EPP-MVSNet: Epipolar-assembling based Depth Prediction for Multi-view Stereo

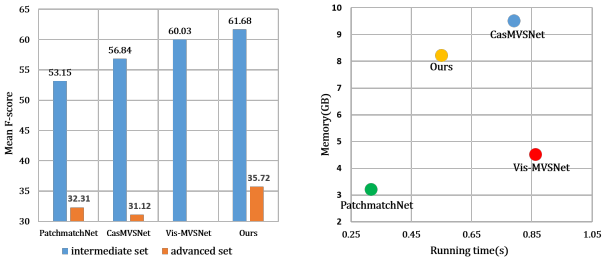
Xinjun Ma[1]* Yue Gong[1]* Qirui Wang[1] Jingwei Huang[1] Lei Chen[2]† Fan Yu[1]
[1]Distributed and Parallel Software Lab, Huawei Technologies
[2]Department of Computer Science and Engineering, Hong Kong University of Science and Technology
{maxinjun1, gongyue1, wangqirui1, huangjingwei6, fan.yu}@huawei.com leichen@cse.ust.hk

Abstract

In this paper, we proposed EPP-MVSNet, a novel deep learning network for 3D reconstruction from multi-view stereo (MVS). EPP-MVSNet can accurately aggregate features at high resolution to a limited cost volume with an optimal depth range, thus, leads to effective and efficient 3D construction. Distinct from existing works which measure feature cost at discrete positions which affects the 3D reconstruction accuracy, EPP-MVSNet introduces an epipolar-assembling-based kernel that operates on adaptive intervals along epipolar lines for making full use of the image resolution. Further, we introduce an entropy-based refining strategy where the cost volume describes the space geometry with the little redundancy. Moreover, we design a light-weighted network with Pseudo-3D convolutions integrated to achieve high accuracy and efficiency. We have conducted extensive experiments on challenging datasets Tanks & Temples(TNT), ETH3D and DTU. As a result, we achieve promising results on all datasets and the highest F-Score on the online TNT intermediate benchmark. Code is available at `https://gitee.com/mindspore/mindspore/tree/master/model_zoo/research/cv/eppmvsnet`.

1. Introduction

Dense 3D reconstruction from multi-view stereo (MVS) is a fundamental problem that has been studied for decades, where dense correspondences are computed among multiple images and used to determine the dense geometry. Typically, correspondences can be established for each patch of a reference image by searching its optimal matching patch among target images [2]. Alternatively, [10] computes matching costs by exhaustively sampling pixels on the epipolar lines and storing them into a cost volume, which is used to determine the final depth map. However, both directions encounter the challenge on how to estimate depth accurately and efficiently, especially in the real world scenarios filled with noises and smooth texture, existing solutions require high computation cost, but often achieve unsatisfactory reconstruction quality. While recent deep learning-based solutions [9, 22] address these issues and further enhance reconstruction quality, they still suffer from high memory and computation requirement for constructing and regularizing cost volumes, which make them unable to make fully usage of high resolution images.



(a) Mean F-score (b) Run-time and memory

Figure 1. Comparison between the proposed EPP-MVSNet and state-of-the-art learning-based multi-view stereo methods [9, 19, 27] on reconstruction quality in (a) and run-time and memory requirement in (b) with input images resolution of 1920×1056 on Tanks & Temples dataset [13].

In this paper, we aim at designing a deep neural network to fully utilize the information of high resolution images. Though high-resolution cost volume is memory consuming, patch-match-based methods can overcome this issue by searching for the minimum cost rather than storing all the cost at the given image resolution. We further develop this idea and propose an epipolar-assembling module that assembles matching cost along the epipolar line. Specifically, our epipolar-assembling module constructs a com-

*Equal contribution.
†Corresponding author.

pact cost volume by assembling densely interpolated features and further reducing the volume size by adaptive pooling. Rather than increasing the resolution, our module assembles high-resolution cost volume into a coarse volume resolution and only introduces an one-dimensional aggregation and pooling complexity. Furthermore, we introduce an entropy-based refining strategy for reducing redundancy and information loss of the constructed fine cost volume.

In proposed deep learning model, the most expensive operator is 3D convolution, thus, in this work, we further reduce the cost by replacing a 3D convolution with a Pseudo-3D convolution and developing a light-weighted structure for cost volume regularization. Our experiments show that such change does not hurt reconstruction accuracy but dramatically improves learning and inference efficiency. We evaluate the EPP-MVSNet on Tanks & Temples(TNT) [13], ETH3D [17] and DTU [1] dataset and show that the proposed network achieve promising performance on both reconstruction quality and efficiency aspects. Ablation study is further conducted for exhibiting advantageous effect brought by each key modules proposed.

To summarize, our major contributions are listed as follows:

- We introduce an epipolar-assembling module for assembling high-resolution information into cost volumes with limited size.
- We propose an entropy-based process that adjusts depth range for reducing redundancy and information loss.
- We apply a light-weighted 3D regularization network which dramatically increases learning and inference efficiency.
- We have conducted extensive experiments to show that EPP-MVSNet outperforms state-of-the-art methods in TNT and ETH3D datasets with respect to effectiveness and efficiency.

2. Related work

MVS has been exploited for decades with traditional methods such as COLMAP [16], ACMM [20] and Gipuma [8] which achieve great and robust results. However, facing the challenge of high-performance large-scale 3D reconstruction, traditional methods fail to leverage the accuracy and the computational cost. With deep learning accomplishing significant achievements in on 2D and 3D vision tasks [4, 6, 28], learning-based MVS methods also demonstrate promising performances.

Ji et al. bring up the first learning based network, SurfaceNet [11], for MVS which utilizes 3D CNN for regularizing disparity. Adopting the divide-and-conquer strategy, [11] is memory expensive, thus can only be used in limited sized scenes. Later, Yao et al. propose MVSNet [25] for large-scale 3D reconstruction and brought up a widely used pipeline compose of feature extraction on 2D images, construction a volume of matching cost between images features, cost regularization and depth regression. However, the memory and computation requirement of MVSNet [25] is determined by the spatial resolution of the image and the depth resolution of the scene and the regularization network of multiple 3D convolution layers.

Motivated by the demand for large-scale reconstruction, [9, 22, 23] propose to further enhance the efficiency of MVSNet which can be categorized into RNN-based method and CNN-based method. RNN-based MVS [22, 23] reduces memory requirement by replacing the regularization network with cubic 3D convolutions with convolutional GRU and LSTM for regularizing the cost volume sequentially along the depth dimension. RNN-based methods suffer from great time consumption in exchange for low memory cost. In contrast, CNN-based MVS [5, 9, 24, 27] preserve regularization with cubic 3D CNN, and adopt a coarse-to-fine structure for depth estimation. CasMVSNet [9] raises a multi-stage pipeline for predicting depth initially with a low-resolution cost volume at the coarse stage and refine predicted coarse depth at high resolution in a narrow depth range. To further increase the reconstruction resolution with limited cost, CVP-MVSNet [24] and UCSNet [5] propose to modify the construction of fine-stage cost volume. To maintain a predictable the computational cost, both methods adopt a fixed number of depths for the cost volume. CVP-MVSNet constructs the cost volume with a proposed optimal depth resolution of half pixel which results in a narrow range for depth prediction in fine stage. Despite achieving high depth resolution, the reconstruction quality of CVP-MVSNet is heavily affected by the narrow range restricted by depth resolution and hypothesis number. UCSNet solve this problem by predicting an appropriate range adaptive to the confidence of depth prediction in the previous stage and adjust the depth resolution accordingly. However, the effectiveness of CVP-MVSNe and UCSNet highly depends on the quality of coarse-stage depth prediction thus is not robust to real-life complicated scenes.

To this end, we propose an EPP-MVSNet for extending the trade-off of reconstruction accuracy and computational cost. Enlightened by [9], we also adopt a coarse-to-fine structure and propose further improvements for constructing compact and non-redundant cost volumes. Distinct from previous works, we optimize the construction of coarse cost volume by assembling high resolution features which enhance the prediction accuracy of coarse stage. Moreover, we estimate the confidence of depth prediction by probability entropy of cost volumes and adjust the hypothesis range accordingly. In addition to our effort on enhancing recon-

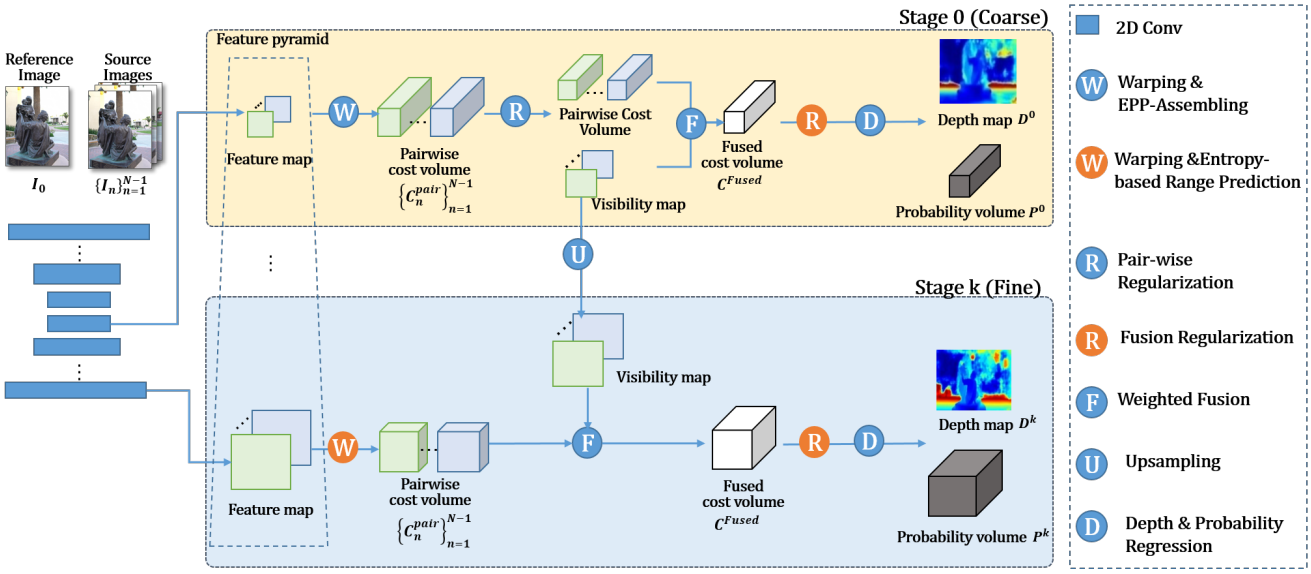


Figure 2. Structure of EPP-MVSNet. The proposed network leverages coarse and fine depth prediction and adopt different network accordingly. The coarse depth map D^0 is predicted by regressing an assembled cost volume constructed with proposed epipolar-assembling method and regularized with pair-wise and fused regularization network. In the fine stage, we adopt the entropy-based refining strategy for constructing cost volume within a non-redundant range. Afterwards, the fine cost volume is regularized with fused regularization network and regressed to generate depth map

struction quality, we also exploit the excessiveness of the cost regularization and build a light-weighted network.

3. Method

In this section, we start with an introduction of the overall structure of the EPP-MVSNet and further present the novel epipolar-assembling module and the entropy-based refining strategy for cost volume construction as well as the proposed light-weighted network for cost regularization. EPP-MVSNet adopts a multi-stage structure for predicting depth in a coarse-to-fine manner (See Figure 2). At each stage k, the depth map D^k and the corresponding probability volume P^k are inferred with four key procedures, feature extraction, cost volume construction, regularization and regression. To start with, given reference image I_0 and source images $\{I_i\}_{i=1}^{N}$, a pyramid feature extraction is applied for generating features maps at coarse or fine spatial resolution. Then, the cost volumes are constructed by firstly build feature volumes by homography warping feature maps at several hypothesized depths and further calculating the matching cost between reference and source feature volumes. Specifically, we utilize the epipolar-assembling kernel (Section 3.1.1) and entropy-based refining strategy (Section 3.1.2) respectively for the cost volume construction at coarse and fine stages. Thirdly, the cost volumes are regularized with a light-weighted network integrated with Pseudo-3D CNN which is specifically introduced in Section 3.2). Finally, the regularized cost volume is regressed to generate the depth map and a corresponding probability map.

3.1. Cost volume construction

The cost volume is constructed by calculating the correlation between reference and source features. We first utilize the differentiable homography [19, 25] for feature volume construction by warping all feature maps into the fronto-parallel plane of reference view at a set of hypothesized depth d:

$$p_n = K_n \cdot (r_{0,n} \cdot (K_0^{-1} \cdot p \cdot d) + t_{0,n}), \qquad (1)$$

where p_n represents the transformed pixel corresponds to pixel p on source image I_n at hypothesized depth d. K_n denotes the intrinsic matrix of source image I_n and $r_{0,n}$ and $t_{0,n}$ denote the relative rotation and transformation parameters between reference image I_0 and source image I_n. Given the feature map and hypothesized depths, a feature volume F_n of source image warped to the reference view can be calculated according to Equation 1. Then, we construct the cost volume by calculating matching cost between reference and source feature volumes using the proposed epipolar-assembling module and entropy-based refining accordingly at coarse and fine stages.

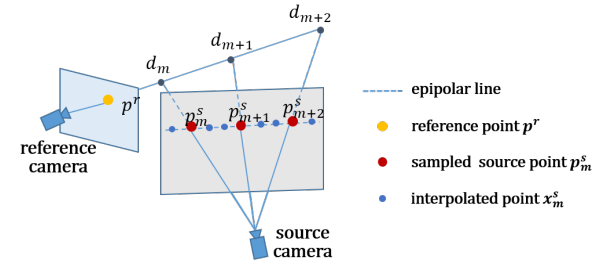
3.1.1 Epipolar-assembling module

For the coarse stage, we propose the epipolar-assembling module for constructing cost volume. According to [3, 22,

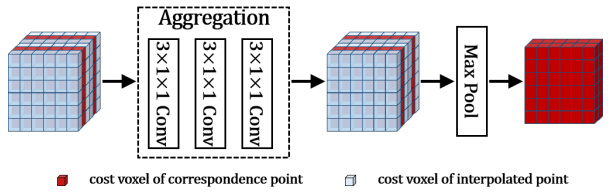
25], constructing a high resolution cost volume with narrow interval between hypothesized depths leads to full utilization of multi-view images and greatly improves reconstruction accuracy. As shown in Figure 3(a) which visualizes the point-correspondence of homography warping source and reference images, we observe that for each reference point p^r, the correspondence source points p^s_m are discretely sampled along the epipolar-line at different depth hypothesis d_m. With the hypothesis range fixed in the coarse stage, the interval between sampled source points p^s_m can be narrowed by increasing hypothesis number M which inevitably results in the growth of volume size and high cost on the memory and computation. To this end, we aim to break the constraint of network efficiency for utilizing high resolution cost volume by integrating features at adaptive interval to scattered sampled points along the epipolar-line:

$$cost^a(p^s_m) = \int_{p^s_m - \frac{\alpha}{2}}^{p^s_m + \frac{\alpha}{2}} \Omega(cost(x))\mathrm{d}x, \qquad (2)$$

where α denotes intervals between the sampled points p^s_m and the $\Omega(\cdot)$ represents the proposed epipolar-assembling kernel. Based on Equation 2, $cost(x)$ within the range of $(p^s_m - \frac{\alpha}{2}, p^s_m + \frac{\alpha}{2})$ are assembled to $cost^a(p^s_m)$. We discretize the Equation 2 for the implementation of epipolar-assembling module.



(a)



(b)

Figure 3. Epipolar-assembling module: Figure 3(a) visualizes the point correspondence between reference and source images at different depth hypothesises and the dense interpolation of points along the epipolar line. Figure 3(b) shows the assembling network for assembled cost volume construction

To start with, the positions of the sampled source points are acquired using Equation 1 at depth hypothesis $\{d_m\}^M_{m=1}$, we further interpolate even number of points along the epipolar line by a maximal interval of half pixel which proposed by [24] as the optimal interval and construct a high resolution cost volume by measuring the group correlation between reference point and the densely interpolated points. It is notable that, the hypothesized depth is generated using the inverse depth setting [22] which leads to a relatively uniform interval between the sampled points. Then, the cost volume is downsized by assembling cost volumes of interpolated points through a network shown in Figure 3(b). We design the assembling network with two steps, aggregation and pooling. Given the high resolution cost volume, each volume aggregates the neighboring features using three convolution layers of $3 \times 1 \times 1$ kernels for an appropriate receptive field. Further, the cost volume is downsized by a max-pooling operation along the depth direction with the window size adaptive to the interpolation rate.

Through the aggregation and pooling process, we manage to assemble the dense feature to the scattered sampled points and construct a compact and limited-sized cost volume. It is notable that the proposed epipolar-assembling kernel not only makes full use of the information from images but also adaptively assembles features at the optimal resolution in spite of the variation of the depth interval caused by the diversity of camera positions. Our experiments in Section 5 confirm that reconstruction using the proposed assembled cost volume achieves comparable results with reconstructing with high resolution cost volumes.

3.1.2 Entropy-based refining strategy

Adopt multi-stage structure, depth map D^{k+1} is predicted by refining D^k in a narrower range. Consequently, the depth hypothesises for fine cost volumes are determined. As shown in Figure 4, for each pixel, the center of the

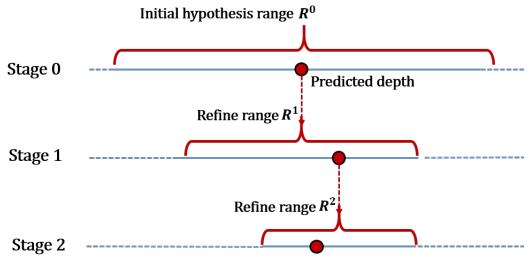


Figure 4. Variation of hypothesized depths at coarse-to-fine stages

hypothesized depths $\{d^{k+1}_m\}^M_{m=1}$ is the predicted depth in stage k and the hypothesis range is narrowed typically by a fixed factor often determined by the experiments [9, 27]. Narrowing range with a fixed factor may either cause the truth depth locating excluded from the refining range in the case of poor coarse depth prediction or introducing redundancy for refining an accurate depth within a wide range.

In contrast, we propose to narrow the hypothesis range with little redundancy based on the confidence of last-stage prediction by using the proposed entropy-based refining strategy. To further present the insight of the philosophy of the entropy-based refinement for depth prediction confidence, we refer to the original definition of entropy [18] that the entropy of variable is the average level of "information" and "surprise" inherent in the variable's possible outcomes. As in our case, given M possible outcomes, E^k estimates the amount of "surprise" in the depth prediction of stage k and $M^{\overline{E^k(p)}}$ is the sufficient number of states for describing the "surprise".

Given the probability volume P^k, the confidence for depth prediction at stage k is estimated by the entropy of the predicted probability for each hypothesis depth:

$$E^k(p) = -\sum_{m=0}^{M-1} P^k(p, d_m^k) \log_M P^k(p, d_m^k), \quad (3)$$

where $P^k(p, d_m^k)$ denotes the probability for the depth value of pixel p being hypothesis depth d_m^k and the number of hypothesis depth for stage k is represented as M. Greater entropy indicates less confidence for D^k which naturally requires a greater range of hypothesis depth. The hypothesis depth range for stage $k+1$ is determined by:

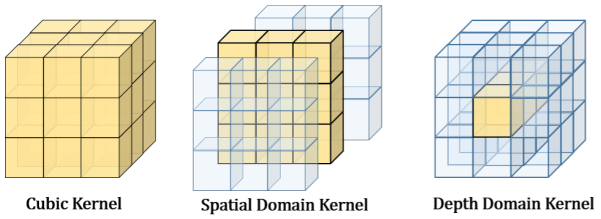
$$r^{k+1} = \left(\frac{M^{\lambda \cdot \overline{E^k(p)}}}{M}\right) \cdot r^k, \quad (4)$$

where r^k is the hypothesis depth range for stage k. Consequently, the cost volume can be constructed by calculating the group-correlation between reference and source feature volumes warped at the determined depth hypothesises. Because the confidence for depth map is approximated by simply averaging pixel-wise entropy, we introduce a hyper-parameter λ for tuning the narrowing factor of hypothesis depth range. Adopting the entropy-based refining strategy enables adaptive adjustment for the hypothesis range with little redundancy. With a fixed number of hypothesis depth, the fine cost volume with non-redundant range and relatively high resolution can be constructed.

3.2. Light-weighted regularization

In this section, the proposed light-weighted cost regularization network is introduced. Enlightened by [27], we adopt two 3D U-nets [15] for regularizing pair-wise and fused cost volumes and further optimize the network. In the coarse stage, given the cost volumes of each pair of reference and source feature volumes, F_0 and F_n, we exert a pair-wise regularization on the cost volume which is a two-block 3D U-net and a visibility map is jointly inferred. The fused cost volume is constructed by the linear combination of pair-wise cost volumes using visibility map as weights. Then, the fused cost volume is further regularized through a two-block 3D U-net network. Finally, the coarse depth is regressed from the cost volume with the soft-argmin operation. As for the fine stage, we directly fuse the pair-wise cost volumes using the up-sampled visibility map inferred at the coarse stage, and the combined cost volume is regularized by the fused regularization network.

Further more, considering the physical interpretation of the cost volume, we argue that the cost volume of neighboring pixels at different depths convolved by normal cubic CNN has little coherence which leads to redundant computation and high cost. Following [14], we replace normal 3D convolution operators with Pseudo-3D convolutions.



(a) (b)

Figure 5. Comparison between the normal 3D convolution in Figure 5(a) and the Pseudo-3D convolution in Figure 5(b)

Illustrated in Figure 5, the proposed Pseudo-3D adopts CNN separately on the spatial and the depth dimension. For the spatial convolution with kernel size of $1 \times 3 \times 3$, the cost volumes of adjacent pixels are convolved and on the depth domain the cost volumes of a pixel at different depth hypothesises are convolved by a convolution with kernel size of $3 \times 1 \times 1$. Evidently, the computational cost is greatly reduced and the reconstruction quality is also improved.

4. Experiments

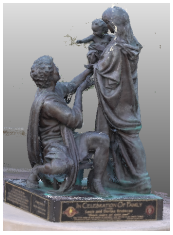
4.1. Implementation

4.1.1 Training

Our network is trained on BlendedMVS dataset [26] for Tanks & Temples [13] and ETH3D [17] benchmarking. We also evaluate our method on DTU [1] evaluation set and the training setting is shown in Section 5. BlendedMVS is a large-scale dataset consist of over 17k high-resolution images and 113 scenes covering various scenes including architectures and sculptures. During the training, we set the image resolution to 512×640 and the number of source images in a group $N = 3$, and output depth map size is 256×320. We adopt three stages structure consists of a coarse stage and two fine stages and train it by the summation L_1 loss of depth maps predicted at all stages and the uncertainty loss [27] of probability volume at the coarse stage. For each stage, the number of hypothesis depths is

Method	intermediate									advanced							DTU(mm)		
	mean	Fam.	Franc.	Horse	Light.	M60	Pan.	Play.	Train	**mean**	Audi.	Ballr.	Courtr.	Museum	Palace	Temple	Acc.	Comp.	Overall
COLMAP [16]	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04	27.24	16.02	25.23	34.70	41.51	18.05	27.94	0.400	0.664	0.532
ACMM [20]	57.27	57.27	**69.24**	51.45	46.97	55.07	57.64	60.08	54.58	34.02	23.41	32.91	**41.47**	48.13	26.17	36.69	-	-	-
CVP-MVSNet [24]	54.03	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54	-	-	-	-	-	-	-	**0.296**	0.406	0.351
CasMVSNet [9]	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	31.12	19.81	38.46	29.10	43.87	27.36	28.11	0.325	0.385	0.355
UCSNet [5]	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	-	-	-	-	-	-	-	0.338	0.349	**0.344**
Vis-MVSNet [27]	60.03	77.40	60.23	47.07	**63.44**	**62.21**	57.28	60.54	52.07	-	-	-	-	-	-	-	0.369	0.361	0.365
PatchmatchNet [19]	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	32.31	**23.69**	37.73	30.04	41.80	28.31	32.29	0.427	**0.277**	0.352
Ours	**61.68**	**77.86**	60.54	**52.96**	62.33	61.69	**60.34**	**62.44**	**55.30**	**35.72**	21.28	**39.74**	35.34	**49.21**	**30.00**	**38.75**	0.413	0.296	0.355

Table 1. F-score (higher is better) results on the Tanks & Temples benchmark [13] and quantitative result (lower is better) on the evaluation set of DTU [1]. The overall best results are marked as bold numbers.



(a) Family

(b) Francis

(c) Train

(d) Panther

(e) Playground

Figure 6. Point clouds on Tanks & Temples intermediate dataset [13].

respectively $M^1 = 32$, $M^2 = 16$, $M^3 = 8$. The learning rate is set to be 0.001, and the network is trained for 10 epochs with a batch size of 4 using Adam [12] as optimizer. The learning rate is reduced by half at epoch 6, 8 and 9, respectively.

4.1.2 Evaluation

We evaluate the EPP-MVSNet on Tanks & Temples, ETH3D and DTU dataset without fine-tuning process and compare it with other state-of-the-art learning-based methods. For evaluation, we set the hypothesis depth number to be $M^1, M^2, M^3 = 32, 16, 8$ and the interval threshold for coarse-stage pixel interpolation is set to be 0.5. We adopt the dynamic consistency checking approach proposed in [23] for generating point cloud from the depth maps.

Tanks & Temples Dataset. Tanks & Temples (TNT) is a benchmarks for realistic scenes includes both outdoor and indoor scenes. We evaluate EPP-MVSNet on the intermediate and advanced TNT dataset. For evaluation, we set the input image size to 1920×1056, and adopt 7 source images for each inference process. As shown in Table 1, our method outperforms others on the overall quality and achieve the highest mean F-score in the intermediate benchmark (until March 17, 2021). For example, EPP-MVSNet shows significant improvement on each scenes compared to the coarse-to-fine method CasMVSNet [9]. While comparing to CVP-MVSNet [24] and UCSNet [5] which also propose to increase the reconstruction resolution, EPP-MVSNet shows respectively 7.65 and 6.85 higher results on mean F-score. Furthermore, compared to SOTA method Vis-MVSNet [27], our method performs better on most scenes. For the most challenging advanced dataset, our method EPP-MVSNet still performs best among all the methods, which contains some traditional MVS methods like [20]. The generated point cloud is shown in Figure 6, it is obvious that the proposed methods manage to generate dense point cloud with fine details well reserved.

ETH3D Dataset. ETH3D provides diverse types of scenes ranging from complicated natural scenes to man-made indoor and outdoor environments with relatively large variation of view-points thus reconstruction on ETH3D dataset requires more robustness and generalization ability of the network. We set the input image size to 3072×2048 and the number of source images N to 7. For most learning-based methods shows poor performance on the ETH3D

benchmark, we further present results of traditional methods for comparison. As is shown in Table 2, our method outperforms learning-based methods PVSNet [21] and PatchmatchNet [19], and shows competitive results with traditional MVS methods [20].

DTU Dataset. DTU dataset contains more than 100 scenes which collected by a robot arm with a structured light scanner. We use a resolution of 1600×1184 and set the number of source images to 4 on evaluation set. As shown in Table 1, Our method shows comparable results with other SOTA methods.

Method	Training	Test
Gipuma [8]	36.48	45.18
PMVS [7]	46.06	44.16
COLMAP [16]	67.66	73.01
ACMM [20]	**78.86**	80.78
PVSNet [21]	67.48	72.08
PatchmatchNet [19]	64.21	73.12
Ours	74.00	**83.40**

Table 2. F_1 score (in %) comparisons of point clouds on ETH3D high-res benchmark at evaluation threshold 2cm.

Memory and Run-time Comparison. The computational cost of EPP-MVSNet is compared with the mentioned learning-based methods by competing memory consumption and run-time. For fair comparison, we use a fixed input size of 1920×1056 and the same source images number 4 for all competing methods. As shown in Table 3, our method EPP-MVSNet reduces 30.3% and 36.1% run-time compared to CasMVSNet [9] and Vis-MVSNet [27] respectively.

Method	Time(ms)	Mem.(GB)
CasMVSNet [9]	792.2	9.5
Vis-MVSNet [27]	864.2	4.5
PatchmatchNet [19]	317.7	3.2
Ours	552.2	8.2

Table 3. Comparison of the running time(*ms* per view) and memory consumption between our methods and other SOTA learning-based multi-view stereo methods [9, 19, 27] on *Tanks & Temples* [13].

5. Ablation study

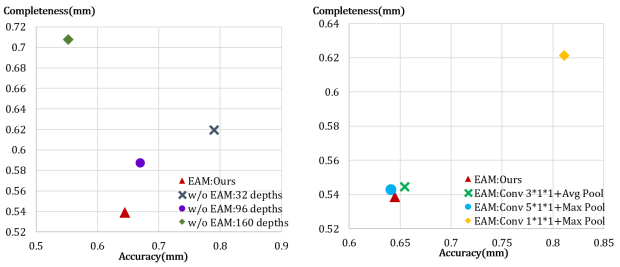
We conduct extensive ablation study for validating enhancements brought up by the proposed modules. Here we use DTU training set [1] to train our method and all test on DTU evaluation set. Training setting is the same as Section 4.1.1 except with a batch size of 8 and a initial learning rate of 0.0015. While testing, we follow the same setting which is shown in Section 4.1.2.

Method	Depth Num.	Acc.(mm)	Comp.(mm)	Overall(mm)
w/o EAM	32	0.7903	0.6195	0.7049
	96	0.6692	0.5871	0.6282
	160	**0.5521**	0.7074	0.6298
Ours	32	0.6451	**0.5389**	**0.5920**

Table 4. Ablation study of reconstructing w/o the epipolar-assembling module(EAM) on the 1^{st} stage.

EAM kernel	Acc.(mm)	Comp.(mm)	Overall(mm)
Conv $3\times1\times1$ + Max pool (ours)	0.6451	**0.5389**	**0.5920**
Conv $3\times1\times1$ + Average pool	0.6541	0.5447	0.5994
Conv $5\times1\times1$ + Max pool	**0.6413**	0.5428	0.5921
Conv $1\times1\times1$ + Max pool	0.8112	0.6213	0.7163

Table 5. Ablation study for different kernels adopted for the epipolar-assembling module(EAM).



(a) Ablation study of EAM (b) Ablation study of EAM kernel

Figure 7. Comparison of reconstruction accuracy(mm) and completeness(mm) for reconstructing w/o the proposed epipolar-assembling module(EAM) and with various assembling kernels on DTU evaluation set [1].

Epipolar-assembling module. To quantitatively measure the effectiveness of aggregation, we only use depth prediction results of 1^{st} stage to reconstruct point clouds. As stated in Section 3.1.1, the epipolar-assembling module aims to construct a compact cost volume with high resolution features and limited size. We compare our epipolar-assembling method with three depth number settings on cost volume without assembling network. Comparing results constructed with same resolution cost volume, we observe that utilizing the proposed epipolar-assembling network significantly improves the reconstruction quality.(0.5920 vs. 0.7049, overall quality) Even compared with $3\times$ and $5\times$ resolution case, our method still performs better on completeness and overall quality. It is shown in Figure 7(a) and Table 4 that utilization of aggregated cost volume benefits the reconstruction quality by a great amount comparing to constructing cost volume with much higher resolution. It is worthy to clarify that although adopting adaptive window size for epipolar-assembling, the average window size for the whole DTU evaluation set is 1.408. Thus, it can be concluded that not only does the proposed method effectively aggregate high-resolution features but

also manage to adaptively aggregates at an optimal resolution.

Kernel for epipolar-assembling. Firstly, we demonstrate the enhancement brought by the proposed epipolar-assembling module by evaluating the point clouds quality reconstructed from 1^{st} stage outputs. By replacing the epipolar-assembling kernel, we experiment on different network architectures for aggregating high-resolution feature. Apart from the adopted kernel which composes of 3 layers of convolutions with kernel size of $3\times1\times1$ and a max pooling operation, we alter the size of convolution kernel and replace the pooling operation. In Table 5 and Figure 7(b), the accuracy and completeness of point cloud reconstructed using different assemble kernel are illustrated. Comparing results of using max and average pooling, kernel with max pooling outperforms kernel with average pooling on both accuracy and completeness. Using max pooling operation, we compare the results with different convolution kernel size. We observe that adopting kernel size of $3\times1\times1$ and $5\times1\times1$ leads to comparable results which are both significantly better than results inferred with convolution kernel size of $1\times1\times1$.

Method	Acc.(mm)	Comp.(mm)	Overall(mm)
1^{st} stage	0.6451	0.5389	0.5920
3^{rd} stage w/o ER	0.4255	**0.2935**	0.3595
3^{rd} stage	**0.4137**	0.2968	**0.3553**

Table 6. Ablation study of entropy based refining strategy (ER) on DTU [1].

Entropy-based refining strategy. We further inspect the entropy-based refine strategy by comparing 3^{rd} stage's reconstructed results based on 1^{st} stage's depth prediction using epipolar-assembling module. Our method achieves 0.4137 on accuracy and 0.3553 on overall quality, which enhanced 0.0188 and 0.0038 compared to baseline (3^{rd} stage w/o ER). It is noticed that improvement on 3^{rd} stage is particularly challenging, which means the entropy-based refine strategy practically reduces redundancy depth range for next stage to generate better depth prediction.

Volume size	Run-time(ms)	
	3D Block	Pseudo-3D Block
$32\times296\times400$	23.6	16.4
$64\times296\times400$	46.6	31.7
$32\times148\times200$	6.0	4.4

Table 7. Run-time ablation study of Pseudo-3D block on DTU [1]. We test run-time of normal 3D block(contains two $3\times3\times3$ convolution layer which followed with BN and ReLU) and our Pseudo-3D block(contains one $1\times3\times3$ convolution layer and one $3\times1\times1$ convolution layer, followed with BN and ReLU at last) on cost volume of different shape($D \times H \times W$).

Method	Acc.(mm)	Comp.(mm)	Overall(mm)	Time(ms)	Mem.(GB)
3D CNN	0.4160	0.2989	0.3575	624.9	5.2
Pseudo-3D CNN	**0.4137**	**0.2968**	**0.3553**	**340.6**	**3.1**

Table 8. Comparison of reconstruction quality and computational cost between integrating Pseudo-3D CNN and normal 3D CNN for cost regularization in the proposed network under the case of a fixed window size of 3 for the epipolar-assembling kernel.

Pseudo-3D convolution. Experiments are further conducted for showing the beneficial effect of regularization using Pseudo-3D convolution. We compare the reconstruction quality and computational cost of adopting Pseudo-3D CNN and 3D CNN in the regularization network. According to Table 8, adopting Pseudo-3D CNN for regularization results in comparable accuracy and completeness with normal 3D CNN. To present the advantageous effect brought by integrating Pseudo-3D convolution, we compare the performance of Pseudo-3D convolution with normal cubic 3D convolution. As shown in Table 7, the 3D block reduce respectively 30.5%, 32.0%, 26.7% for operating on volume sizes of $32\times296\times400$, $64\times296\times400$, $32\times148\times200$ run-time in each cases. To fully analyse the computational enhancement relative to the whole network, we also exhibit the time and memory consumption comparison between 3D CNN and Pseudo-3D CNN. With a fixed window size for epipolar-assembling kernel of 3, the run-time reduced by 45.5% and the memory requirement is reduced by 40.4% while using Pseudo-3D instead of the normal 3D CNN.

6. Conclusion

We present EPP-MVSNet, a novel learning-based method with the proposed light-weighted coarse-to-fine network for effective and efficient high resolution depth prediction. Initially, we adopt the adaptive epipolar-assembling kernel for the construction of a compact cost volume with aggregated high resolution feature resulting in the high accuracy of coarse depth prediction. Then, we refine the depth prediction in narrow range predicted by the proposed entropy-based range prediction strategy. We further enhance the efficiency of the network by optimizing the cost regularization network and integrating Pseudo-3D operation. The proposed EPP-MVSNet achieves the highest F-Score on the Tanks & Temples benchmark and achieves relatively low memory and time consumption comparing to the state-of-the-art learning-based methods.

Acknowledgements We thank MindSpore[1] for the great support of this work, which is an open AI framework with friendly design, efficient running experience and flexible deployment.

[1]https://www.mindspore.cn/

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016.

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[3] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1538–1547, 2019.

[4] Xinlei Chen and Abhinav Gupta. An implementation of faster rcnn with study for region sampling. *arXiv preprint arXiv:1702.02138*, 2017.

[5] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020.

[6] Ziqing Feng and Qijun Zhao. Robust face recognition with deeply normalized depth images. In *Chinese Conference on Biometric Recognition*, pages 418–427. Springer, 2018.

[7] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

[8] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015.

[9] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020.

[10] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):504–511, 2012.

[11] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2307–2315, 2017.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[14] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[16] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[17] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[19] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. *arXiv preprint arXiv:2012.01411*, 2020.

[20] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5483–5492, 2019.

[21] Qingshan Xu and Wenbing Tao. Pvsnet: Pixelwise visibility-aware multi-view stereo network. *arXiv preprint arXiv:2007.07714*, 2020.

[22] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European Conference on Computer Vision*, pages 674–689. Springer, 2020.

[23] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European Conference on Computer Vision*, pages 674–689. Springer, 2020.

[24] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020.

[25] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[26] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020.

[27] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *arXiv preprint arXiv:2008.07928*, 2020.

[28] Zisha Zhong, Yusung Kim, Leixin Zhou, Kristin Plichta, Bryan Allen, John Buatti, and Xiaodong Wu. 3d fully convolutional networks for co-segmentation of tumors on pet-ct images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 228–231. IEEE, 2018.