mathbf{x}_i, \mathbf{z}_j\right) = e^{\theta(\mathbf{x}_i)^T \varphi(\mathbf{z}_j)} \tag{4}$$

where $\theta\left(\mathbf{x}_i\right) = \mathbf{w}_\theta \mathbf{x}_i$, $\varphi\left(\mathbf{z}_j\right) = \mathbf{w}_\varphi \mathbf{Z}_j$ and the scalar factor $C(\mathbf{x}, \mathbf{z})$ is set to $\sum_{\forall j} f\left(\mathbf{x}_i, \mathbf{z}_j\right)$. The framework of scale agnostic attention is shown in Figure 2.

C. THE FRAMEWORK OF NEURAL NETWORK

1) COARSE DEPTH PREDICTION

Because learning-based MVS methods produce a large amount of memory and time consumption when calculating cost volume, we predict a low-resolution cost volume for coarse depth prediction when given the images and corresponding camera parameters. In the coarse prediction network, the cost volume is built by 1/8 the size of the reference images which contains 48 or 96 virtual depth planes for training and evaluation.

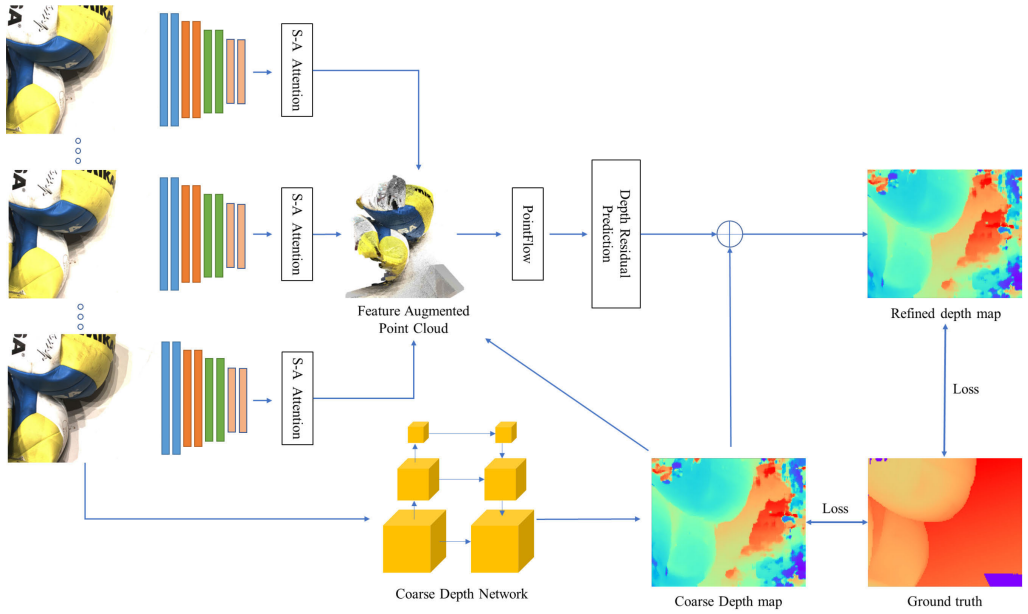


FIGURE 1. The main architecture of our neural network. A pyramid attention module is added before the feature augmented point cloud.

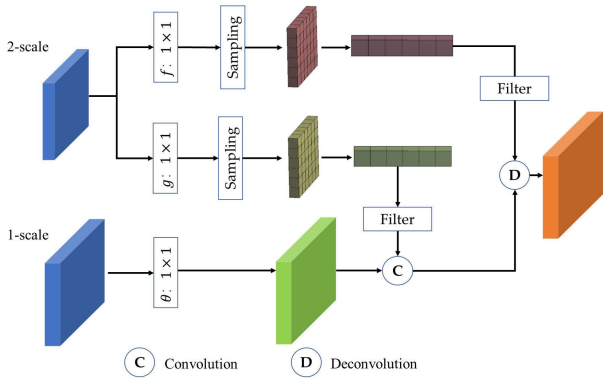


FIGURE 2. The framework of scale agnostic (S-A) attention. The S-A operation is performed between two feature maps of different scales.

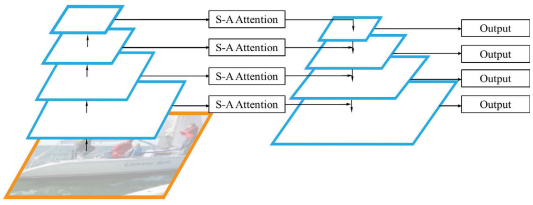


FIGURE 3. The framework of pyramid attention module. The scale agnostic attention operation is added and we capture outputs from different scales in the top-down pathway of the feature pyramid.

2) 2D-3D FEATURE ENHANCEMENT

Studies shows that it is important to take advantage of learning-based image features to improve dense pixel correspondence quality. In order to enlarge the receptive fields of points and make them have more multi-scale contextual information, we construct a 4-scale feature pyramid. We apply 2D convolution to downsample the feature map and finally feature pyramid is expressed as $\mathbf{F}_i = \left[\mathbf{F}_i^1, \mathbf{F}_i^2, \mathbf{F}_i^3, \mathbf{F}_i^4\right]$ for image $\mathbf{I}_i$.

As shown in Figure 3, feature maps at different levels are manipulated by a scale agnostic module. Instead of capturing the result in the down-top pathway, we fetch the feature maps from four different levels in the feature pyramid as the input of next layer. In addition, feature pyramids are shared among all the input. We use a differential unprojection to obtain image appearance features of each point from the multi-view feature maps. The camera intrinsic matrix is transformed at different level of feature maps for feature warp because the image resolution of features $\mathbf{F}_i^1, \mathbf{F}_i^2, \mathbf{F}_i^3, \mathbf{F}_i^4$ are various. A variance-based cost metric, aggregating features from an arbitrary number of views, is calculated by following formula:

$$\mathbf{C}_j = \frac{\sum_{i=1}^{N} \left(\mathbf{F}_i^j - \overline{\mathbf{F}}^j\right)^2}{N}, (j = 1, 2, 3, 4) \tag{5}$$

here j represents the different level of pyramid features. Considering the normalized 3D coordinates in world space $\mathbf{X}_p$, we conduct a concatenation as shown in formula (6):

$$\mathbf{C}_p = \text{concat}\left[\mathbf{C}_p^j, \mathbf{X}_p\right], \quad (j = 1, 2, 3, 4) \tag{6}$$

considering that the entire network iteratively predicts the depth residual, the position is updated after each iteration.

3) ITERATIVE REFINEMENT

A coarse depth map generated by coarse depth prediction network need to be refined iteratively because of the low resolution of 3D cost volume. Given the camera parameters, we convert the depth map to a point cloud by unprojection. Following the Point-MVSNet, we generate a sequence of

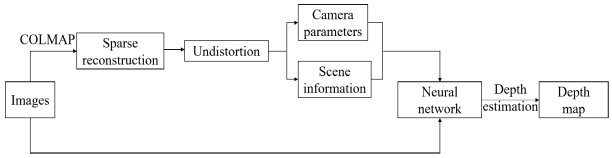


FIGURE 4. The schematic diagram of processes for depth estimation of images taken by mobile devices.

point with different displacement along the reference camera. Inspired by dynamic graph CNN (DGCNN) [48], edge convolution is used to enrich feature aggregation between neighboring points. At the step of flow prediction, we use four edge convolution layers to aggregate point features at different scales of the neighborhood for obtaining a depth residual map. Then the output is sent to the initial depth map for depth refinement. During iterative refinement with upsampling, we use nearest neighbor unsample the depth map $\mathbf{D}^{(i)}$ and obtain $\mathbf{D}^{(i+1)}$ by flow prediction.

4) TRAINING LOSS

we use L_1 loss that measure the absolute difference between the ground truth depth map and the estimated depth map as our training loss. Meanwhile, we take the initial depth map and iteratively refined ones into account. The training loss is calculated in formula (7):

$$\text{Loss} = \sum_{i=0}^{l} \left(\frac{\lambda^{(i)}}{S^{(i)}} \sum_{p \in \mathbf{P}_{\text{valid}}} \left\| \mathbf{D}_{GT}(p) - \mathbf{D}^{(i)}(p) \right\|_1 \right) \quad (7)$$

here $\mathbf{P}_{\text{valid}}$ denotes the valid ground truth pixel set and l is the iteration number, while parameter $\lambda^{(i)}$ is set to 1.0 in experiments.

D. EXPERIMENTS ON SELF-COLLECTED DATA

We design an experiment on raw image sequences without any data preprocessing. In this work, traditional MVS reconstruction method is utilized to implement sparse reconstruction and obtain relevant parameters needed by the neural network to predict. Then these information and images are taken as a scan of dataset for prediction and the depth estimation of the scene can be calculated by the neural network. The complete experimental process is shown in Figure 4.

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

1) TRAINING

We train our neural network on the DTU dataset, a large-scale MVS dataset which consists of 124 different indoors scenes with different lighting conditions and is split into training, validation and evaluation sets. Our neural network is trained on the training set and make an evaluation on the evaluation set. We generate depth maps from the given ground truth as in MVSNet for data pre-processing. Similar to Point-MVSNet, the input image resolution is set to $W \times H = 640 \times 512$, and number of views to $N = 3$ during the training stage. The 3D cost volume, sampled from 425mm to 921mm, with $D = 48$ depth planes is to construct for coarse prediction. We set flow iteration $l = 2$ and depth intervals 8mm and 4mm for depth refinement. Meanwhile, the number of nearest neighbor points is set to 16.

Our network is implemented on Pytorch [49] and trained end-to-end using RMSProp with an initial learning rate 0.0005 which is decreased by 0.9 for every 2 epochs. The coarse prediction step is trained for 4 epochs separately and the model is trained for another 12 epochs. Batch size is set to 4 on 2 NVIDIA RTX 2080Ti graphics cards.

2) EVALUATION

We set image view number $N = 5$ and $D = 96$ depth layers for initial prediction. Meanwhile, flow iteration is set to $l = 3$ for depth refinement. Following the same approach of post-processing as in MVSNet, we fuse all depth maps to point clouds and the input image resolution is set to 1280×960.

TABLE 1. Quantitative results of reconstruction quality dataset (lower is better).

Methods	Acc. (mm)	Comp. (mm)	Overall (mm)
Camp	0.835	0.554	0.695
Furu	0.613	0.941	0.777
Tola	0.342	1.190	0.766
Gipuma	**0.283**	0.873	0.578
Colmap	0.400	0.664	0.532
SurfaceNet	0.450	1.040	0.745
MVSNet	0.396	0.527	0.462
R-MVSNet	0.385	0.459	0.422
Point-MVSNet	0.361	**0.421**	0.391
Ours	0.313	0.437	**0.375**

B. BENCHMARKS RESULTS

We evaluate our network on the DTU evaluation set. Quantitative results are shown in Table 1. The accuracy and completeness are calculated using the official code provided by the DTU dataset. To evaluate the overall reconstruction quality, the overall score is computed by the average of the accuracy and the completeness. The accuracy of our approach is 0.313, which is better than other MVSNet and Point-MVSNet, although our completeness is 0.437, 0.016 higher than Point-MVSNet. While Gipuma performs the best regarding to accuracy, our method performs the best in overall quality compared with previous methods which a score of 0.375. Qualitative results are shown in Figure 5. Our method generates a more complete and detailed point cloud.

C. ABLATION STUDY

In this section, we provided ablation experiments for quantitative and qualitative analysis to evaluate the strength of the key components in our work. For following studies, experiments are implemented and evaluated on the DTU dataset. Moreover, both accuracy and completeness are used to evaluate the performance of our network.

1) SCALE LEVELS

In this part, we investigate the influences of pyramid levels to verify the effectiveness of pyramid attention. We conduct

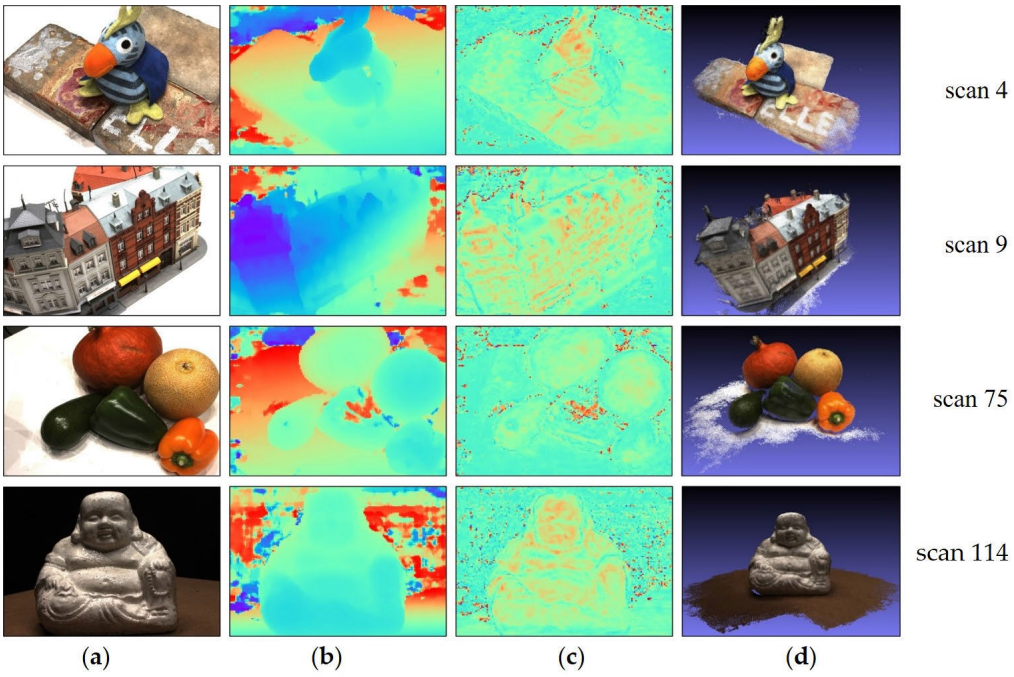


FIGURE 5. Illustration on predicted depth map, probability map and point cloud representation. (a) One reference image of scan 4, 9, 75 and 114 of DTU dataset; (b) the predicted depth map; (c) the probability map; (d) the generated point cloud by depth map fusion for post-processing.

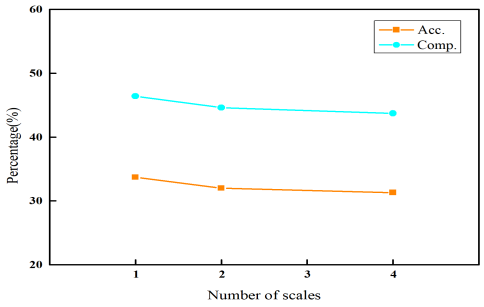


FIGURE 6. The results of the accuracy and completeness with different pyramid scales on the DTU dataset.

control experiments by adding levels to the feature pyramid with 3mm threshold. The final pyramid consists of four scales. The quantitative results are shown in Figure 6, which shows that the reconstruction quality partly is optimized with the increase of pyramid levels.

2) PYRAMID ATTENTION

he key difference between the classic non-local operation and pyramid attention is that our module allows the network to utilize correspondences at multiple scales. To verify the effectiveness of our network, We construct a network with non-local operation by replacing the scale agnostic module. The reconstruction quality The quantitative results are shown in Table 2, which demonstrates that although the classic non-local operation can improve the reconstruction results compared with original Point-MVSNet, our method shows better.

TABLE 2. Ablation study between non-local attention and scale pyramid attention.

	Acc. (mm)	Comp. (mm)	Overall (mm)
Non-local	0.320	0.446	0.383
Ours	**0.313**	**0.437**	**0.375**

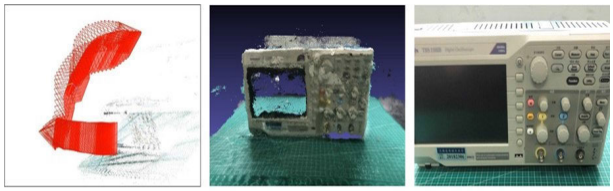


FIGURE 7. Illustration on 3D reconstruction of data collected by mobile devices. (a) The results of sparse reconstruction by COLMAP where camera poses and parameters are computed; (b) the fused point cloud; (c) one of the images taken by mobile devices.

D. PERFORMANCE ON SELF-COLLECTED DATA

We use the iPhone to collect two-dimensional image sequences of indoor objects from different angles and light conditions. General self-collected data only contains RGB information without camera parameters. We utilize COLMAP to conduct a sparse reconstruction, undistortion, calculating the extrinsic and intrinsic parameters and the camera poses during the reconstruction, then we obtain information of camera files and sparse scene as the input of the neural network. Finally, we predict the data and achieve the inferred depth map and point cloud. The reconstruction results are shown in Figure 7.

Both COLMAP and learning-based methods are used in the experiment for comparison. The reconstruction results

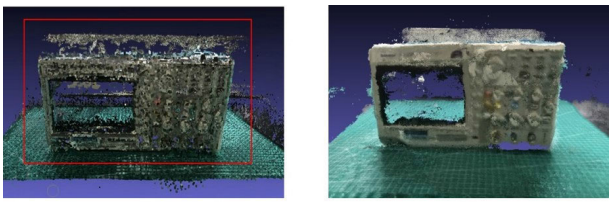


FIGURE 8. The results of point cloud generated by (a) COLMAP and (b) MVS neural network.

of point cloud are shown in Figure 8. It is obvious that the result using traditional method (COLMAP) has more noise and poor quality.

V. CONCLUSION

In this paper, we introduce a pyramid attention module to improve the deep learning architecture based on point cloud for MVS reconstruction. Inspired by the self-attention mechanism, the pyramid attention module captures non-local relationships at multiple scales. Experiments on the DTU dataset show that our proposed neural network performs better than previous methods and produces high-quality point clouds. Additionally, we use the neural network to implement 3D reconstruction on the data collected by the mobile phone and generate the reconstruction point cloud, providing ideas for the development of 3D reconstruction technology on mobile devices.

ACKNOWLEDGMENT

The project was supported by Innovation Project of Guangxi Graduate Education (YCBZ2021019). The authors would like to thank all the participants of the study for their time and useful comments. They appreciate the support from their colleagues and classmates.

REFERENCES

[1] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 519–528.

[2] A.-A. Iddrisu and I. P. Alagidede, "Monetary policy and food inflation in South Africa: A quantile regression analysis," *Food Policy*, vol. 91, Feb. 2020, Art. no. 101816, doi: 10.1016/j.foodpol.2019.101816.

[3] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, Jul. 2017.

[4] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 153–168, Nov. 2016.

[5] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3260–3269.

[6] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.

[7] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 767–783.

[8] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5525–5534.

[9] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1538–1547.

[10] F. Wang and D. M. J. Tax, "Survey on the attention based RNN model and its applications in computer vision," 2016, *arXiv:1601.06823*. [Online]. Available: http://arxiv.org/abs/1601.06823

[11] Y. Mei, Y. Fan, Y. Zhang, J. Yu, Y. Zhou, D. Liu, Y. Fu, T. S. Huang, and H. Shi, "Pyramid attention networks for image restoration," 2020, *arXiv:2004.13824*. [Online]. Available: http://arxiv.org/abs/2004.13824

[12] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[14] S. N. Sinha, P. Mordohai, and M. Pollefeys, "Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[15] A. Hornung and L. Kobbelt, "Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 503–510.

[16] G. Vogiatzis, P. H. S. Torr, and R. Cipolla, "Multi-view stereo via volumetric graph-cuts," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 391–398.

[17] S. Tran and L. Davis, "3D surface reconstruction using graph cuts with surface constraints," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 219–231.

[18] C. H. Esteban and F. Schmitt, "Silhouette and stereo fusion for 3D object modeling," *Comput. Vis. Image Understand.*, vol. 96, pp. 367–392, Dec. 2004.

[19] A. Zaharescu, E. Boyer, and R. Horaud, "TransforMesh: A topology-adaptive mesh-based approach to surface evolution," in *Proc. Asian Conf. Comput. Vis.*, 2007, pp. 166–175.

[20] Y. Furukawa and J. Ponce, "Carved visual hulls for high-accuracy image-based modeling," in *Proc. ACM SIGGRAPH Sketches (SIGGRAPH)*, 2005, pp. 564–577.

[21] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.

[22] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz, "Multi-view stereo for community photo collections," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[23] M. Habbecke and L. Kobbelt, "Iterative multi-view plane fitting," in *Proc. Int. Fall Workshop Vis., Modeling, Vis.*, 2006, pp. 73–80.

[24] E. Tola, C. Strecha, and P. Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Mach. Vis. Appl.*, vol. 23, no. 5, pp. 903–920, Sep. 2012.

[25] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2008, pp. 766–779.

[26] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 873–881.

[27] Y. Yao, S. Li, S. Zhu, H. Deng, T. Fang, and L. Quan, "Relative camera refinement for accurate dense reconstruction," in *Proc. Int. Conf. 3D Vis. (DV)*, Oct. 2017, pp. 185–194.

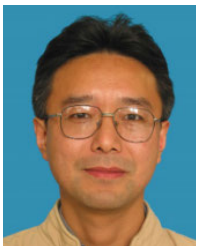
[28] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm, "PatchMatch based joint view selection and depthmap estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1510–1517.

[29] J. L. Schönberger, E. Zheng, J. M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 501–518.

[30] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4104–4113.

[31] P. Moulon, P. Monasse, and R. Marlet, "La bibliothèque openMVG: Open source multiple view geometry," Center Vis. Comput., ENPC, Univ. Paris-Est, LIGM (UMR CNRS), Champs-sur-Marne, France, Tech. Rep., 2013.

[32] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "SurfaceNet: An end-to-end 3D neural network for multiview stereopsis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2307–2315.
[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.
[34] C. B. Choy, D. Xu, J. Y. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 628–644.
[35] S. Im, H.-G. Jeon, S. Lin, and I. S. Kweon, "DPSNet: End-to-end deep plane sweep stereo," 2019, *arXiv:1905.00538*. [Online]. Available: http://arxiv.org/abs/1905.00538
[36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[37] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*. [Online]. Available: http://arxiv.org/abs/1805.10180
[38] G. Ren, T. Dai, P. Barmpoutis, and T. Stathaki, "Salient object detection combining a self-attention module and a feature pyramid network," 2020, *arXiv:2004.14552*. [Online]. Available: http://arxiv.org/abs/2004.14552
[39] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 60–65.
[40] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[41] M. Shokri, A. Harati, and K. Taba, "Salient object detection in video using deep non-local neural networks," *J. Vis. Commun. Image Represent.*, vol. 68, Apr. 2020, Art. no. 102769.
[42] D. Wen, B. Fan, Y. Loy, C. C. Huang, and T. S. Huang, "Non-local recurrent network for image restoration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1673–1682.
[43] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," 2019, *arXiv:1903.10082*. [Online]. Available: http://arxiv.org/abs/1903.10082
[44] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 593–602.
[45] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[46] B. Bryan, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3760–3769.
[47] C.-T. Liu, C.-W. Wu, Y.-C. Frank Wang, and S.-Y. Chien, "Spatially and temporally efficient non-local attention network for video-based person re-identification," 2019, *arXiv:1908.01683*. [Online]. Available: http://arxiv.org/abs/1908.01683
[48] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, Nov. 2019.
[49] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NIPS Autodiff Workshop*, 2017.



KE ZHANG received the Ph.D. degree in mechanical engineering from Donghua University, in 2005. He is currently a Professor with the Department of Mechanical Engineering with the Shanghai Institute of Technology, China. His research interests include precision measurement, and mechatronics and sensor technology.

MENGYU LIU received the B.S. degree from Jimei University, Xiamen, China, in 2016. He is currently pursuing the master's degree with the Shanghai Institute of Technology, Shanghai. His main research interests include image processing and 3D reconstruction.

JINLAI ZHANG received the B.S. degree from the Changsha University of Science and Technology, Changsha, China, in 2017. He is currently pursuing the Ph.D. degree with the College of Mechanical Engineering, Guangxi University. His research interests include 3D deep learning and time series forecasting.

ZHENBIAO DONG received the Ph.D. degree from the School of Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, in 2019. He is currently a Postgraduate Tutor with the Shanghai Institute of Technology. His research interests include material processing, 3D printing, semiconductor materials, and semiconductor device technology.