Applied Intelligence (2023) 53:4289–4301
https://doi.org/10.1007/s10489-022-03754-3

MFNet: Multi-level fusion aware feature pyramid based multi-view stereo network for 3D reconstruction

Youcheng Cai[1] · Lin Li[1] · Dong Wang[1] · Xiaoping Liu[1,2]

Accepted: 10 May 2022 / Published online: 7 June 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

We present an efficient multi-view stereo (MVS) network for 3D reconstruction from multi-view images. While the existing state-of-the-art methods have achieved satisfactory results, the accuracy and scalability remain an open problem due to unreliable dense matching and memory-consuming cost volume regularization. To this end, we propose a multi-level fusion aware feature pyramid based multi-view stereo network (MFNet) for reliable depth inference. First, we adopt a coarse-to-fine strategy that achieves high-resolution depth estimation based on the coarse depth map. This strategy gradually narrows the depth search interval by using the prior information from the previous stage, which dramatically reduces memory consumption. Second, we conduct multi-level fusions to construct the feature pyramid such that the different level features receive information from each other, thus enabling rich multi-level feature representations. Finally, the group-wise correlation similarity measure is introduced to replace the variance-based approach used in previous works for cost volume construction, resulting in a lightweight and effective cost volume representation. Experimental results on the DTU, Tanks & Temples, and BlendedMVS benchmark datasets show that MFNet achieves better results than the state-of-the-art methods.

Keywords Multi-view stereo · Multi-level fusions · Feature pyramid · Group-wise correlation

1 Introduction

Multi-view stereo (MVS) is a classic and fundamental computer vision problem that has been extensively studied for decades [1]. It aims to reconstruct the dense representation of a scene from a collection of images with calibrated camera parameters, which has been widely applied in 3D visualization, virtual/augmented reality, robotics, etc. Before the deep learning era, traditional MVS methods usually utilized hand-crafted photo-consistency metrics, such as the sum of absolute differences (SAD) and the normalized cross-correlation (NCC), to find corresponding pixels based on the projection relationship among multiple views. Although traditional MVS methods [2–4] have achieved impressive reconstruction performance, recent works [5–7] show that learning-based methods can produce comparable or even better results.

In learning-based methods, the deep features extracted by convolutional neural networks (CNNs) encode global and local information for reliable dense matching, which is then regulized to obtain high-quality 3D geometries. In particular, Yao et al. [5] propose an end-to-end network named MVSNet that builds a cost volume based on differentiable homography warping and utilizes a multi-scale 3D CNN for regularization. The network infers the depth map estimation for each view, achieving a more accurate and complete reconstruction compared to traditional methods. However, the regularization process based on the 3D CNN is memory-consuming, as the memory requirement is cubic to the image resolution, which therefore limits it to low-resolution input images. To enable high-resolution

✉ Xiaoping Liu
liu@hfut.edu.cn

Youcheng Cai
youchengcyc@mail.hfut.edu.cn

Lin Li
lilinjulia@hfut.edu.cn

Dong Wang
dongwang7@mail.hfut.edu.cn

[1] The School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230009, China

[2] The Engineering Research Center of Safety Critical Industrial Measurement and Control Technology, Ministry of Education, Hefei, 230009, China

image processing, R-MVSNet [6] is proposed to replace the 3D CNN with a convolutional gated recurrent unit (GRU) to regularize the cost volume sequentially. Then, D2HC-RMVSNet [8] further reduces the memory consumption through a novel Hybrid U-LSTM. Additionally, other methods typically use a coarse-to-fine strategy [7, 9, 10] to achieve high-resolution outputs. For example, Point-MVSNet [7] first predicts a coarse depth map and then iteratively refines the point cloud to result in the full resolution depth map. Yang et al. [9] designed CVPMVSNet to build a cost volume pyramid and iteratively estimate the pixelwise depth residual to refine the depth map.

In this work, we propose a multi-level fusion aware feature pyramid based multi-view stereo network (MFNet) for accurate and reliable depth inference. First, our network architecture follows a coarse-to-fine approach that starts by estimating a coarse depth map and then iteratively refines the depth to achieve high-resolution outputs, significantly reducing memory consumption. Second, instead of using the U-shape feature extractor, which is commonly adopted in previous works [5–7], we design a novel and efficient multi-level fusion based feature extraction module (MLF) that exchanges information between multi-level features to construct the feature pyramid. The proposed backbone can boost the multi-level feature representations for reliable dense matching. Finally, the group-wise correlation is conducted to construct a lightweight cost volume for regularization, which can improve both efficiency and effectiveness.

In summary, our main contributions are listed as follows:

1. We propose a novel coarse-to-fine framework for MVS that integrates the group-wise correlation to construct cost volumes, thereby reducing memory consumption and achieving high-resolution inference.
2. We design a novel and efficient feature extraction module that utilizes multi-level fusion to boost the dense matching robustness.
3. Our method outperforms the state-of-the-art learning-based MVS methods on the standard benchmarks DTU [11], Tanks & Temples [12], and BlendedMVS [13].

2 Related work

2.1 Traditional MVS

According to the taxonomy of Seitz et al. [1], traditional MVS methods can be divided into four categories: volumetric-based [14, 15], surface evolution-based [16, 17], patch expansion-based [4, 18], and depth map-based [2, 19] methods. The basic idea of volumetric-based methods is to divide the 3D space into small voxel grids and then decide whether each voxel belongs to the surface. Sinha et al. [14] proposed generating a multiresolution volumetric mesh based on photo consistency and then using a graph cut algorithm to produce a triangulated surface. Ulusoy et al. [15] introduced an object-level shape prior that formulated a probabilistic model using 3D shape information from multiple objects to generate a dense 3D reconstruction. Surface evolution-based methods typically start from a good initial guess of the scene surface and then reconstruct the 3D surface by a surface evolution scheme. Cremers et al. [16] proposed defining the reconstruction problem as one of minimizing a convex function with silhouette consistency. To recover fine-scale details, Li et al. [17] presented a detail-preserving and content-aware method that utilized reprojection error minimization and mesh denoising. Patch expansion-based methods, however, consider the surface as a dense set of small patches. Furukawa et al. [4] proposed a match, expand, and filter procedure that starts from a sparse set of matched features and then repeatedly expands to cover the surface. Locher et al. [18] presented a computational budget in a progressive manner, which makes efficient use of computational power and allows large-scale reconstruction. Comparatively, the depth map-based approach seems more concise and flexible and consists of two parts: per-view depth map estimation and point cloud fusion. Galliani et al. [2] presented Gipuma, an algorithm that formulates Patchmatch in the scene space to aggregate image similarities between multiple views for depth map estimation. To improve the performance, Xu et al. proposed the ACMM [19], which adopted adaptive checkerboard sampling and multihypothesis joint view selection. While traditional MVS methods have achieved remarkable results, learning-based methods have recently demonstrated comparable results or even surpassed the performance of traditional methods on several datasets [11, 12].

2.2 Learning-based MVS

Recently, deep CNNs have demonstrated great success on many vision tasks of image recognition [20, 21], object detection [22, 23], and semantic segmentation [24]. For MVS problems, deep CNNs can be easily adapted to learn a similarity measurement between multiple patches and introduce a learned cost metric [25]. There are basically two types of learning-based MVS methods: volume-based and depth map-based methods.

2.2.1 Volumetric-based methods

To apply CNNs to 3D reconstruction, LSM [26] first encodes the camera parameters by applying the differentiable projection and then uses a 3D CNN for reasoning

about surface voxels. It allows end-to-end learning for the MVS task. Concurrently, Ji et al. presented SurfaceNet [27], which introduced a novel representation of the colored voxel cube (CVC) to implicitly encode the camera parameters. The key advantage of this approach is accounting for both photo consistency and geometric relations. Furthermore, SurfaceNet+ [28] applied a novel volume-wise view selection approach and a multi-scale strategy to refine and recover 3D geometry, yielding a promising result under the extreme sparse-MVS settings. However, all these volumetric-based methods suffer from the common limitation of volumetric representation: the huge memory requirement precludes the processing of large-scale scenes.

2.2.2 Depth map-based methods

Compared with volumetric-based methods, depth map-based methods have been proven more effective. Yao et al. [5] proposed an end-to-end network named MVSNet, which is considered the de facto standard pipeline. The network constructs a cost volume followed by regularization to infer the depth map for each view. Then, the 3D point cloud is fused by the estimated depth maps. However, the regularization process implemented by 3D CNN requires large GPU memory, resulting in low-resolution depth map output. To overcome this limitation, some methods replace 3D CNN with 2D sequential regularization. For example, R-MVSNet [6] adopted the convolutional GRU to regularize the cost volume. D2HC-RMVSNet [8] presented a hybrid recurrent network, named Hybrid U-LSTM, for cost volume regularization. Recently, AA-RMVSNet [29] introduced adaptive aggregation modules to adaptively extract image features and aggregate cost volumes before applying an RNN-CNN hybrid network for recurrent regularization. These recurrent multi-view stereo networks allow high-resolution reconstruction and finer depth hypothesis sampling. Another attempt to reduce memory consumption is based on the coarse-to-fine strategy. Point-MVSNet [7] first predicts a coarse depth map and then iteratively refines the point cloud by estimating the 3D flow for each point. This point-based architecture can generate detailed point clouds. Fast-MVSNet [30] introduced a novel sparse-to-dense framework that densifies the sparse high-resolution depth map by a data-adaptive propagation method, which achieves fast and accurate depth estimation. Furthermore, CasMVSNet [10] and CVP-MVSNet [9] built upon a multi-stage framework that iteratively constructs new cost volumes to achieve high-resolution depth map inference, which provides promising reconstruction results. Instead of using a fixed depth hypothesis, UCS-Net [31] designed adaptive thin volumes (ATVs) to efficiently partition local depth ranges, which enables a high accuracy depth map. Moreover, Xu et al. [32] proposed CIDER, which constructs a lightweight cost volume by an average group-wise correlation similarity measure and achieves subpixel estimation with an inverse depth regression task. In our work, the group-wise correlation similarity measure is utilized jointly with the coarse-to-fine strategy, further boosting GPU memory efficiency and computational speed.

3 Method

In this section, we describe the detailed framework of the proposed MFNet. The full architecture is depicted in Fig. 1. As in existing works, the network takes a reference image I_0 and neighboring source images $\{I_i\}_{i=1}^{N-1}$ with corresponding camera parameters $\{\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i\}_{i=0}^{N-1}$ as input to infer the depth map for I_0, where $\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i$ refer to the camera intrinsics, rotation, and translation matrix, and N is the total number of input images.

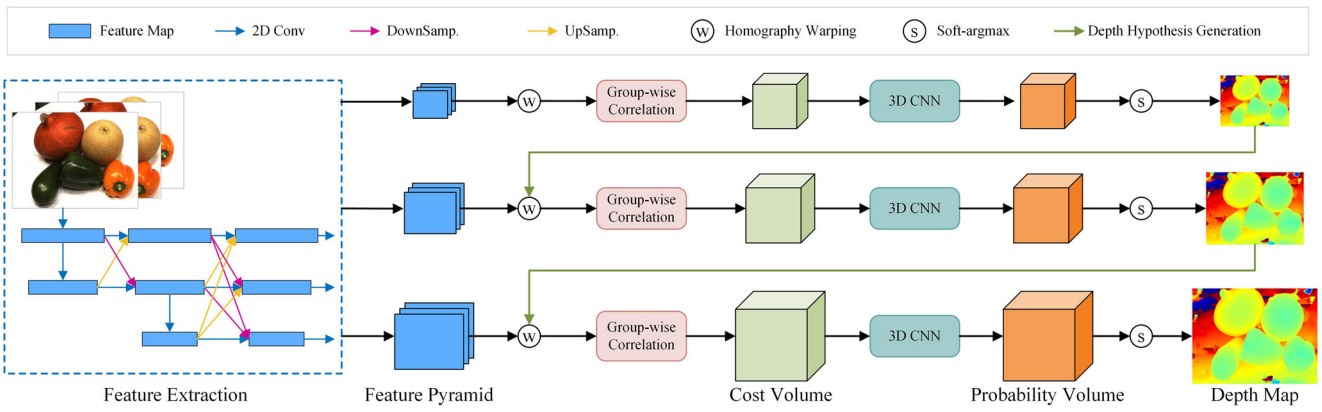


Fig. 1 The network architecture of MFNet. First, our feature extraction module conducts multi-level fusions to construct the feature pyramid. Then, the feature maps go through warping and group-wise correlation to generate cost volumes. Finally, the 3D CNN regularizes the cost volumes for depth regression

3.1 Feature pyramid

The learnable features have been demonstrated to be robust for extracting dense feature correspondences, especially in intractable regions such as low-textured, specular, and reflective regions. The common approach of existing work is to extract multi-level features from the high-resolution image in a high-to-low process or extract multi-level features by building image pyramids. To further improve the reliability of the features, we conduct multi-level fusions by exchanging contextual information across the multi-level features. Our approach is partially inspired by deep fusion [33–35], and they are not applicable to MVS due to the clear computational limitations and huge GPU memory consumption. An important objective of MVS is to achieve large-scale reconstruction, which requires memory-efficient feature extraction modules. Without constructing deep layers, our approach is to design a lightweight but efficient network for reliable dense matching.

Specifically, we design a multi-level fusion based feature extraction module (MLF) to extract an L-level feature pyramid $\{\mathbf{F}_i^l\}_{l=1}^L$ for each input image I_i, gradually increasing the spatial resolution and decreasing the feature channels to accommodate the coarse-to-fine strategy. Let $H \times W$ be the resolution of each input image. As shown in Fig. 2, MLF connects convolutional layers in a high-to-low process, and multi-level fusions are introduced to aggregate different level representations. We adopt 3×3 convolutions with the stride 2 or 4 for $2\times$ or $4\times$ downsampling, respectively. For upsampling, we adopt 3×3 convolutions following the simple nearest neighbor sampling. Here, the feature maps at l-level are denoted as $\mathbf{F}_i^l \in R^{H^l \times W^l \times C^l}$, where C^l is the number of feature channels and corresponding spatial resolutions of $H^l \times W^l = H/2^{(L-l)} \times W/2^{(L-l)}, l \in \{1, \cdots, L\}$. We set the feature pyramid to have $L = 3$ levels.

Furthermore, to better demonstrate the superiority of the proposed MLF, we also compare the proposed MLF with 1) the multi-level feature extraction based on U-Net [36] used in MVSNet [5] (Fig. 3a), 2) the FPN [37] used in CasMVSNet [10] (Fig. 3b). The proposed MLF (Fig. 3c) is superior in three aspects. First, instead of recovering the resolution from the low-level representations, it preserves the high-level representation while fusing the multi-level semantic information, which accordingly obtains potentially more reliable high-level features. Second, the multi-level fusions strengthen the connection between multi-level features, which can be well adapted to the multi-stage network structure for better depth inference. Third, the backbone of the module is lightweight but efficient for feature extraction and dense matching. We will show that the proposed MLF is able to significantly improve performance.

3.2 Cost volume construction

After obtaining the multi-level features, we construct cascade cost volumes for all stages. Similar to previous methods [9, 10], the coarsest resolution depth map is first estimated and then iteratively refined to achieve high-resolution depth inference. We define $L = 3$ as the total stage number of the network.

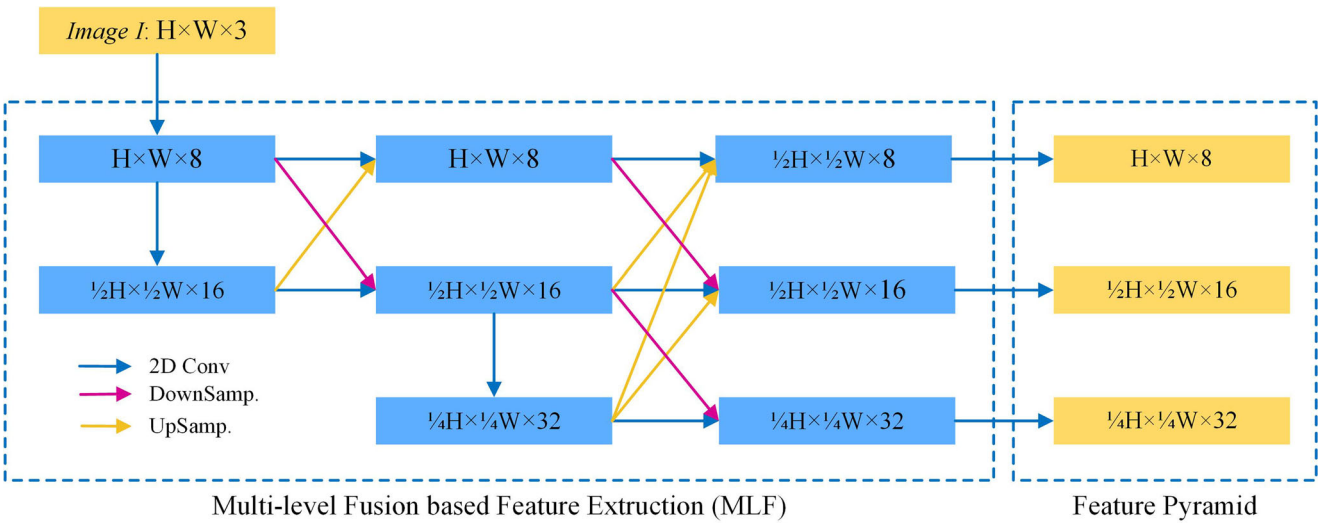


Fig. 2 The network architecture of **MLF**, which is designed by integrating multi-level fusions to aggregate the information from high-, medium-, and low-level feature maps. **DownSamp.**= 3×3 convolution, and **UpSamp.**= 3×3 convolution following nearest neighbor upsampling

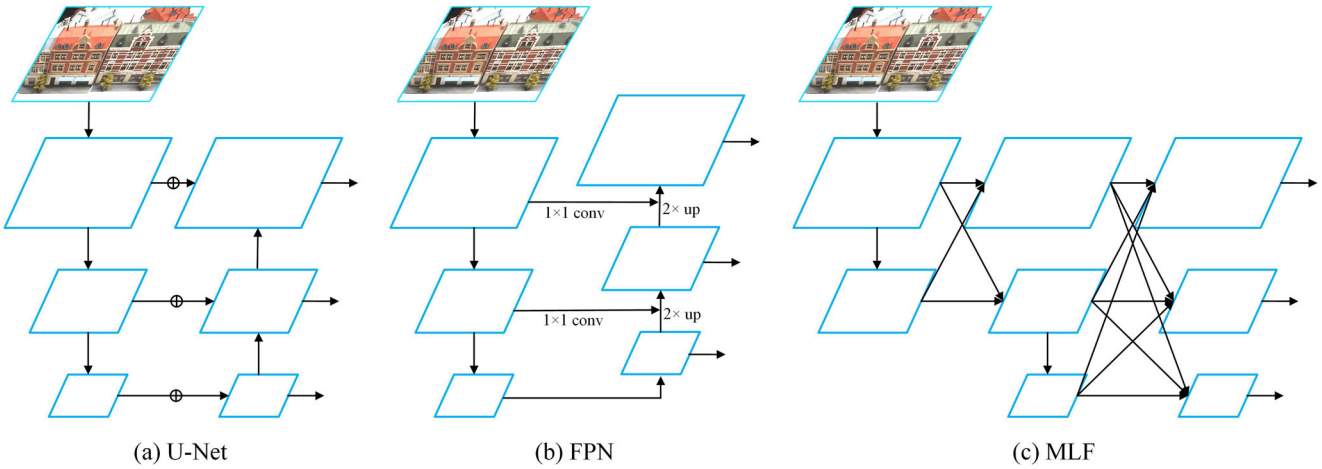


Fig. 3 Different design choices for multi-level feature extraction

3.2.1 Homography warping and group-wise correlation

We construct a cost volume at the l^{th} stage using the differentiable homography warping operation, which implicitly encodes extracted features with the camera geometries. Specifically, based on multiple sampled depth hypotheses d, the coordinate mapping from the feature representations of the i^{th} source view to the reference view at the l^{th} stage is determined by the homography:

$$\mathbf{H}_i^l(d) = \mathbf{K}_i^l \mathbf{R}_i \left(\mathbf{I} - \frac{(\mathbf{t}_1 - \mathbf{t}_i)\, \mathbf{n}_0^T}{d} \right) \mathbf{R}_0^T \left(\mathbf{K}_0^l \right)^{-1} \tag{1}$$

where $\mathbf{I}$ is the identity matrix, and n_0 refers to the principal axis of the reference camera.

When the multiple warped feature volumes are obtained, we employ an average group-wise correlation similarity [32, 38] to aggregate them to one cost volume rather than the variance-based similarity [5, 7, 9, 10]. In this way, the number of channels in the cost volume can be reduced, thus enabling the lightweighting of the cost volume. The basic idea is that the features are divided into groups, and the similarity map is calculated group by group. Specifically, we first divided the reference image feature $\boldsymbol{F}_0^l$ and the warped source image feature $\tilde{\boldsymbol{F}}_i^l$ into G groups along the channel dimension, and the group similarity between $\mathbf{F}_0^l$ and $\tilde{\boldsymbol{F}}_i^l$ at g^{th} group can be computed as:

$$\mathbf{S}_i^{l,g} = \frac{1}{C^l / G^l} \left\langle \mathbf{F}_0^{l,g}, \tilde{\boldsymbol{F}}_i^{l,g} \right\rangle, g \in \{0, 1, \cdots, G-1\} \tag{2}$$

where $\langle \cdot, \cdot \rangle$ refers to the inner product. Next, the group similarity maps of all G^l groups are compressed into a G-channel similarity map $\mathbf{S}_i^l$ of size $G^l \times H^l \times W^l \times D^l$. D^l is the number of depth hypotheses at the l^{th} stage. Finally, the similarity maps of all the views are averaged to one cost volume to adapt an arbitrary number of input images:

$$\mathbf{V}^l = \frac{1}{N-1} \sum_{i=1}^{N-1} \mathbf{S}_i^l \tag{3}$$

Note that this lightweight cost volume can simultaneously meet the memory consumption and computational requirements in the subsequent procedure.

3.2.2 Depth map inference

Recall that our ultimate goal is to estimate $\mathbf{D} = \mathbf{D}^L$ for I_0. In the first stage ($l = 1$), the depth hypotheses are uniformly sampled in the depth range $[d_{min}, d_{max}]$, which can be expressed as:

$$d(m) = d_{\min} + m \left(d_{\max} - d_{\min} \right) / M, m \in \{0, 1, \cdots, M-1\} \tag{4}$$

where M is the number of depth hypotheses. Similar to MVSNet [5], the 3D CNN are then used to regularize the cost volume $\mathbf{V}^{l=1}$ to predict a depth probability volume $\mathbf{P}^{l=1}$, which indicates the probability distribution of the depth. Then, the depth map is computed by the soft-argmax operation. Mathematically, the predicted depth for each pixel $\boldsymbol{p}$ is expressed as:

$$\mathbf{D}_{\boldsymbol{p}}^1 = \sum_{m=0}^{M-1} d(m) \mathbf{P}_{\boldsymbol{p}}^1(d(m)) \tag{5}$$

where $\mathbf{P}_{\boldsymbol{p}}^1(d(m))$ denotes the probability of pixel $\boldsymbol{p}$ for the depth hypothesis $d(m)$ at the first stage.

To further estimate a finer depth $\mathbf{D}^l$ in the following $l^{th} (l > 1)$ stage, we define the depth hypothesis with a

refined depth interval based on the previous depth map $\mathbf{D}^{l-1}$:

$$d(m) = \mathbf{D}^{l-1} \uparrow + m\Delta d_{\boldsymbol{p}}, m \in \{-M/2, \cdots, M/2-1\} \quad (6)$$

where $\mathbf{D}^{l-1} \uparrow$ denotes the upsampled $\mathbf{D}^{l-1}$ via bicubic interpolation. Here, we use the approach proposed by [9] to determine the depth interval $\Delta d_{\boldsymbol{p}}$. We project $\boldsymbol{p}$ into the source images, and the distance between two neighbor pixels along the polar line projected into the 3D ray will be defined as the depth interval, thus avoiding the projected points in the image that are too close to provide extra information for feature matching. Ultimately, the updated depth of each pixel $\boldsymbol{p}$ at the next level is computed as:

$$\mathbf{D}_{\boldsymbol{p}}^{l} = \sum_{m=-M/2}^{M/2-1} \left(\mathbf{D}_{\boldsymbol{p}}^{l-1} \uparrow + m\Delta d_{\boldsymbol{p}}\right) \mathbf{P}_{\boldsymbol{p}}^{l} \left(\left(\mathbf{D}_{\boldsymbol{p}}^{l-1} \uparrow + m\Delta d_{\boldsymbol{p}}\right)\right), l \in \{2, \cdots, L\} \quad (7)$$

where $\mathbf{P}^{l}$ denotes the depth probability volume at the l^{th} stage, which is regularized from the cost volume $\mathbf{V}^{l}$ by the 3D CNN.

3.3 Loss function

For the depth regression, we construct the multi-level ground truth depth $\left\{\mathbf{D}_{gt}^{l}\right\}_{l=1}^{L}$ as the supervisory signal. The total loss is based on the smooth $L1$ loss, which measures the absolute difference between the ground truth and the estimated depth. Considering all stages, our loss function is:

$$Loss = \sum_{l=1}^{L} \sum_{\boldsymbol{p} \in \boldsymbol{p}_{valid}} \left\| \mathbf{D}_{gt}^{l}(\boldsymbol{p}) - \mathbf{D}_{pred}^{l}(\boldsymbol{p}) \right\|_1 \quad (8)$$

where $\mathbf{p}_{valid}$ denote the set of valid ground truth pixels and $\mathbf{D}_{pred}^{l}$ denote the predicted depth maps at l^{th} stage.

4 Experiments

4.1 Implementation details

4.1.1 Datasets

We evaluate our method on the DTU dataset [11], Tanks & Temples dataset [12], and BlendedMVS dataset [13]. The DTU dataset is an indoor MVS dataset containing 124 scenes scanned from 49 or 64 views under 7 different lighting conditions. The Tanks & Temples dataset contains both indoor and outdoor scenes captured in more complex environments. The BlendedMVS dataset provides a variety of 113 scenes, including outdoor buildings, architectures, sculptures, and small objects.

4.1.2 Training

We train our MFNet on the DTU training set [11] and the BlendedMVS training set [13]. The resolution of the input image is set to 640×512 for the DTU training set and 678×576 for the BlendedMVS training set. The number of views is set to $N = 3$ during training as previous methods [5, 6]. For different stages, the number of feature channels is set to $C = [32, 16, 8]$, the number of depth hypothesis is set to $M = [48, 32, 8]$, and the number of groups is set to $G = [8, 4, 4]$. We implement our network by using PyTorch with Adam optimizer. The network trained for 16 epochs with the batch size of 2 on an NVIDIA GTX 2080Ti graphics card. The learning rate is set to 0.001 and then downscaled by 2 at the 10th, 12th, 14th epoch.

4.1.3 Filter and fusion

We followed Yao et al. [5] to generate a dense point cloud with two steps: depth map filtering and depth map fusion. For the filter step, the photometric and geometric consistencies are considered to remove outliers. For the fusion step, all depth maps are fused into a consistent point cloud based on the method developed in [2].

4.2 Benchmark performance

4.2.1 Evaluation on DTU dataset

In our experiments, we first evaluate our MFNet on the DTU evaluation set [11]. We use the model trained on the DTU training set. Similar to previous methods, the number of views is set to 5, and the input image size is set to 1600×1184. We compare the proposed MFNet with state-of-the-art traditional methods and learning-based methods. The distance metrics in terms of accuracy, completeness, and overall score are adopted to evaluate the quality of the final reconstructions. Here, the metrics measure the distance between the reconstruction results and the ground truth. The evaluation protocol is provided by the DTU dataset and conducted via the MATLAB code [11].

The quantitative comparison results are shown in Table 1. Our MFNet outperforms other methods in both completeness and overall score while remaining quite competitive in accuracy. As a consequence, MFNet achieves the best overall performance and ranks 1st in the DTU benchmark to the best of our knowledge. Figure 4 shows the qualitative comparison results, where the blue rectangle indicates that MFNet can reconstruct a more complete

Table 1 Quantitative results of reconstruction quality on the DTU evaluation set [11]

Methods	acc.	comp.	overall
Camp [39]	0.835	0.554	0.695
Furu [4]	0.613	0.941	0.777
Tola [3]	0.342	1.190	0.766
Gipuma [2]	**0.283**	0.873	0.578
SurfaceNet [27]	0.450	1.040	0.745
MVSNet [5]	0.396	0.527	0.462
R-MVSNet [6]	0.383	0.452	0.417
P-MVSNet [40]	0.406	0.434	0.420
CIDER [32]	0.406	0.434	0.420
Point-MVSNet [7]	0.342	0.411	0.376
CVPMVSNet [9]	0.296	0.406	0.351
CasMVSNet [10]	0.346	0.351	0.348
UCSNet [31]	0.338	0.349	0.344
ADR-MVSNet [41]	0.354	0.317	0.335
MFNet	0.339	**0.304**	**0.321**

Our MFNet outperforms all methods with the best completeness and overall score
The best results are marked in bold

3D point cloud. All reconstruction results on the DTU evaluation set are illustrated in Fig. 5.

4.2.2 Evaluation on tanks & temples dataset

To evaluate the generalization ability of MFNet, we test the proposed method on the intermediate set of Tanks & Temples dataset [12], using the model trained on BlendedMVS [13]. We use $N = 7$ and $W \times H = 1920 \times 1056$ for this experiment. We use the F-score metric to evaluate the reconstruction quality, which considers both accuracy and completeness. The metric is calculated at the official site of Tanks & Temples (https://www.tanksandtemples.org/).

Quantitative results are presented in Table 2, which shows that MFNet achieves the highest mean F-score (59.46) compared with other state-of-the-art methods. The reconstruction results are shown in Fig. 6. MFNet can reconstruct a more continuous and complete 3D point cloud, especially for the Horse scene, which contains a large amount of textureless areas. This might be because the proposed feature extraction module enables better dense matching. The performance of MFNet is excellent in relatively near scenes and is still competitive in larger scenes (e.g., Lighthouse, Playground). Note that as the larger scenes have large depth ranges of interest, P-MVSNet [40] provides more sampled depth hypotheses, which can increase the depth prediction quality.

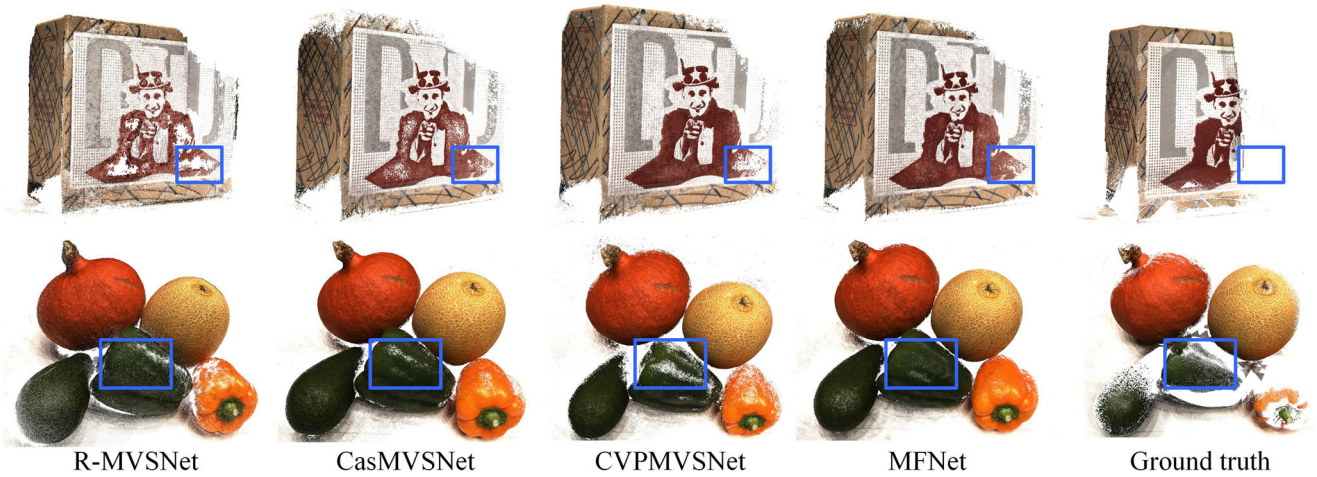


Fig. 4 Qualitative results of scan 16 and scan 75 of DTU dataset [11]. As shown in the blue rectangle, the reconstruction results of the proposed MFNet contain better fine detailed structures than other methods

Fig. 5 Reconstruction results on the evaluation set of DTU [11]

4.2.3 Evaluation on BlendedMVS dataset

BlendedMVS [13] is a new large-scale MVS dataset that covers various scenes containing small objects and large outdoor scenes. Since the BlendedMVS dataset does not provide ground truth reconstructed point clouds, we qualitatively compare the reconstruction results and quantitatively compare the depth inference quality. For a fair comparison, we set input images with $N = 7$ and $W \times H = 678 \times 576$ for all methods. The depth maps established from R-MVSNet and MVSNET are upsampled to the same size as the depth maps from other methods.

In general, the end point error (EPE), which is calculated by using (8), a ratio of <1 pixel accuracy, and a ratio of <3 pixel accuracy are considered to evaluate the depth map inference quality. Table 3 shows the quantitative results, which show that MFNet trained on the BlendedMVS dataset achieves the best performance.

Table 2 Reconstruction results of intermediate set in the Tank and Temples dataset [12]

Methods	Mean	Family	Francis	Horse	Lighthouse	M60	Panther	Playground	Train
COLMAP [42]	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04
SurfaceNet [27]	49.38	62.38	32.35	29.35	62.86	54.77	54.14	56.13	43.10
MVSNet [5]	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
R-MVSNet [6]	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
P-MVSNet [40]	55.62	70.04	44.64	40.22	**65.20**	55.08	55.17	**60.37**	54.29
CIDER [32]	46.76	56.79	32.39	29.89	54.67	53.46	53.51	50.48	42.85
Point-MVSNet [7]	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06
CVPMVSNet [9]	54.03	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54
CasMVSNet [10]	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51
UCSNet [24]	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89
ADR-MVSNet [41]	52.47	72.61	52.81	37.43	53.46	51.88	48.71	59.11	43.76
MFNet	**59.46**	**79.97**	**59.10**	**51.77**	53.28	**60.96**	**56.98**	58.20	**55.43**

The proposed MFNet achieves the best result compared with state-of-the-art methods

The best results are marked in bold

Fig. 6 Reconstruction results of the intermediate set of the Tanks & Temples dataset [12]

Moreover, to explore the generalization ability of MFNet, we also provide the results trained on the DTU dataset. Overall, the proposed MFNet outperforms other state-of-the-art methods in all experiments, which demonstrates the superiority of MFNet in terms of depth map inference and generalization (Fig. 7).

The qualitative comparison of a large-scale outdoor scene is visualized in Fig. 8, and the reconstruction results are visualized in Fig. 7. It is obvious that the proposed MFNet can generate the highest quality reconstruction results that have a complete and detailed structure.

4.3 Ablation studies

In this section, we conduct ablation studies to analyze the components of the network architecture. For all the following studies, experiments are performed on the DTU dataset, Tanks & Temples dataset, and BlendedMVS dataset. The statistical comparison of the results is listed to measure the reconstruction quality.

4.3.1 Ablations on feature pyramid

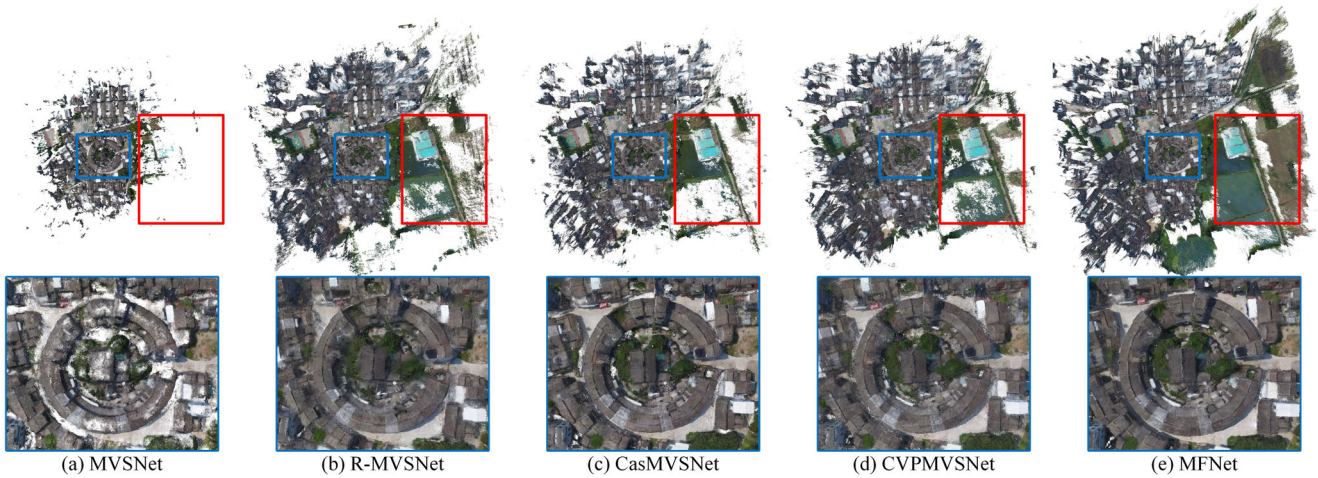
Our designed multi-level feature extraction module (MLF) aims to learn reliable features for dense matching. As shown in Table 4, compared with other common feature extraction modules, including U-Net and FPN, the proposed MLF can improve both the accuracy and completeness of the DTU dataset and achieve better results on the Tanks & Temples dataset and BlendedMVS dataset. This might be because integrated multi-level fusion can generate more robust multi-level features and strengthen the connections between multi-stage network structures, allowing the network to estimate depth values precisely.

Table 3 Quantitative results of depth inference quality on the validation set of BlendedMVS [13]

Methods	Training dataset	EPE	<1 pix. acc	<3 pix. acc
MVSNet [5]	DTU	5.97	56.4%	76.2%
R-MVSNet [6]	DTU	7.58	56.2%	77.0%
CVPMVSNet [9]	DTU	9.92	65.5%	73.5%
CasMVSNet [10]	DTU	4.51	76.4%	83.7%
MFNet	DTU	**2.80**	**83.6%**	**90.1%**
MVSNet [5]	BlendedMVS	2.53	71.7%	88.1%
R-MVSNet [6]	BlendedMVS	3.14	72.1%	87.9%
CVPMVSNet [9]	BlendedMVS	6.80	69.2%	79.1%
CasMVSNet [10]	BlendedMVS	2.89	83.3%	90.1%
MFNet	BlendedMVS	**1.10**	**93.5%**	**97.1%**

MFNet trained on the BlendedMVS dataset achieves the best performance
The best results are marked in bold

Fig. 7 Reconstruction results on the validation set of BlendedMVS [13]



(a) MVSNet (b) R-MVSNet (c) CasMVSNet (d) CVPMVSNet (e) MFNet

Fig. 8 Comparison of the reconstruction results on the large-scale outdoor scene of BlendedMVS [13]. MFNet can reconstruct complete and finer results compared with other methods

Table 4 Ablation study on the DTU evaluation dataset

Methods	U-Net	FPN	MLF	Var	GC	DTU			Tanks & Temples	BlendedMVS
						acc.	comp.	overall	F-score	EPE
Model A	✓			✓		0.365	0.318	0.342	56.30	2.69
Model B		✓		✓		0.334	0.342	0.338	57.31	2.53
Model C			✓	✓		**0.323**	0.331	0.327	58.95	1.49
MFNet			✓		✓	0.339	**0.304**	**0.321**	**59.46**	**1.10**

MLF+GC achieves the best overall quality, which demonstrates the effectiveness of different components of our method
The best results are marked in bold

Table 5 Comparison of runtime and GPU memory on the DTU dataset with the same input sizes

Methods	Input Sizes	Depth Map Sizes	GPU Memory	Runtime
MVSNet [5]	1600*1184	400*296	15.4GB	1.18s
R-MVSNet [6]	1600*1184	400*296	6.7GB	2.35s
Point-MVSNet [7]	1600*1184	800*592	8.7GB	5.44s
CVP-MVSNet [9]	1600*1184	1600*1184	6.5GB	1.89s
CasMVSNet [10]	1600*1184	1600*1184	7.5GB	1.03s
Model A (**U-Net+Var**)	1600*1184	1600*1184	6.8 GB	1.11s
Model B (**FPN+Var**)	1600*1184	1600*1184	7.5 GB	1.00s
Model C (**MLF+Var**)	1600*1184	1600*1184	6.7 GB	1.15s
MFNet (**MLF+GC**)	1600*1184	1600*1184	**5.7 GB**	**0.96s**

The proposed MFNet has the smallest GPU memory requirement and the fastest running speed
The best results are marked in bold

4.3.2 Ablations on group-wise correlation similarity

To further study the effectiveness of group-wise correlation similarity (**GC**), we evaluated the network with common variance-based similarity (**Var**) as in previous methods [5–7]. Table 4 demonstrates that group-wise correlation similarity can improve the overall performance, especially in terms of completeness. Although the accuracy is decreased, this might be expected as the channels of the cost volumes are reduced.

4.4 Runtime and GPU memory

To analyze the efficiency of the proposed method, we further compare MFNet with related learning-based methods in terms of runtime and GPU memory usage. As shown in Table 5, MFNet has the smallest GPU memory requirement and the fastest running speed while achieving the best reconstruction quality. Moreover, both MLF and GC can reduce memory consumption, which demonstrates the efficiency and effectiveness of MFNet.

5 Conclusion

In this paper, we have proposed MFNet, a novel learning-based architecture for multi-view stereo. The proposed MFNet can estimate high accuracy and high-resolution depth maps in an end-to-end fashion for 3D reconstruction. Specifically, we designed a new feature extraction module (MLF) to construct the feature pyramid, which adopts multi-level fusions to aggregate all levels of contextual information. Compared with the classical feature extractors of U-Net and FPN, MLF can fully exploit the multi-level information to improve matching robustness, leading to more accurate and complete reconstruction point clouds. In addition, MFNet adopts a coarse-to-fine approach that starts by estimating a coarse depth map and then iteratively generates higher-resolution depth maps. Furthermore, MFNet also integrates the group-wise correlation similarity to aggregate the lightweight cost volumes. This reduces memory consumption and boosts computational speed. Experimental results on the benchmark datasets of DTU and Tanks & Temples show that MFNet achieves the best performance compared with state-of-the-art methods. Additionally, comprehensive experiments on BlendedMVS demonstrate the great generalization of MFNet.

Acknowledgements This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFC1523100 and in part the National Natural Science Foundation of China under Grant 61877016.

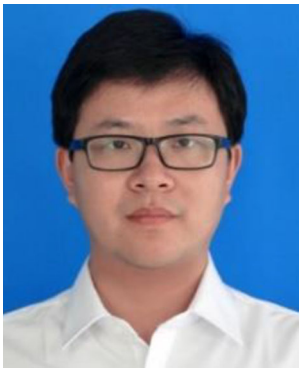
References

1. Seitz SM, Curless B, Diebel J, Scharstein D, Szeliski R (2006) A comparison and evaluation of multi-view stereo reconstruction algorithms. In: IEEE Conference on computer vision and pattern recognition. vol 1, pp 519–528
2. Galliani S, Lasinger K, Schindler K (2015) Massively parallel multiview stereopsis by surface normal diffusion. In: IEEE Conference on computer vision and pattern recognition. pp 873–881
3. Tola E, Strecha C, Fua P (2012) Efficient large-scale multi-view stereo for ultra high-resolution image sets. Mach Vis Appl 23:903–920
4. Furukawa Y (2010) Accurate, dense, and robust multiview stereopsis. IEEE Trans Pattern Anal Mach Intell 32(8):1362–1376
5. Yao Y, Luo Z, Li S, Fang T, Quan L (2018) Mvsnet: depth inference for unstructured multi-view stereo. In: European conference on computer vision. pp 785–801
6. Yao Y, Luo Z, Li S, Shen T, Fang T, Quan L (2019) Recurrent mvsnet for high-resolution multi-view stereo depth inference. In: IEEE Conference on computer vision and pattern recognition. pp 5525–5534

7. Chen R, Han S, Xu J, Su H (2019) Point-based multi-view stereo network. In: IEEE International conference on computer vision. pp 1538–1547
8. Yan J, Wei Z, Yi H, Ding M, Zhang R, Chen Y, Wang G, Tai Y-W (2020) Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In: European conference on computer vision. pp 674–689
9. Yang J, Mao W, Alvarez JM, Liu M (2021) Cost volume pyramid based depth inference for multi-view stereo. IEEE transactions on pattern analysis and machine intelligence
10. Gu X, Fan Z, Zhu S, Dai Z, Tan F, Tan P (2020) Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: IEEE Conference on computer vision and pattern recognition. pp 2495–2504
11. Aanaes H, Jensen RR, Vogiatzis G, Tola E, Dahl AB (2016) Large-scale data for multiple-view stereopsis. Int J Comput Vis 120(2):153–168
12. Knapitsch A, Park J, Zhou Q-Y, Koltun V (2017) Tanks and temples: benchmarking large-scale scene reconstruction. ACM Trans Graph 36(4):1–13
13. Yao Y, Luo Z, Li S, Zhang J, Ren Y, Zhou L, Fang T, Quan L (2020) Blendedmvs: a large-scale dataset for generalized multi-view stereo networks. In: IEEE Conference on computer vision and pattern recognition. pp 1790–1799
14. Sinha SN, Mordohai P, Pollefeys M (2007) Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. In: IEEE Conference on computer vision and pattern recognition. pp 1–8
15. Ulusoy AO, Black MJ, Geiger A (2017) Semantic multi-view stereo: jointly estimating objects and voxels. In: IEEE Conference on computer vision and pattern recognition. pp 4531–4540
16. Cremers D, Kolev K (2011) Multiview stereo and silhouette consistency via convex functionals over convex domains. IEEE Trans Pattern Anal Mach Intell 33(6):1161–1174
17. Li Z, Wang K, Zuo W, Meng D, Zhang L (2016) Detail-preserving and content-aware variational multi-view stereo reconstruction. IEEE Transactions on Image Processing 25(2):864–877
18. Locher A, Perdoch M, Gool LV (2016) Progressive prioritized multi-view stereo. In: IEEE Conference on computer vision and pattern recognition. pp 3244–3252
19. Xu Q, Tao W (2019) Multi-scale geometric consistency guided multi-view stereo. In: IEEE Conference on computer vision and pattern recognition. pp 5483–5492
20. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE Conference on computer vision and pattern recognition. pp 770–778
21. Qian K, Tian L, Liu Y, Wen X, Bao J (2021) Image robust recognition based on feature-entropy-oriented differential fusion capsule network. Appl Intell 51(2):1108–1117
22. Xie E, Ding j, Wang W, Zhan X, Xu H, Sun P, Li Z, Luo P (2021) Detco: unsupervised contrastive learning for object detection. In: IEEE International conference on computer vision. pp 8392–8401
23. Pal SK, Pramanik A, Maiti J, Mitra P (2021) Deep learning in multi-object detection and tracking: state of the art. Appl Intell 51(9):6400–6429
24. Zhang X-L, Du B-C, Luo Z-C, Ma K (2021) Lightweight and efficient asymmetric network design for real-time semantic segmentation. Applied Intelligence. pp 1–16
25. Hartmann W, Galliani S, Havlena M, Van Gool L, Schindler K (2017) Learned multi-patch similarity. In: IEEE International conference on computer vision. pp 1586–1594
26. Kar A, Hane C (2017) Learning a multi-view stereo machine. In: Neural information processing systems. pp 365–376
27. Ji M, Gall J, Zheng H, Liu Y, Fang L (2017) Surfacenet: an end-to-end 3d neural network for multiview stereopsis. In: IEEE International conference on computer vision. pp 2326–2334
28. Ji M, Zhang J, Dai Q, Fang L (2020) surfacenet+: an end-to-end 3d neural network for very sparse multi-view stereopsis. IEEE Trans Pattern Anal Mach Intell 43(11):4078–4093
29. Wei Z, Zhu Q, Min C, Chen Y, Wang G (2021) Aa-rmvsnet: adaptive aggregation recurrent multi-view stereo network. In: IEEE International conference on computer vision. pp 6187–6196
30. Yu Z, Gao S (2020) Fast-mvsnet: sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In: IEEE Conference on computer vision and pattern recognition. pp 1949–1958
31. Cheng S, Xu Z, Zhu S, Li Z, Li LE, Ramamoorthi R, Su H (2020) Deep stereo using adaptive thin volume representation with uncertainty awareness. In: IEEE Conference on computer vision and pattern recognition. pp 2524–2534
32. Xu Q, Tao W (2020) Learning inverse depth regression for multi-view stereo with correlation cost volume. In: National conference on artificial intelligence. vol 34, pp 12508–12515
33. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: IEEE Conference on computer vision and pattern recognition. pp 5693–5703
34. Zhang T, Qi G-J, Xiao B, Wang J (2017) Interleaved group convolutions. In: IEEE International conference on computer vision. pp 4383–4392
35. Zhao L, Li M, Meng D, Li X, Zhang Z, Zhuang Y, Tu Z, Wang J (2018) Deep convolutional neural networks with merge-and-run mappings. In: International joint conference on artificial intelligence. pp 3170–3176
36. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention. pp 234–241
37. Lin T-Y, Dollar P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: IEEE Conference on computer vision and pattern recognition. pp 936–944
38. Guo X, Yang K, Yang W, Wang X, Li H (2019) Group-wise correlation stereo network. In: IEEE Conference on computer vision and pattern recognition. pp 3273–3282
39. Campbell ND, Vogiatzis G, Hernández C, Cipolla R (2008) Using multiple hypotheses to improve depth-maps for multi-view stereo. In: European conference on computer vision. pp 766–779
40. Luo K, Guan T, Ju L, Huang H, Luo Y (2019) P-mvsnet: learning patch-wise matching confidence aggregation for multi-view stereo. In: IEEE International conference on computer vision. pp 10451–10460
41. Li Y, Zhao Z, Fan J, Li W (2022) Adr-mvsnet: a novel cascade network for 3d point cloud reconstruction with pixel occlusion. Pattern recognition 108516
42. Schonberger JL, Frahm J-M (2016) Structure-from-motion revisited. In: IEEE Conference on computer vision and pattern recognition. pp 4104–4113

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Youcheng Cai received the B.S. degree in information and computing science from the Hefei University of Technology, Anhui, China, in 2008, where he is currently pursuing the Ph.D. degree in computer science. His research interests include 3D reconstruction, computer vision, and machine learning.



Dong Wang received the B.S. degree in electronic information engineering from the Anhui Normal University, Wuhui, China, in 2016, and the M.S. degree in electronic and communication engineering from the Hefei University of Technology, Hefei, China, in 2019. He is currently pursuing the Ph.D. at the Hefei University of Technology. His current research interests include 2D/3D human pose estimation, computer vision, and machine learning.

Lin Li received the M.S. degree in computer application technology and the Ph.D. degree in computer application technology from the Hefei University of Technology, Hefei, China, in 2014 and 2016, respectively. She is currently an Assistant Professor with the School of Computers and Information, Hefei University of Technology. Her research interests include computer graphics and computer animation.

Xiaoping Liu received the master's and Ph.D. degrees in computer science from the Hefei University of Technology, Hefei, China, respectively. He is currently working as a Professor with the School of Computer and Information, Hefei University of Technology. His research interests include 3D reconstruction, computer animation, and cooperative computing.