Machine Vision and Applications (2023) 34:93
https://doi.org/10.1007/s00138-023-01449-4

ORIGINAL PAPER

# Multi-distribution fitting for multi-view stereo

Jinguang Chen[1] · Zonghua Yu[1] · Lili Ma[1] · Kaibing Zhang[1]

Received: 12 August 2022 / Revised: 10 April 2023 / Accepted: 12 August 2023 / Published online: 30 August 2023


**Abstract**

We propose a multi-view stereo network based on multi-distribution fitting (MDF-Net), which achieves high-resolution depth map prediction with low memory and high efficiency. This method adopts a four-stage cascade structure, which mainly has the following three contributions. First, view cost regularization is proposed to weaken the influence of matching noise on building the cost volume. Second, it is suggested to adaptively calculate the depth refinement interval using multi-distribution fitting (MDF). Gaussian distribution fitting is used to refine and correct depth within a large interval, and then Laplace distribution fitting is used to accurately estimate depth within a small interval. Third, the lightweight image super-resolution network is applied to upsample the depth map in the fourth stage to reduce running time and memory requirements. The experimental results on the DTU dataset indicate that MDF-Net has achieved the most advanced results. It has the lowest memory consumption and running time among the high-resolution reconstruction methods, requiring only approximately 4.29G memory for predicting a depth map with the resolution of 1600 × 1184. In addition, we validate the generalization ability on Tanks and Temples dataset, achieving very competitive performance. The code has been released at https://github.com/zongh5a/MDF-Net.

**Keywords** Deep learning · Depth estimate · High resolution · Multi-view stereo · Point cloud

## 1 Introduction

Multi-view stereo (MVS) aims to recover the spatial structure of a scene through images acquired from multiple viewpoints of the same scene and information about the recorded internal and external camera parameters. With the rise of the convolutional neural network, the extracted image features have been used for stereo matching, and the reconstruction outperforms many traditional algorithms [1–4]. Algorithms that build and regularize cost volumes with image features [5–13] have achieved even more excellent results.

Although the algorithm of building the cost volume has produced many excellent results, its potential has not been fully exploited. First, the matching noise from different perspectives affects the reliability of the cost volume. Second, this algorithm requires many depth samplings within a preset depth range. The sampling density is directly proportional to the reconstruction quality and inversely proportional to the inference efficiency. Therefore, it is advantageous that the cascade structure methods [10, 12, 13] estimate the rough depth map in low-resolution dense sampling and refine the depth value in high-resolution sparse sampling. In this structure, obtaining depth refinement ranges adaptively at all scales is worth studying. Finally, this algorithm usually uses a 3D CNN network to regularize the cost volume, which consumes considerable memory and reduces efficiency. At this time, some networks [12], when the depth accuracy is sufficient, use a residual upsampling structure to upsample the depth map instead of building the cost volume at high resolution, which reduces the memory footprint and running time.

Based on the above, we use the following methods for optimization. We improve the inner product method to aggregate the feature volumes from each viewpoint. First, the feature group is preprocessed with the softmax function to make the

✉ Jinguang Chen
xacjg@163.com

Zonghua Yu
xpuyzh@163.com

Lili Ma
xamll@163.com

Kaibing Zhang
zhangkaibing@xpu.edu.cn

[1] School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China

value range of the inner product within [1], reducing the fitting process of regularization. Then, the cost volume of each view is regularized to obtain its probability volume. The probability volume is used as the weight of cost aggregation to weaken the matching noise.

Our method adopts a four-scale cascade structure. The depth map is refined in the second and third stages using the probability volumes from the previous level. In the first stage, the cost volume is constructed via uniform sampling in depth space, and the probability volume is obtained by regularizing the cost volume. Then, the depth map is regressed. At this time, the sampling spacing is wide, and the depth is imprecise. We perform a depth-probability Gaussian distribution fitting using probability volume and depth sampling, refining and correcting the depth in the second stage based on probability thresholds over a wide range. After refinement, the accuracy of the depth has improved, and the probability distribution tends to a sharp single-peaked distribution. At this point, it is fitted to a Laplace distribution, and in the third stage, an exact depth estimate is made in a small interval based on the probability threshold.

To improve the completeness of the reconstruction, the fourth stage should obtain a depth map with the same resolution as the input image. Due to the high accuracy of the depth value obtained in the third stage, the optimization of depth by constructing the cost volume again is limited and affects the efficiency. For this purpose, we designed a lightweight image super-resolution network to upsample the depth map.

The main contributions of our work are as follows:

- View cost aggregation (VCR) is proposed, which regularizes the cost volume from each view constructed by the inner product into a probability volume and uses it as a weight for cost aggregation, weakening the matching noise.
- The depth value is corrected and refined adaptively in an appropriate range by Gaussian distribution fitting and Laplace distribution fitting. High-precision depth estimation is obtained with few depth planes.
- To improve the completeness of the reconstruction, we build a lightweight image super-resolution network to upsample the depth map, which sweetens the network performance effectively.

## 2 Related work

### 2.1 Deep learning-based multi-view stereo

The relevant algorithms for MVS can be broadly organized into volumetric methods and depth map fusion methods [14]. Deep fusion algorithms are flexible and competitive in handling large datasets and are a prominent research direction at present. In the pioneering contribution, Yao et al. [5] used differentiable homography warping and bilinear interpolation to build cost volume, which significantly enhanced the performance of stereo matching and showed the great potential of deep learning technology. The method has four main steps: feature extraction with shared weights, deep hypothesis sampling, cost volume construction, and cost volume regularization. Most current approaches are improvements on the network structure and these steps. For the cost volume construction, there are mainly the variance method and inner product method. Variance-based methods [5–8, 13] treat all perspectives equally and evaluate the similarity of all perspective features. The inner product methods [12] group image features, calculate the inner product of feature groups between each source view and reference view, and weight view visibility through various attention methods [9, 12, 15]. To reduce the memory footprint of cost volume regularization, R-MVSNet [6] regularizes the 2D cost volume sequentially along the depth direction by GRU (gated recurrent unit), and PatchmatchNet [12] simplifies the computation using adaptive spatial cost aggregation. In terms of network structure, the mainstream is the multi-scale structure with a coarse-to-fine manner. CVP-MVSNet [11] uses the cost volume pyramid structure [10, 12, 13]. use a cascade structure to refine the depth map in multiple stages.

### 2.2 Adaptive depth sampling

Generally, multi-scale depth estimate methods need to preset the depth refinement interval in each stage. A reasonable refinement interval should be as narrow as possible while including the ground truth depth value. In this case, a small amount of depth samples can obtain a high-precision depth value. At present, most depth hypothesis methods take a fixed symmetric interval centered on the previous depth values. CasMVSNet [10] uses a preset educing factor to narrow the depth sampling interval at each stage. This requires manually designing a reasonable value and adjusting when the dataset changes. Therefore, an adaptive depth hypothesis approach is necessary. Inspired by the optical flow estimation, CVP-MVSNet [11] calculates the mean value of depth variation corresponding to unit pixel length on epipolar as the depth interval so that the projection points of depth assumptions fall on different pixel points as much as possible, reducing useless depth assumptions. UCSNet [13] proposes a variance-based uncertainty estimate that uses probability volume information to estimate depth intervals. This method effectively improves the quality of depth estimation, proving that the probability volume can be used to refine the depth map. We use the probability volume information to fit Gaussian and Laplace distributions and obtain refinement intervals based on probability thresholds.

### 2.3 Depth map refinement

When the depth map is sufficiently accurate, building the cost volume again for refinement has finite improvement and can consume considerable memory and time. PatchMatchNet [12] designs a residual network based on MSG-Net [16], which adds the estimated residual to the upsampled depth map for refinement. We note that in recent years, learning-based image super-resolution networks have achieved significant performance improvements [17–19], which can effectively improve the resolution of images. Therefore, an attempt is made to enhance the depth map resolution using a lightweight super-resolution network. The experimental results show that the method has a good optimization effect on the depth map.

## 3 Method

Figure 1 illustrates the network architecture of MDF-Net, which adopts a four-scale cascade structure in a coarse-to-fine manner. The network processes images from one reference view and $N$-1 source views, along with their parameter matrices, and then outputs the depth map of the reference view. In the first stage, the coarse depth map is estimated by using the dense equidistant depth hypothesis. In the second and third stages, the probability volume of the previous stage is used to fit the Gaussian distribution and Laplace distribution, respectively. Then, the depth is refined or corrected in a small depth range according to the probability threshold. In the fourth stage, the depth map is upsampled to the resolution of the input image using a lightweight image super-resolution network. In addition, to reduce the resolution limitation of the input image, we reduced the downsampling process in the network. The 3D CNN network in the first stage is only downsampled twice.

### 3.1 Multi-scale feature map

Similar to PatchMatchNet, we use the FPN [20] structure to extract multi-scale features. A weight sharing approach is used to extract multi-scale feature maps for input images of all viewpoints. The channels of the output feature map are 64, 32, 16, and the resolutions are 1/8 × 1/8, 1/4 × 1/4, 1/2 × 1/2, which are used as the input of each stage.

### 3.2 View cost regularization for aggregation

At each stage, the feature of each source view is mapped to the reference view based on the hypothetical depth using the differentiable homography warping and bilinear interpolation. For the pixel position $p_0(x,y)$ in the reference view, when the depth is $d_0$, the projection coordinates in the $i$th source view are calculated as

$$P_i = K_i R_i (R_0^T K_0^{-1} p_0 d_0 + (R_i^T T_i - R_0^T T_0)) \tag{1}$$

where $K_0$, $R_0$, $T_0$, $K_i$, $R_i$, and $T_i$ are the camera intrinsics, rotation matrix, and translation vector of the reference view and the $i$th source view, respectively. $P_i$ is the homogeneous coordinate after $p_0$ mapping.

After warping the features, the similarity of the features across perspectives needs to be measured to determine the reliability of the depth hypothesis. We first preprocess the feature groups with the softmax function and then use the inner product operation for measurement. Meanwhile, to weaken the effect of noise points, we designed the viewpoint cost regularization network (VCR-Net), which regularizes the cost volumes of each view with continuous 1 × 1 convolution to obtain the probability volumes. It is used as a weight to aggregate the cost volumes in a weighted average manner. The network structure of VCR-Net is shown in "Appendix".

Let $F \in \Re^{C\times D\times H\times W}$ be the feature volume after warping, where $C$, $D$, $H$, and $W$ are the number of channels, depth samples, height, and width of the feature map, respectively. The feature channels are divided into $G$ groups, and each feature group is normalized by the softmax operation, $F \in \Re^{G\times \frac{C}{G}\times D\times H\times W}$.

Next, the similarity of each perspective feature is measured. Let $F_0$ and $F_i$ be the feature volumes of the reference view and the $i$th source view, respectively. The inner product of the reference view and the $i$th source view in the $g$th group is calculated as follows:

$$S_i^g = \left\langle F_0^g, F_i^g \right\rangle \tag{2}$$

where $\langle\cdot,\cdot\rangle$ is the inner product.

Then, the weighted average of all perspectives is calculated to obtain the cost volume.

$$S(p) = \frac{\sum_{i=1}^{N-1} W_i S_i}{\sum_{i=1}^{N-1} W_i} \tag{3}$$

where the view weights $W_i = \mathrm{VCR{-}Net}(S_i)$.

### 3.3 Regression depth and confidence map

Similar to the general MVS algorithm, to filter the cost volume noise, the 3D CNN is used to gather neighborhood information from different receptive field. In the first three stages, we regularize the cost volume with the 3D UNet [21] structure and normalize the depth dimension with the softmax function to obtain the probability volume. To reduce the network's constraints on the input image resolution, we reduce downsampling in the network. The number of 3D UNet downsamplings for the first three stages is 2, 3, and

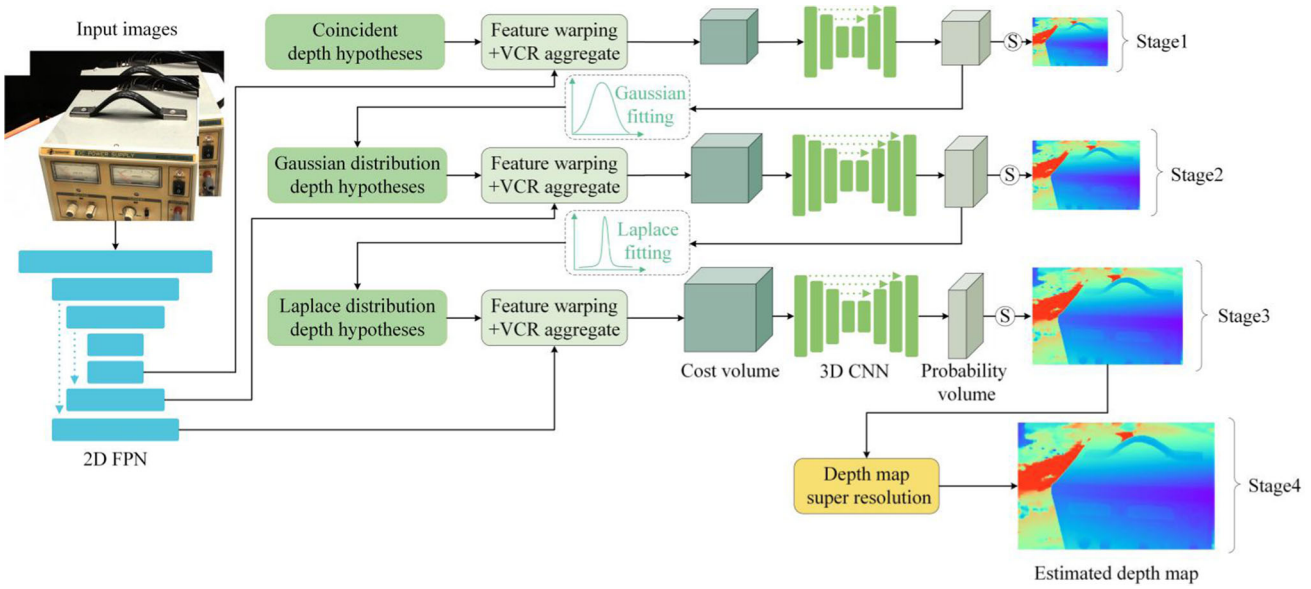



**Fig. 1** Architecture of MDF-Net

3. Once the probability volume is obtained, the depth is regressed using soft argmin [22] along the depth dimension, and the confidence map is calculated in the same way as MVSNet.

### 3.4 Multi-distribution fitting

The probability distribution of the probability volume reflects the reliability of each depth hypothesis, which can guide the next stage of depth refinement. In the first and second stages, we use Gaussian and Laplace distributions to fit the depth probability curve separately. The first stage is sampled at the maximum depth range to obtain a coarse depth map when the probability distribution of depth is more dispersed. Therefore, fit it to a Gaussian distribution. In the second stage, the probability distribution after one optimization has been more concentrated, showing a sharp unimodal distribution, which is fitted to the Laplace distribution. After obtaining the depth probability curve, the depth interval corresponding to the greater than probability threshold is used as the new depth hypothesis interval for equally spaced sampling. Meanwhile, the probability distributions of adjacent locations are similar, making the depth hypotheses of local regions more regular in 3D space, which provides useful contextual information for regularization. This process is shown in Fig. 2c, where the depth refinement interval obtained by Laplace fitting in the second stage is only approximately 2 mm, allowing the third stage to obtain very accurate depth values.

*(1) Gaussian distribution fitting:* For simplicity, any pixel position is discussed. We take the depth hypothesis $d$ as the independent variable and the probability volume value $p$ as the dependent variable for curve fitting.

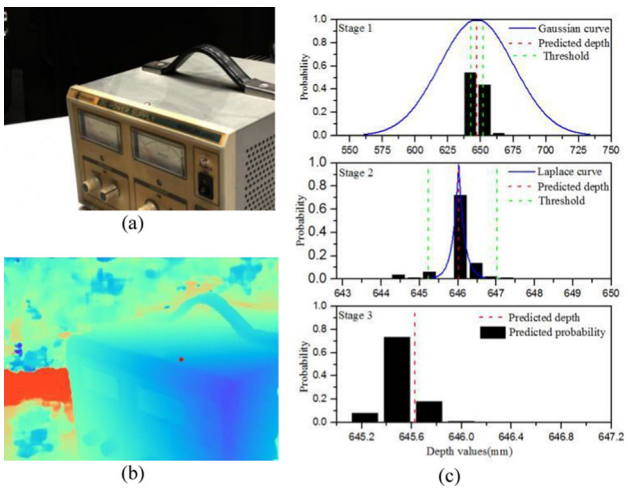



**Fig. 2** Deep refinement process of the red pixel position. **a** RGB image, **b** predicted depth, **c** depth hypothesis at different stages

For Gaussian distribution fitting, the target curve is set as

$$p(d) = \exp\left(-\frac{(d-\hat{d})^2}{\sigma}\right) \tag{4}$$

To enhance the fitting power, the depth values $\hat{d}$ obtained by regression were replaced by the parameter $\mu$ during fitting.

$$p(d) = \exp\left(-\frac{(d-\mu)^2}{\sigma}\right) \tag{5}$$

First, perform a logarithmic transformation,

$$Ln(p) = -\frac{d^2}{\sigma} + \frac{2\mu d}{\sigma} - \frac{\mu^2}{\sigma} \tag{6}$$

Then, the following matrices are constructed:

$$Z = \begin{bmatrix} Ln(p_0) \\ \vdots \\ Ln(p_n) \end{bmatrix} \quad X = \begin{bmatrix} d_0^2 & d_0 & 1 \\ \vdots & \vdots & \vdots \\ d_n^2 & d_n & 1 \end{bmatrix} \quad B = \begin{bmatrix} -\frac{1}{\sigma} \\ \frac{2\mu}{\sigma} \\ -\frac{\mu^2}{\sigma} \end{bmatrix} \tag{7}$$

From the least square method,

$$B = (X^T X)^{-1} X^T Z = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} \tag{8}$$

$$\sigma = -\frac{1}{b_0} \tag{9}$$

At this time, the parameter $\hat{d}$ is used to replace $\mu$. The probability distribution is expressed as

$$p(d) = \exp(b_0(d - \hat{d})^2) \tag{10}$$

Given a probability threshold $p_{\text{Th}}$, the new depth hypothesis interval is $[\hat{d} - (Ln(p_{Th}) / b_0)^{1/2}, \hat{d} + (Ln(p_{Th}) / b_0)^{1/2}]$.

*(2) Laplace distribution fitting:* The target curve fitted by the Laplace distribution is

$$p(d) = \exp\left(-\frac{|d - \hat{d}|}{\gamma}\right) \tag{11}$$

Performing logarithmic transformation on Eq. (11)

$$Ln(p) = -\frac{|d - \hat{d}|}{\gamma} \tag{12}$$

Then, the following matrices are constructed:

$$Z = \begin{bmatrix} Ln(p_0) \\ \vdots \\ Ln(p_n) \end{bmatrix} \quad X = \begin{bmatrix} |d_0 - \hat{d}| \\ \vdots \\ |d_n - \hat{d}| \end{bmatrix} \quad B = \begin{bmatrix} -\frac{1}{\gamma} \end{bmatrix} \tag{13}$$

From the least square method,

$$B = (X^T X)^{-1} X^T Z = [b_0] \tag{14}$$

$$\gamma = -\frac{1}{b_0} \tag{15}$$

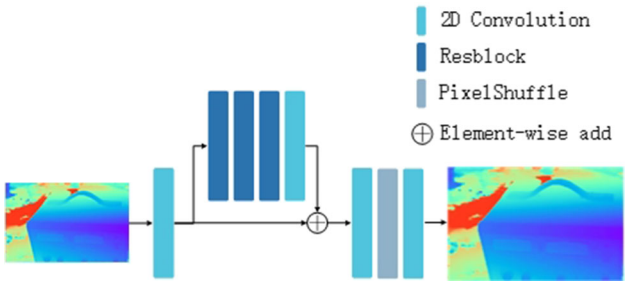



**Fig. 3** Super-resolution network for depth map upsampling

The probability distribution is expressed as

$$p(d) = \exp(b_0 |d - \hat{d}|) \tag{16}$$

Given a probability threshold $p_{\text{Th}}$, the new depth hypothesis interval is $[\hat{d} - Ln(p_{Th}) / b_0, \hat{d} + Ln(p_{Th}) / b_0]$.

### 3.5 Depth map refinement

The third stage obtained very accurate depth values in the depth hypothesis interval fitted by the Laplace distribution. At this time, the effect of building the cost volume in the fourth stage to refine the depth is small. Therefore, an image super-resolution network is used here to upsample the depth map and try to maintain depth accuracy. As shown in Fig. 3, inspired by the structure of EDSR [23], we design a lightweight image super-resolution network.

The depth map is first normalized, and then the depth information is transformed into feature information by a convolutional layer. Then, the higher-level semantic information is obtained by successive residual modules, and they are summed at the element level. Finally, upsampling is performed using the PixelShuffle layer, whose main function is to convolute the low-resolution feature map and recombine the multi-channel feature map to obtain a high-resolution feature map. Then, the normalized values are mapped back to the depth interval to obtain a high-resolution depth map.

### 3.6 Loss function

The loss function of the network is defined as the smooth L1 loss of the estimated depth map in all stages and the ground truth depth map $D_{\text{GT}}$.

$$L_{\text{depth}} = \sum_{l=1}^{4} \sum_{p \in \Omega} ||\mathbf{D}^l(p) - \mathbf{D}^l_{\text{GT}}(p)||_1 \tag{17}$$

where $\Omega$ is the set of valid pixel in $D_{\text{GT}}$, and $D^l$ represents the estimated depth map of the $l$th stage.

## 4 Experiments

This section presents the details of MDF-Net training and evaluation. Then, its effectiveness and generality are verified.

### 4.1 Datasets

The DTU dataset [24] provides 124 indoor scene data, and each scene contains 49 fixed-view images and camera information under seven lighting conditions. For DTU datasets, we follow the same dataset division method as other MVS methods [5, 10–13].

The BlendedMVS dataset [25] provides more than 17 k high-quality training samples covering various scenes for multi-view depth estimation. The model trained on this dataset has better generalization ability.

The Tanks and Temples dataset [26] contains advanced set and intermediate set that are provided by a collection of video sequences and serve as a benchmark for measuring network performance.

We train the model on DTU training set and evaluate the performance on DTU test set. On Tanks and Temples dataset, we use the model trained on the BlendedMVS dataset to verify the generalization ability.

### 4.2 Training

For the DTU dataset, images with a resolution of 640 × 512 are used for training, the initial depth range is [425.0 mm, 935.0 mm], and the batch size is set to 4. The input image resolution of the BlendedMVS dataset is 768 × 576, using the respective initial depth range of each viewpoint, and the batch size is set to 6. In the first, second, and third stages, 48, 24, and 8 sampling planes are taken, respectively. The probability thresholds for the Gaussian and Laplace distributions are set to 0.95 and 1e-5, respectively. Furthermore, to enrich the training set and enhance the stability of VCR-Net, a robust training strategy [12] is used to randomly select four viewpoints from the ten best source viewpoints. We implement the network model with the PyTorch [27] framework and optimize the parameters with the Adam [28] optimizer. The initial learning rate is set to 0.001, and the training is performed on an NVIDIA GeForce RTX 3090 GPU for 30 epochs using the same learning rate decay strategy as DB [29].

### 4.3 Fusion

For the DTU dataset, the depth map fusion process is performed using the method of [4] for a fair comparison. Since the scenarios of the Tanks and Temples dataset are large, the method in [4] requires high GPU memory. Therefore, we use the method in [30] to fuse depth maps and optimize the dynamic geometric consistency checking method proposed by [31]. The dynamic geometric consistency checking method sets different thresholds. The strict threshold uses a small number of views to check; in contrast, a large number of views to check, and then all the depths that meet the conditions are regarded as effective depths. Although this strategy obtains more reliable depth estimates, it also introduces depth with large error, resulting in point cloud noise. Therefore, we regard the depth that meets more than half of the conditions as the credible depth, which effectively reduces the point cloud noise.

**Table 1** Quantitative results of reconstruction quality on the DTU test set (lower is better)

| Methods | Acc. (mm) | Comp. (mm) | Overall (mm) |
|---|---|---|---|
| Camp [1] | 0.835 | 0.554 | 0.695 |
| Gipuma [4] | **0.283** | 0.873 | 0.578 |
| MVSNet [5] | 0.396 | 0.527 | 0.462 |
| R-MVSNet [6] | 0.383 | 0.452 | 0.417 |
| Point-MVSNet [7] | 0.342 | 0.411 | 0.376 |
| Fast-MVSNet [8] | 0.336 | 0.403 | 0.370 |
| PVA-MVSNet [9] | 0.379 | 0.336 | 0.357 |
| CasMVSNet [10] | 0.325 | 0.385 | 0.355 |
| CVP-MVSNet [11] | <u>0.296</u> | 0.406 | 0.351 |
| PatchMatchNet [12] | 0.427 | **0.277** | 0.352 |
| UCSNet [13] | 0.338 | 0.349 | 0.344 |
| MDF-Net (ours) | 0.349 | <u>0.303</u> | **0.326** |

### 4.4 Benchmark performance

*(1) Evaluation on DTU dataset:* We use input images with a resolution of 1600 × 1184 to estimate a depth map with information from 5 viewpoints. Afterward, the point cloud is fused. Then, the accuracy (Acc), completeness (Comp), and overall quality of the estimated point cloud are calculated by the MATLAB script provided by the DTU dataset. Accuracy is measured by the distance from the estimated point cloud to the ground truth cloud, while completeness refers to the distance from the ground truth point cloud to the estimated point cloud. The overall score is the average of accuracy and completeness. The experimental environment is a Linux server with Quadro RTX 5000 GPU, PyTorch version 1.10.0, and CUDA version 11.4. Table 1 shows the quantitative results of our method and other advanced methods. Underlined text indicates the second-best result for each column, while bold text indicates the best result for each column. Our method is superior to all other methods in overall quality and has excellent completeness. Compared with the other adaptive depth

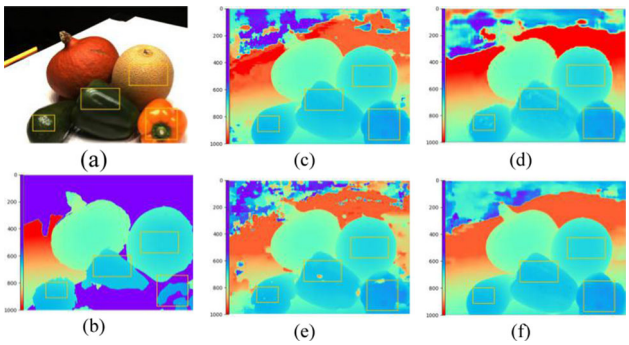



**Fig. 4** Comparison of depth map results on the DTU dataset. **a** RGB image. **b** Ground truth depth. **c** CasMVSNet. **d** CVP-MVSNet. **e** UCSNet. **f** Ours

sampling methods, the overall quality is improved by approximately 7.12% and 5.23% over CVP-MVSNet and UCSNet, respectively.

To verify the effectiveness of the method, we tested the reconstruction effect with the output depth map from each stage on the DTU test set, and the quantitative results are shown in Table 2. In the first stage, 48 depth planes are used to estimate the depth map in the preset maximum depth range when the depth map resolution is low and the depth error is large, with an overall quality of 0.751 mm. After refinement with Gaussian distribution fitting, the overall quality is improved by 0.290 mm in the second stage, achieving performance comparable to that of MVSNet. After fitting with the Laplace distribution, it improves again by 0.112 mm, reaching a performance comparable to advanced methods. After upsampling by the image super-resolution network in the fourth stage, the overall quality surpasses other methods and achieves state-of-the-art results. This demonstrates the effectiveness of fitting the distribution to refine the depth and image super-resolution networks.

Figure 4 shows the depth map performance of MDF-Net and other advanced methods on the 75th scene of the DTU dataset. We use the codes and pretrained models provided in [10, 11, 13] for testing, filtering the depth map obtained by each method to make it in the same depth range. It can be seen that our method has good results in weakly textured areas and reflective surfaces, and the depth map behaves smoothly. At the same time, the accuracy is extremely high, and even on low-curvature surfaces, there are significant differences in depth.

*(2) Generalization on tanks and temple datasets:* To verify the generalization ability, the model is trained on the BlendedMVS dataset with flexible camera trajectories and different scene depths and is evaluated on Tanks and Temples dataset. The height of the input image is adjusted to 1056, while 10 views of the image are input for depth estimate. The reconstructed point clouds were submitted to the benchmark website of Tanks and Temples (https://www.tanksandtemples.org/leaderboard/) for evaluation to obtain the F-Score of each scene.

In addition, to explore the potential of MDF-Net, this section ignores some performance costs and designs the network as a three-scale structure similar to CasMVSNet and UCSNet (Without using image super-resolution networks, directly extract feature maps at 1/4 × 1/4, 1/2 × 1/2 and 1 × 1 input resolution to construct cost volumes), trained and evaluated under the same conditions. Tables 3 and 4 present the published F-scores on Tanks and Temples website for our method versus the other methods. MDF-Net achieves the best performance in the Lighthouse, Panther, Playground, Train, Courtroom, and Temple scenarios, and the mean F-Score is comparable to CasMVSNet, PatchMatchNet, and UCSNet, with competitive generalization capability. At the same time, the reconstruction effect of MDF-Net under the three-scale structure has significantly improved, achieving the highest F-score in both the intermediate set and advanced set.

*3) Memory and Run-time Comparison:* Fig. 5 shows the running time and GPU memory usage for each method with different output depth map resolutions. Compared with other methods, MDF-Net uses the least number of depth planes, only 80. Second, VCR-Net uses a 1 × 1 convolutional kernel, which is extremely computationally minimal. Meanwhile, the lightweight image super-resolution network is used in the fourth stage, which greatly saves memory and running time. Therefore, the performance of our method is optimal in high-resolution scenes, with the lowest memory usage and the fastest running speed. Inferring a depth map with a resolution of 1600 × 1184 takes only approximately 4396 M GPU memory and 0.376 s time. As the resolution increases, the memory footprint and runtime grow at a gentle rate.

## 4.5 Ablation study

*(1) View cost regularization for aggregation:* To verify the effectiveness of view cost aggregation, we remove the softmax operation and VCR-Net in building the cost volume, and use a fixed optimal perspective to train the model in the same experimental environment for comparison. As shown in Table 5, when cost volume build is performed without the VCR module, the accuracy and completeness are reduced by approximately 0.009 mm and 0.011 mm, respectively. This suggests that view cost aggregation can improve the quality of the cost volume to increase the effectiveness of the reconstruction. Additionally, when the VCR module is not used, the runtime is reduced by 0.06 s and becomes faster.

*(2) Probability threshold:* According to Eqs. (10) and (16), the width of the depth refinement interval is negatively correlated with the probability threshold. That is to say, the smaller the probability threshold is, the larger the interval range is. An interval that is too large will affect the accuracy of depth prediction, while an interval that is too narrow may reduce

**Table 2** Quantitative results of MDF-Net on the DTU test set at different stages (lower is better)

| Stage | Resolution | Acc.(mm) | Comp.(mm) | Overall(mm) |
|---|---|---|---|---|
| 1 | 200 × 148 | 0.871 | 0.631 | 0.751 |
| 2 | 400 × 296 | 0.431 | 0.491 | 0.461 |
| 3 | 800 × 592 | 0.391 | 0.306 | 0.349 |
| 4 | 1600 × 1184 | 0.349 | 0.303 | 0.326 |

**Table 3** F-score on the intermediate set of Tanks and Temples Dataset (higher is better)

| Method | Family | Francis | Horse | Lighthouse | M60 | Panther | Playground | Train | Mean |
|---|---|---|---|---|---|---|---|---|---|
| MVSNet | 55.99 | 28.55 | 25.07 | 50.79 | 53.96 | 50.86 | 47.90 | 34.69 | 43.48 |
| R-MVSNet | 69.96 | 46.65 | 32.59 | 42.95 | 51.88 | 48.80 | 52.00 | 42.38 | 48.40 |
| Point-MVSNet | 61.79 | 41.15 | 34.20 | 50.79 | 51.97 | 50.85 | 52.38 | 43.06 | 48.27 |
| PVA-MVSNet | 69.36 | 46.80 | 46.01 | 55.74 | 57.23 | 54.75 | 56.70 | 49.06 | 54.46 |
| PatchMatchNet | 66.99 | 52.64 | 43.24 | 54.87 | 52.87 | 49.54 | 54.21 | 50.81 | 53.15 |
| CVP-MVSNet | 76.50 | 47.74 | 36.34 | 55.12 | 57.28 | 54.28 | 57.43 | 47.54 | 54.03 |
| UCSNet | 76.09 | 53.16 | 43.03 | 54.00 | 55.60 | 51.49 | 57.38 | 47.89 | 54.83 |
| CasMVSNet | 76.37 | 58.45 | 46.26 | 55.81 | 56.11 | 54.06 | 58.18 | 49.51 | 56.84 |
| MDF-Net (ours) | 75.22 | 50.12 | 45.38 | 57.97 | 54.93 | 55.06 | 58.27 | 52.47 | 56.18 |
| MDF-Net (ours, three-scale) | **78.64** | **61.72** | **49.37** | **60.70** | **58.77** | **56.31** | **59.97** | **56.43** | **60.24** |

**Table 4** F-score on the advanced set of Tanks and Temples Dataset (higher is better)

| Method | Auditorium | Ballroom | Courtroom | Museum | Palace | Temple | Mean |
|---|---|---|---|---|---|---|---|
| R-MVSNet | 12.55 | 29.09 | 25.06 | 38.68 | 19.14 | 24.96 | 24.91 |
| PatchMatchNet | 23.69 | 37.73 | 30.04 | 41.80 | 28.31 | 32.29 | 32.31 |
| CasMVSNet | 19.81 | 38.46 | 29.10 | 43.87 | 27.36 | 28.11 | 31.12 |
| MDF-Net (ours) | 24.61 | 40.14 | 34.50 | 41.79 | 30.51 | 36.64 | 34.70 |
| MDF-Net (ours, three-scale) | **25.42** | **42.50** | **36.29** | **46.40** | **33.72** | **39.55** | **37.31** |

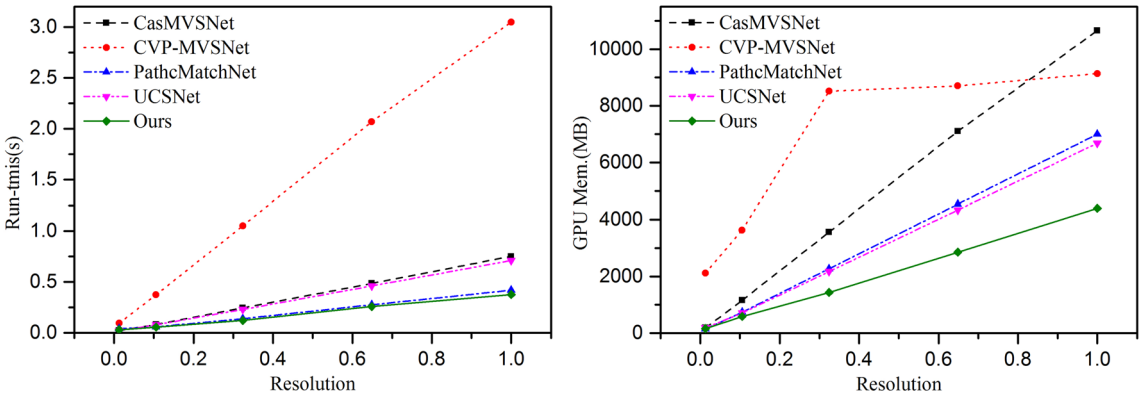



**Fig. 5** Comparison of runtime and GPU memory usage on the DTU dataset for different output sizes (Resolution 1.0 is 1600 × 1184)

**Table 5** Comparison of using VCR on the DTU test set (lower is better)

| VCR | Acc.(mm) | Comp.(mm) | Overall(mm) | Time.(s/view) |
|---|---|---|---|---|
| × | 0.358 | 0.314 | 0.336 | 0.316 |
| √ | 0.349 | 0.303 | 0.326 | 0.376 |

**Table 6** Performance of different probability thresholds on the DTU test set (lower is better)

| Gaussian threshold | Laplace threshold | Acc.(mm) | Comp.(mm) | Overall(mm) |
|---|---|---|---|---|
| 0.925 | 1e−5 | 0.353 | 0.303 | 0.328 |
| 0.95 | 1e−5 | 0.349 | 0.303 | 0.326 |
| 0.975 | 1e−5 | 0.347 | 0.307 | 0.327 |
| 0.95 | 1e−3 | 0.348 | 0.310 | 0.329 |
| 0.95 | 1e−1 | 0.347 | 0.321 | 0.334 |

the rectified ability. Therefore, a reasonable threshold is necessary. The reconstruction results under different probability thresholds are shown in Table 6. When the Gaussian probability threshold is less than 0.95, the depth refinement interval is too large, which reduces the accuracy. When the Gaussian probability threshold is greater than 0.95, the interval is too narrow, which has a greater impact on completeness. With the increase in the Laplacian threshold, the refinement interval becomes narrower, and the prediction accuracy increases, but the rectified ability decreases, which reduces the completeness. It is more appropriate to take 1e-5.

## 5 Conclusion

This paper proposes MDF-Net, which uses appropriate distribution functions to fit the probability of depth hypotheses at different stages to adaptively obtain high-quality depth refinement intervals, which greatly improves the quality of depth estimation. At the same time, we propose view cost regularization to weaken the matching noise, which effectively improves the reliability of the cost volume. Finally, high-resolution depth estimation is obtained with minimal performance cost using image super-resolution techniques. Experimental results show that MDF-Net achieves state-of-the-art results in reconstruction quality, memory consumption, and running time, with competitive generalization ability. In the future, we will consider using learned probability thresholds to obtain more efficient depth refinement intervals.

**Funding** This work was supported by the National Natural Science Foundation of China under Grant 61971339 and 61471161, the Natural Science Basic Research Program of Shaanxi under Grant 2023-JC-YB-826, the Scientific Research Program Funded by Shaanxi Provincial Education Department under Grant 22JP028, and the Postgraduate Innovation Fund of Xi'an Polytechnic University under Grant chx2022019.

## Appendix

### Why use softmax to preprocess feature groups?

Ideally, the more similar the features of different perspectives on the same depth plane, the closer the plane is to the true depth, and the higher the probability. Therefore, we believe that a good cost measurement method should satisfy the following points. First, the performance of the similarity measurement is good. Second, the more similar the features of the depth plane are, the larger the cost is, which is proportional. Third, the value range of the cost metric is consistent with the probability value range, which is between [1]. We choose the inner product operation as the main method of the cost metric (meeting the first point). In addition, compared to vector normalization, the gradient calculation of softmax normalization is simpler, and the value range of the inner product is kept in [1]. The feature group pretreated by softmax function reduces the fitting process, making VCR-Net and 3D CNN more efficient.

### Why can VCR-Net improve the cost volume quality?

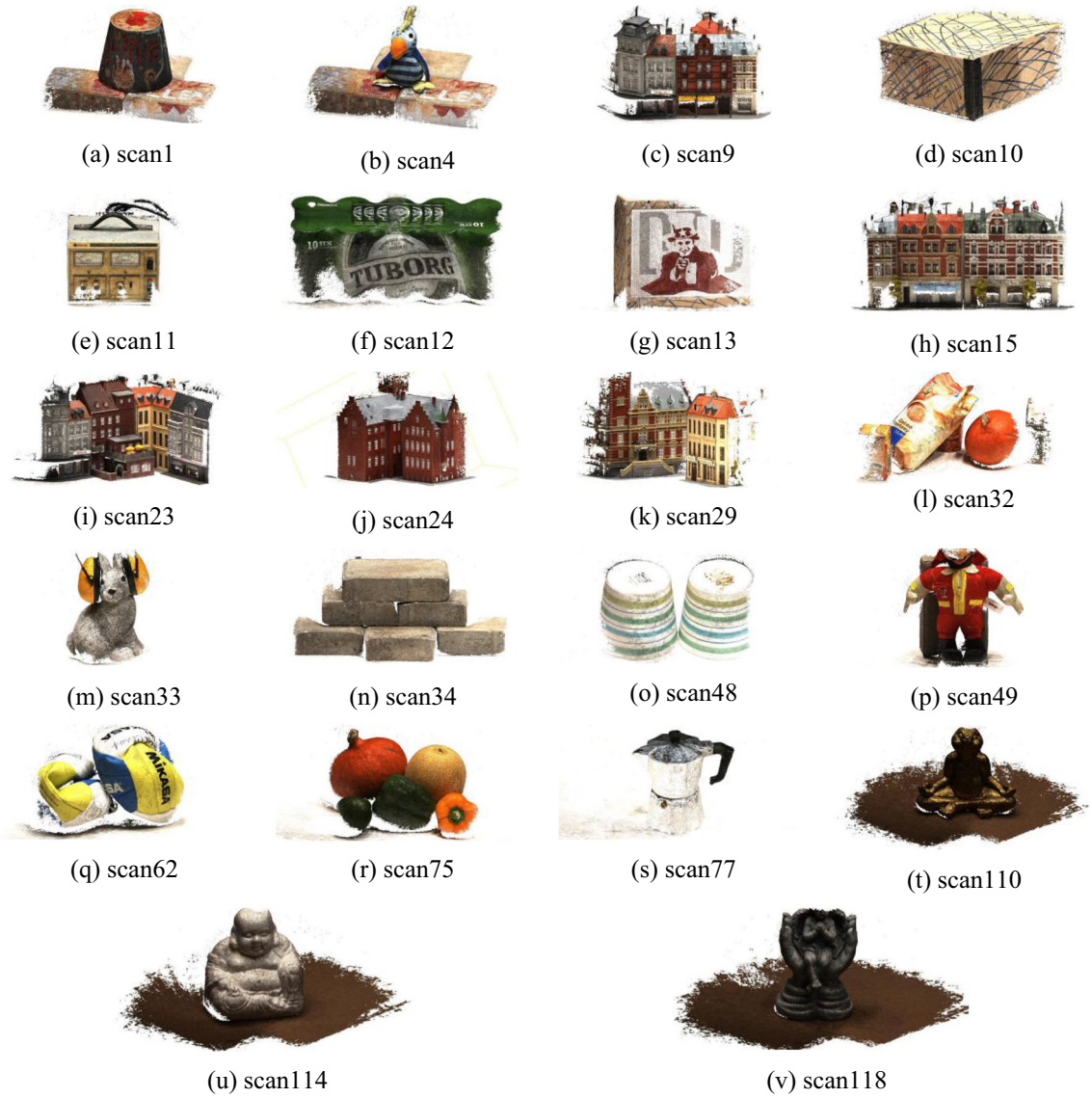
The functions of VCR-Net and the 3D CNN regularization network are similar. Both regularize the cost volume to obtain the probability value of each depth plane. The difference is that VCR-Net processes the cost volume of each view separately, takes the probability volume as the weight, and uses the sigmoid activation function. The features of the noise locations are mismatched, with low similarity, and a small weight is obtained by VCR-Net. When calculating the weighted average, a low weight is given to the noise to reduce its matching cost. The VCR-Net network structure is shown in Table 7.

### Visualization of point cloud results

All qualitative results of our method are shown in Figs. 6 and 7.

**Table 7** Details of VCR-Net, where G is the number of feature groups, D is the number of depth samples, and H and W are the width and height of the feature map, respectively

| Name | Operation | Kernel size | Stride | Batch normalization | Activation function | Output Size |
|---|---|---|---|---|---|---|
| Input | | | | | | G × D × H × W |
| Conv | 3D CNN | 1 | 1 | √ | ReLU | 1 × D × H × W |
| Output | 3D CNN | 1 | 1 | × | Sigmoid | 1 × D × H × W |




(a) scan1 (b) scan4 (c) scan9 (d) scan10

(e) scan11 (f) scan12 (g) scan13 (h) scan15

(i) scan23 (j) scan24 (k) scan29 (l) scan32

(m) scan33 (n) scan34 (o) scan48 (p) scan49

(q) scan62 (r) scan75 (s) scan77 (t) scan110

(u) scan114 (v) scan118

**Fig. 6** All qualitative results on the DTU test set. Our method achieves excellent completeness and the best overall quality


**Fig. 7** All qualitative results of the Tanks and Temples dataset. The results are comparable to those of advanced methods

## References


1. Campbell, N.D.F., Vogiatzis, G., Hernández, C., Cipolla, R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. In: Proceedings of European Conference on Computer Vision (ECCV), pp. 766–779 (2008)
2. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. **32**(8), 1362–1376 (2010)
3. Tola, E., Strecha, C., Fua, P.: Efficient large-scale multi-view stereo for ultra high-resolution image sets. Mach. Vis. Appl. **23**(5), 903–920 (2012)
4. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 873–881 (2015)
5. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: MVSNet: depth inference for unstructured multi-view stereo. In: Proceedings of European Conference on Computer Vision (ECCV), pp. 785–801 (2018)
6. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent MVSNet for high-resolution multi-view stereo depth inference. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5520–5529 (2019)
7. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 1538–1547 (2019)
8. Yu, Z., Gao, S.: Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1946–1955 (2020)
9. Yi, H., et al.: Pyramid multi-view stereo net with self-adaptive view aggregation. In: Proceedings of European Conference on Computer Vision (ECCV), pp. 766–782 (2020)
10. Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P.: Cascade cost volume for high-resolution multi-view stereo and stereo matching. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2492–2501 (2020)
11. Yang, J., Mao, W., Alvarez, J.M., Liu, M.: Cost volume pyramid based depth inference for multi-view stereo. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4876–4885 (2020)
12. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: PatchmatchNet: learned multi-view patchmatch stereo. In: Proceedings

of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14189–14198 (2021)
13. Cheng, S., et al.: Deep stereo using adaptive thin volume representation with uncertainty awareness. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2521–2531 (2020)
14. Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L.: SurfaceNet: an end-to-end 3d neural network for multiview stereopsis. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 2326–2334 (2017)
15. Zhang, J., Yao, Y., Li, S., Luo, Z., Fang, T.: Visibility-aware multi-view stereo network. Int. J. Comput. Vis. **131**(1), 199–214 (2023)
16. Hui, T., Loy, C.C., Tang, X.: Depth map super-resolution by deep multi-scale guidance. In: Proceedings of European Conference on Computer Vision (ECCV), pp. 353–369 (2016)
17. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1646–1654 (2016)
18. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1637–1645 (2016)
19. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 105–114 (2017)
20. Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936–944 (2017)
21. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Proceedings of International Conference MICCAI, pp. 234–241 (2015)
22. Kendall, A., et al.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 66–75 (2017)
23. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1132–1140 (2017)
24. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. Int. J. Comput. Vis. **120**(2), 153–168 (2016)
25. Yao, Y., et al.: BlendedMVS: a large-scale dataset for generalized multi-view stereo networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1787–1796 (2020)
26. Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V.: Tanks and temples: benchmarking large-scale scene reconstruction. ACM Trans. Graph. **36**(4), 1–13 (2017)
27. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: Proceedings of NIPS Autodiff Workshop (2017)
28. Kingma, D.P., Ba, J.: Adam: a Method for Stochastic Optimization. arXiv, Jan. 29, 2017. Accessed: May 28, 2022. [Online]. http://arxiv.org/abs/1412.6980
29. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. AAAI **34**(07), 11474–11481 (2020)
30. Merrell, P., et al.: Real-time visibility-based fusion of depth maps. In: Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 1–8 (2007)
31. Yan, J., et al.: Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In: Proceedings of European Conference on Computer Vision (ECCV), pp. 674–689 (2020)



**Jinguang Chen** received B.S. degree and M.Sc. degree from Xi'an Polytechnic University in 2000 and 2005, respectively. In 2011, he obtained Ph.D. degree in Intelligent Information Processing from Xidian University. He is currently a Professor at the School of Computer Science, Xi'an Polytechnic University. His interests include information fusion, target tracking, machine learning, etc.

**Zonghua Yu** received the B.S. degree in Electronic Science and Technology from Xi'an University of Technology, Xi'an, China, in 2018. He is currently pursuing his M.Sc. degree at the School of Computer Science, Xi'an Polytechnic University, Xi'an, China. His research interests include deep learning and computer vision.

**Lili Ma** received B.S. degree from Northwest Normal University and M.Sc. degree from Xi'an Polytechnic University in 2002 and 2007, respectively. She is currently an Associate Professor at the School of Computer Science, Xi'an Polytechnic University. Her interests include information fusion, target tracking, machine learning, etc.

**Kaibing Zhang** received the M.Sc. degree in Computer Software and Theory from Xihua University, Chengdu, China, in 2005, and the Ph.D. degree in Pattern Recognition and Intelligent System from Xidian University, Xi'an, China, in 2012, respectively. He is currently a Professor at the School of Electronics and Information, Xi'an Polytechnic University, Xi'an, China. His main research interests include pattern recognition, computer vision, and image super-resolution reconstruction. In these areas, he has published around 40 technical articles in refereed journals and proceedings including IEEE TIP, IEEE TCSVT, IEEE TNNLS, signal processing (Elsevier), neurocomputing, applied soft computing, neural networks, CVPR, ICIP, etc.