Efficient Edge-Preserving Multi-View Stereo Network for Depth Estimation

Wanjuan Su, Wenbing Tao*

National Key Laboratory of Science and Technology on Multispectral Information Processing
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
{suwanjuan, wenbingtao}@hust.edu.cn

Abstract

Over the years, learning-based multi-view stereo methods have achieved great success based on their coarse-to-fine depth estimation frameworks. However, 3D CNN-based cost volume regularization inevitably leads to over-smoothing problems at object boundaries due to its smooth properties. Moreover, discrete and sparse depth hypothesis sampling exacerbates the difficulty in recovering the depth of thin structures and object boundaries. To this end, we present an Efficient edge-Preserving multi-view stereo Network (EPNet) for practical depth estimation. To keep delicate estimation at details, a Hierarchical Edge-Preserving Residual learning (HEPR) module is proposed to progressively rectify the upsampling errors and help refine multi-scale depth estimation. After that, a Cross-view Photometric Consistency (CPC) is proposed to enhance the gradient flow for detailed structures, which further boosts the estimation accuracy. Last, we design a lightweight cascade framework and inject the above two strategies into it to achieve better efficiency and performance trade-offs. Extensive experiments show that our method achieves state-of-the-art performance with fast inference speed and low memory usage. Notably, our method tops the first place on challenging Tanks and Temples advanced dataset and ETH3D high-res benchmark among all published learning-based methods. Code will be available at https://github.com/susuwj/EPNet.

Introduction

Multi-View Stereo (MVS) aims to reconstruct the 3D scene geometry from multiple calibrated images which is a fundamental topic in computer vision. Benefiting from the coarse-to-fine architecture (Cheng et al. 2020; Yang et al. 2020; Gu et al. 2020), learning-based MVS methods have achieved significant progress concerning both efficiency and quality of the reconstruction in recent years (Wang et al. 2021; Peng et al. 2022; Su, Xu, and Tao 2022). However, existing learning-based MVS methods still struggle to recover the depth at thin structures and object boundaries and are hard to balance efficiency and generalization ability.

The core idea of learning-based methods is building cost volumes by the differentiable homography (Yao et al. 2018; Gu et al. 2020) with sampled depth hypotheses. And 3D CNNs are applied on cost volumes for regressing the depth map of the reference view. The 3D CNN-based cost volume regularization usually leads to over-smoothing problems at object boundaries (Tosi et al. 2021) caused by its smooth properties. Moreover, as the sampled depth hypotheses are discrete and the number of hypotheses is limited, it is hard to capture precise depths of thin structures and object boundaries, resulting in blur artifacts in the estimated depth maps. This problem is further exacerbated in the widely used coarse-to-fine architectures which first estimate initial depth maps with coarse depth hypotheses at low resolution and gradually refine the initial depth maps with finer cost volumes at higher resolution. As initial depth hypotheses are very sparse and the depth searching range in the finer stage is around the depth map estimated in the previous stage, it is hard to recover depth at object boundaries at the first stage, and errors in the initial depth map will be propagated to the finer stage. Additionally, in the coarse-to-fine architecture, simple bilinear upsampling is generally used in existing methods to upsample the low-resolution depth map to the higher one, which produces cross-edge interpolation.

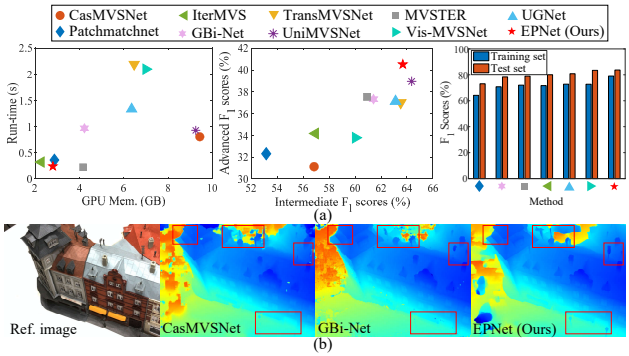

Figure 1: (a) left: comparison results of GPU memory and run-time consumption (image size 1920×1024, 5 views), middle and right: comparison with state-of-the-art learning-based methods on Tanks & Temples and ETH3D datasets (higher is better). (b) Qualitative comparison with CasMVSNet (Gu et al. 2020) and GBi-Net (Mi, Di, and Xu 2022) on Scan9 of DTU dataset. Our method can recover edge-preserving depth maps with rich details.

*Corresponding author

This further aggravates the problem of blur artifacts in the estimated depth maps. Blur artifacts not only reduce the accuracy of the depth map but also reduce the completeness of the reconstructed point cloud by destroying the geometric consistency between views.

To address the problems above, we resort to utilizing the context information in the image, which is a crucial clue to reflect the structure of a scene but is ignored by most MVS methods. We thus propose a Hierarchical Edge-Preserving Residual learning (HEPR) module which incorporates the context information in the reference image to perform edge-aware depth refinement by learning residual maps with high-frequency details. Instead of directly upsampling the depth map to a higher resolution with the bilinear upsampling in the coarse-to-fine architecture, we hierarchically use residual learning networks in the HEPR under the guidance of image context information to estimate residual maps that are used to blend high-frequency details into depth maps predicted by the backbone network, so as to achieve edge-preserving upsampling while refining the erroneous regions of the depth maps predicted by the backbone network. In this way, blur-free depth maps can be obtained which can boost the reconstruction quality.

Furthermore, we noticed that some extreme outliers resulting from hard samples in depth domain prevent the further optimization for more accurate depth estimation of fine-grained regions. As shown in Figure 2 (a), errors between Ground Truth (GT) and estimated depth maps in the depth domain are dominated by extreme errors of texture-less areas (highlighted by the black box). Such dominant errors are harmful to the optimization of interested regions but also are useless for texture-less areas as in theory they are nearly unpredictable (Ding et al. 2022b). To this end, we propose an auxiliary Cross-view Photometric Consistency (CPC) loss to optimize depth errors in image domain instead of depth domain only, so as to enhance the optimization direction towards the interested regions. Specifically, the proposed CPC loss measures the difference between the images synthesized by GT and estimated depth maps as shown in Figure 2 (b). This inverse warping from depth domain to image domain results in that the minor errors of interested areas in depth domain can be magnified due to the non-smooth change of colors in the image, and the excessive outliers from texture-less areas are suppressed, thus generating a more uniform error map. This can be observed by comparing error maps between Figure 2 (a) and (b). As a result, we can enhance the gradient flow of fine regions to boost performance.

In addition, most of the current learning-based methods generally construct cost volumes with high-resolution, which induces high consumption of memory and run-time (Peng et al. 2022; Ding et al. 2022a) as these grow cubically with the increase of cost volumes' resolution (Yao et al. 2019). This inevitably hinders the practical applications which generally require the algorithms to be resource friendly and efficient. Although some efficient methods are proposed recently (Wang et al. 2021, 2022a,b), they show unpleasant performance compared with non-efficient methods (Peng et al. 2022; Ding et al. 2022a). To relief this high consumption of resource, we delicately design a lightweight

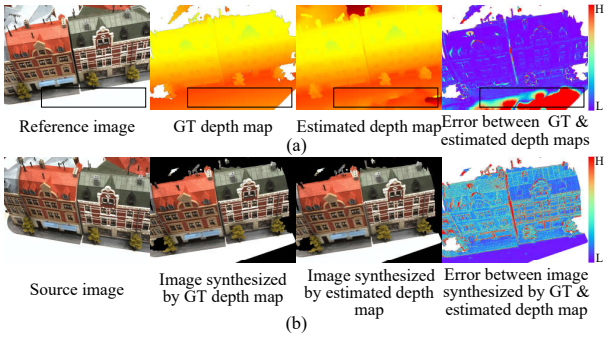

Figure 2: Comparison of error maps in depth domain and image domain, where error maps are obtained by using absolute error between the GT and estimated depth map or images synthesized by the GT and estimated depth map. Note that there are pixels without ground truth depth in GT and invalid pixels in the synthesized image, we masked them with zeros in error maps and synthesized images which are presented in white in GT and error maps, and black in synthesized images. The L and H denote the error is low and high.

cascade framework by stacking two stages at the same resolution, which contributes to maximizing depth hypothesis sampling accuracy as much as possible with a low overhead of memory and run-time.

By embedding the proposed HEPR module and CPC loss into our lightweight cascade framework, an Efficient edge-Preserving multi-view stereo Network (EPNet) is presented in this paper which can ensure high performance while maintaining high efficiency. Extensive experiments on various MVS benchmarks show that our method achieves state-of-the-art performance. Particularly, as shown in Figure 1 (a), we achieve competitive results compared with Trans-MVSNet (Ding et al. 2022a) and UniMVSNet (Peng et al. 2022) with the 56.66% and 69.68% reduction in GPU memory consumption, 88.90% and 74.00% reduction in run-time consumption, and achieve comparable GPU memory and run-time consumption compared to efficient methods Iter-MVS (Wang et al. 2022a) and MVSTER (Wang et al. 2022b) with much better generalization ability on Tanks and Temples and ETH3D high-res benchmarks.

To summarize, our contributions are as follows:

- We present a novel efficient edge-preserving multi-view network that exploits context information in the image for high-quality edge-aware depth estimation with friendly memory and run-time consumption;
- We propose a hierarchical edge-preserving residual learning module to perform the depth refinement which supports blur-free depth upsampling;
- We introduce a cross-view photometric consistency loss to effectively enhance the gradient flow of detailed regions during training.

Related Works

Learning-based MVS Learning-based MVS methods have achieved encouraging performance on various MVS

benchmarks, where most of these methods mainly follow the pipeline of MVSNet (Yao et al. 2018). To reduce the memory consumption caused by the use of 3D CNNs, the coarse-to-fine strategy (Cheng et al. 2020; Yang et al. 2020; Gu et al. 2020) is widely used in recent methods. This kind of method generally first estimates the initial depth map with low-resolution and coarse depth hypotheses, then upsamples the initial depth map to a higher one with the bilinear upsampling and generates finer depth hypotheses based on the previous estimation to refine the initial depth map. Recently, several variants based on multi-stage methods are proposed, e.g., introducing pixel-wise visibility information of source views into cost volumes (Zhang et al. 2020; Xu et al. 2022), applying epipolar-assembling-based kernel and entropy-based refining strategy to adaptively aggregate matching costs along epipolar lines (Ma et al. 2021), embedding the transformer into MVS (Ding et al. 2022a), constructing the cost volume by non-parametric depth distribution modeling (Yang, Alvarez, and Liu 2022) and so on. Although these methods have solved some problems in the vanilla multi-stage methods, they still have difficulty in solving the problem of blur artifacts.

Some other methods are dedicated to improving efficiency. PatchmatchNet (Wang et al. 2021) introduces the idea of Patchmatch which can reduces the overhead of memory and run-time. GBiNet (Mi, Di, and Xu 2022) formulates MVS as a binary search problem that decreases memory consumption by reducing the cost volume size. IterMVS (Wang et al. 2022a) and Effi-MVS (Wang, Li, and Dai 2022) present iterative GRU-based optimizers which are memory and run-time consumption friendly. MVSTER (Wang et al. 2022b) can estimate depth with fewer depth hypotheses by using the epipolar Transformer and optimal transport. However, these methods show limited generalization ability.

Depth Refinement and Upsampling in MVS In previous methods, the refinement/upsampling modules generally are involved as a final step of MVS which is used to refine/upsample the depth maps obtained from the backbone network, so as to get depth maps with better quality or higher resolution. MVSNet (Yao et al. 2018) uses a refinement module to refine the initial depth maps. To alleviate the stair effect, R-MVSNet (Yao et al. 2019) proposes a variational depth map refinement module. PatchmatchNet (Wang et al. 2021) designs a refinement module based on (Hui, Loy, and Tang 2016). IterMVS (Wang et al. 2022a) introduces a spatial upsampling module (Teed and Deng 2020) to upsample the depth to a higher resolution. Distinct from these methods, we propose a HEPR module to perform the depth refinement on the intermediate pyramid stages, so as to enforce the entire network achieve edge-aware depth estimation.

Loss Function in MVS Most of the learning-based MVS methods are supervised by L1 loss only (Yao et al. 2018; Gu et al. 2020; Wang et al. 2021). The recurrent-based methods (Yao et al. 2019; Yan et al. 2020; Wei et al. 2021) and GBiNet (Mi, Di, and Xu 2022) use the cross entropy loss. TransMVSNet (Ding et al. 2022a) introduces the focal loss (Lin et al. 2017) to handle ambiguous predictions. IterMVS (Wang et al. 2022a) proposes a hybrid training strategy with the L1 loss and cross entropy loss. UniMVSNet (Peng et al. 2022) unifies the advantages of regression and classification by the unified focal loss. However, these loss functions all act in the depth domain, which may easily affected by some extreme outliers during training. Note that the proposed CPC loss is different from the photometric consistency loss in self-supervised/unsupervised methods (Xu et al. 2021), which is vulnerable to the adverse effects of illumination changes and occlusion from different views. As CPC loss aims to transfer the difference between the GT and predicted depth from the depth domain to the image domain, it implicitly avoids the adverse effects of these factors.

Method

Given a reference image X_0 and its N-1 source images $\{X_i\}_{i=1}^{N-1}$ with their intrinsic and extrinsic camera parameters, MVS aims to recover the depth of the reference view. Figure 3 illustrates the overall pipeline of the EPNet, which is mainly composed of two modules: the Multi-Scale Depth Estimation (MSDE) module and the HEPR module. The MSDE adopts the designed lightweight cascade structure with the coarse-to-fine strategy (Gu et al. 2020) to estimate depth maps in an efficient manner. During training, the proposed CPC loss is used with L1 loss to supervise the MSDE for enhancing the gradient flow of detail structures. And the HEPR is embedded to the MSDE to gradually achieve edge-preserving depth refinement.

Hierarchical Edge-Preserving Residual Learning

The drawback of estimating depth solely based on geometric information is that some areas of the estimated depth map are not aligned with the context of the reference image, namely, there are blur artifacts in the depth maps. Moreover, the coarse-to-fine architecture depends on the estimation of coarse stages, which has the problems of error accumulation and aggravating blur artifacts. We approach these problems by performing edge-preserving refinement on the intermediate output of the coarse-to-fine architecture with the proposed HEPR module. Specifically, we task this module with learning a depth residual progressively to aid the coarse estimation to recover fine details at higher resolution with the guidance of the reference image.

As shown in Figure 3 (b), the HEPR module first adopts a context encoder to extract multi-scale context features of the reference image which is used to guide the depth residual learning network to learn fine details. Then, the depth residual learning networks are applied to hierarchically refine the intermediate output of the MSDE, so that the sophisticated structure can be progressively recovered. The input of the depth residual learning network is features of the depth maps extracted by the depth feature extraction network and context features extracted by the context encoder. The depth feature extraction network has a three-layer shallow CNN followed by a deconvolution layer with stride 2 to extract features from the normalized depth maps. The depth map is normalized by

$$\hat{D}_n = (\hat{D}_o - \text{mean}(\hat{D}_o))/\text{std}(\hat{D}_o), \tag{1}$$

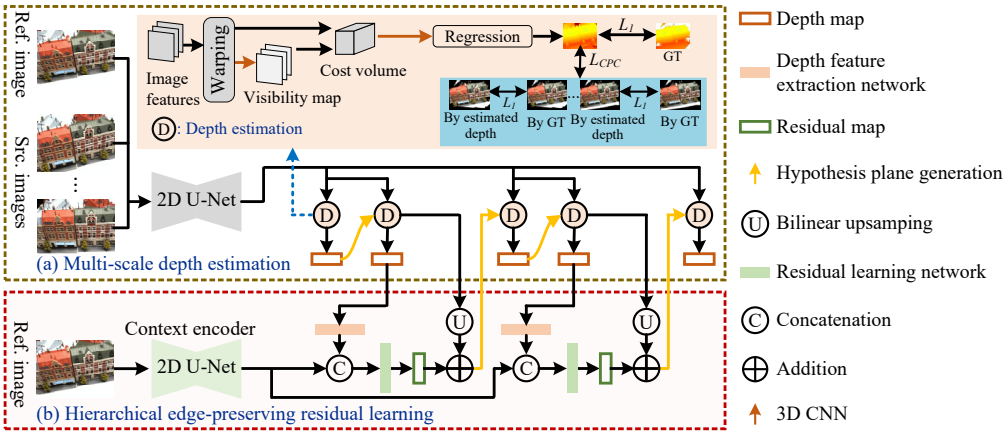

Figure 3: The overview of EPNet. (a) illustrates the pipeline of multi-scale depth estimation which adopts the coarse-to-fine strategy to estimate depth maps. The orange box in this module gives the specific process of stage-wise depth estimation where the proposed cross-view photometric consistency loss is applied during training. (b) shows the details of hierarchical edge-preserving residual learning module.

where $\hat{D}_n$ denotes the normalized depth, $\hat{D}_o$ denotes the depth estimated by MSDE, mean$(\cdot)$ and std$(\cdot)$ denote calculating the mean and standard deviation of the depth map.

The concatenated depth features and context features are processed by the depth residual learning network that mainly consists of an encoder and a decoder with skip connections to couple context features and depth map features. The encoder and decoder of the depth residual learning network consist of several residual blocks for learning a better coupling. Finally, the depth residual is output and added to the upsampled and normalized depth maps, generating an edge-preserving depth map with higher resolution, this process can be formulated by:

$$\hat{D}_{rn} = \text{Up}(\hat{D}_n) + \Delta\hat{D}, \tag{2}$$

where $\hat{D}_{rn}$ denotes the refined normalized depth map, Up$(\cdot)$ denotes using bilinear upsampling to upsample the input to twice its original size, and $\Delta\hat{D}$ denotes the learned depth residual. The $\hat{D}_{rn}$ is unnormalized by the original mean and standard deviation, and the edge-aware depth map $\hat{D}_r$ is:

$$\hat{D}_r = \text{std}(\hat{D}_o) \cdot \hat{D}_{rn} + \text{mean}(\hat{D}_o). \tag{3}$$

Cross-view Photometric Consistency Loss

The core idea of cross-view photometric consistency is to magnify gradient flows of detailed regions by measure the difference between synthesized images at the reference view by GT and estimated depth based on the images of source views. For a pixel p_j in the reference image X_0 with the depth value d, its corresponding pixel $\hat{p}_j$ in the source view is computed as:

$$\hat{p}_j = K_i(R_{0,i}(K_0^{-1}p_j d) + t_{0,i}), \tag{4}$$

where K_0 and K_i denote the intrinsic camera parameters of the reference view and i-th source view, $R_{0,i}$ and $t_{0,i}$ denote the relative rotation and translation between the reference view and i-th source view.

With the transformation above, the image $I_{0,i}$ synthesized by the i-th source image based on the depth maps D at the reference view can be obtained by differentiable bilinear sampling:

$$I_{0,i}(p_j) = I_i(\hat{p}_j). \tag{5}$$

A binary mask M_i is also generated during this process which indicates the invalid pixels in the synthesized image $I_{0,i}$, i.e., pixels projected to the outside regions of the image.

In this case, the cross-view photometric consistency is

$$L_{CPC} = \sum_{i=1}^{N} \sum_{p \in \mathbb{M}_i} \frac{|I_{0,i}^{gt}(p) - \hat{I}_{0,i}(p)|}{|\mathbb{M}_i|}, \tag{6}$$

where $I_{0,i}^{gt}$ and $\hat{I}_{0,i}$ denote images synthesized by the i-th source image based on the GT and estimated depth maps, respectively, N denotes the number of views, $\mathbb{M}_i$ denotes the valid pixels in both synthesized images and the GT depth map which is produced by $M_i \cdot \Omega$, Ω indicates the valid pixels in the GT depth map.

We use CPC loss and L1 loss in depth domain to constrain the MSDE module of EPNet:

$$\mathcal{L}_B = \sum_{s=0}^{4} \lambda^s \Big(\sum_{x \in \Omega} \frac{|D_{gt}^s(p) - \hat{D}_o^s(p)|}{|\Omega|} + \mu \mathcal{L}_{CPC}^s \Big) \tag{7}$$

where D_{gt}^s and $\hat{D}_o^s$ denote the GT and estimated depth maps at stage s, respectively, λ^s and μ are weight coefficients. The HEPR part is constrained by the L1 loss in depth domain alone, so the total loss of EPNet is formulated as:

$$\mathcal{L} = \mathcal{L}_B + \mathcal{L}_R = \mathcal{L}_B + \sum_{s=1,3} \eta^s \sum_{x \in \Omega} \frac{|D_{gt}^{s+1}(p) - \hat{D}_r^s(p)|}{|\Omega|} \tag{8}$$

where $\hat{D}_r^s$ denotes the depth output by HEPR at stage s, η^s is a weight coefficient for stage s.

Edge-Preserving Multi-View Stereo Network

As shown in Figure 3, EPNet combines the MSDE with HEPR. The MSDE applies the coarse-to-fine strategy to progressively estimate high-resolution depth maps, which first extracts multi-scale image features $\left\{F_i^k\right\}_{i=0}^{N-1}$ $(k = 0, 1, 2)$ with resolution $1/2^{3-k}H \times 1/2^{3-k}W$ for all images with resolution $W \times H$ by using a 2D U-Net with shared weights. The encoder and decoder of the 2D U-Net (Ronneberger, Fischer, and Brox 2015) are formed by multiple residual blocks. Then, the stage-wise depth estimation is performed as (Gu et al. 2020) does. As the stage-wise depth estimation is the same for all stages, we omit subindices to ease notation. For each stage, all feature maps of source views are warped into a set of fronto-parallel planes of the reference view at the sampled depth hypotheses to form the $N-1$ warped feature volumes $\left\{\acute{F}_i\right\}_{i=1}^{N-1}$ by the differentiable homography (Yao et al. 2018). And the group-wise correlation similarity (Xu and Tao 2020) with 8 groups is used to construct the two-view cost volume C_i by using the reference feature F_0 and i-th warped source feature $\acute{F}_i$. A two-layer 3D CNN followed by a Sigmoid activation and a max-pooling layer is imposed on the C_i (Xu et al. 2022) at the first stage to obtain the visibility map V_i of each source view, which is used to aggregate the two-view cost volumes into a unified cost volume C in a weighed fusion manner for all stages.

Unlike previous methods which directly use bilinear sampling to upsample the estimated depth maps to the next scale, we adopt the proposed HEPR module to further refine and upsample the estimated depth $\hat{D}_o$. In this way, the edge-preserving depth $\hat{D}_r$ can be obtained that is used as the basis for determining the depth sampling range in the next stage. We adopt the depth hypothesis sampling strategy of CasMVSNet (Gu et al. 2020) which uniformly samples the depth hypotheses with a decreasing depth sampling range and a decreasing depth sampling number. Note that different from the most of learning-based methods where the resolution of the estimated depth map in the next stage generally is twice that of the previous stage, the resolution of the depth maps estimated in the stage 0 and stage 1, stage 2 and stage 3 is the same in EPNet, so we can increase the number of depth hypothesis samples as much as possible without adding a lot of overhead of memory and run-time. In this case, the EPNet consists of 5 stages which estimates the depth with resolution of $1/8H \times 1/8W$ at the coarsest stage and estimates the depth with resolution of $1/2H \times 1/2W$ at the finest stage. Note that, avoiding estimating the depth map at full resolution can greatly decrease the consumption of memory and run-time which is also the main reason for the high efficiency of some methods (Wang et al. 2021, 2022a).

Experiments

Datasets and Evaluation Metrics

Datasets The DTU dataset (Aanæs et al. 2016), BlendedMVS (Yao et al. 2020) dataset, Tanks and Temples dataset (Knapitsch et al. 2017), and ETH3D high-res benchmark (Schöps et al. 2017) are used. DTU (Aanæs et al. 2016) consists of more than 100 scenes captured under 7 different lighting conditions, which is split into the training set, validation set and evaluation set as (Ji et al. 2017) does and preprocessed as (Yao et al. 2018) does. The BlendedMVS (Yao et al. 2020) is a large-scale dataset containing 17k samples of 113 scenes, which is divided into the training set and validation set. Tanks and Temples (Knapitsch et al. 2017) is composed of both realistic outdoor and indoor scenes, which is further divided into *intermediate* subset and *advanced* subset. ETH3D (Schöps et al. 2017) also contains outdoor and indoor scenes captured in realistic environments, but the baseline between views is wider compared with that in Tanks and Temples whose data are presented as video sequences.

Evaluation Metrics The *Accuracy* (Acc.) and *Completeness* (Comp.) of the distance metric are used to measure the quality of reconstructed point clouds for DTU dataset, while the accuracy and completeness of the percentage metric are adopted for Tanks and Temples dataset and ETH3D high-res benchmark. We calculate the average of the mean accuracy and the mean completeness as the *overall* score for DTU dataset and F_1 score for the other two datasets.

Implementation Details

Following (Peng et al. 2022), EPNet is first trained on DTU and then fine-tuned on BlendedMVS. During training, the resolution of images for DTU and BlendedMVS is set to 640×512, and the number of views N is 5 for DTU and 7 for BlendedMVS. The EPNet is composed of 5 stages, the number of depth hypotheses for each stage is set to 32, 16, 8, 8, 8, and the corresponding depth sampling range decays by 0.5 for the second stage and 0.25 for the rest.

The proposed method is implemented by PyTorch (Paszke et al. 2019). Adam (Kingma and Ba 2014) is used as the optimizer. For the model including the HEPR module, we first train the MSDE alone for 1 epoch, then train the HEPR module alone for 2 epochs to warm up this branch, and finally train the full model for another 9 epochs, and the initial learning rate 0.001 is decreased by half at 8-th, 10-th and 11-th epoch. While for the model excluding the HEPR module, it is trained for 10 epochs with the initial learning rate 0.001 decreased by half at 6-th, 8-th and 9-th epochs. When fine-tuned on BlendedMVS, the model is trained for 10 epochs with the initial learning rate 0.0001 decreased by half at 6-th and 8-th epochs. The robust training strategy (Wang et al. 2021) is used in all models for better learning of pixel-wise visibility. The experiments are performed on one GeForce RTX 2080Ti GPU. When testing, we first use the proposed model to predict depth maps for all input views, then use the probability map and geometric consistency as previous methods do (Zhang et al. 2020; Yan et al. 2020) to filter and fuse the depth maps to a unified 3D point cloud.

Benchmark Performance

Results on DTU We evaluate EPNet on the evaluation set of DTU dataset using the model trained on the training set of DTU dataset only, where the number of views is 5 and the resolution of images is 1600×1152. Quantitative results of reconstructed point clouds are given in Table 1. As

	Method	Acc. ↓	Comp.↓	Overall ↓
Tra.	Gipuma	**0.283**	0.873	0.578
	COLMAP	0.411	0.657	0.534
Learning	MVSNet	0.396	0.527	0.462
	Vis-MVSNet	0.369	0.361	0.365
	IterMVS	0.373	0.354	0.363
	CasMVSNet	0.325	0.385	0.355
	EPP-MVSNet	0.413	0.296	0.355
	PatchmatchNet	0.427	0.277	0.352
	UGNet	0.334	0.330	0.332
	PVSNet	0.337	0.315	0.326
	Effi-MVS	0.321	0.313	0.317
	NP-CVP-MVSNet	0.356	0.275	0.315
	UniMVSNet	0.352	0.278	0.315
	TansMVSNet	0.321	0.289	0.305
	MVSTER	0.340	**0.266**	0.303
	GBi-Net	0.327	0.268	**0.298**
	EPNet	0.299	0.323	0.311

Table 1: Quantitative results of reconstructed point clouds on DTU evaluation set by using the distance metric [*mm*] (lower is better).

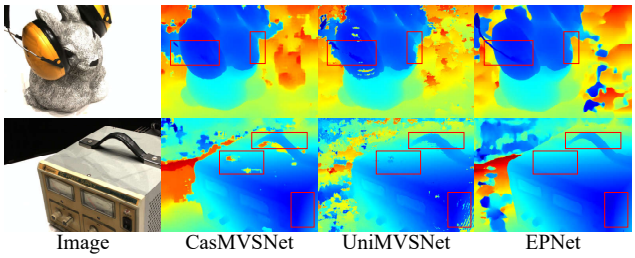

Figure 4: Qualitative comparisons of estimated depth maps with CasMVSNet (Gu et al. 2020) and UniMVSNet (Peng et al. 2022) in terms of Scan33 and Scan11 in DTU dataset.

shown, our method achieves competitive results compared with existing methods, and the *Accuracy* of our method is second only to Gipuma (Galliani, Lasinger, and Schindler 2015) among all methods. We also compare our method with state-of-the-art methods CasMVSNet (Gu et al. 2020) and UniMVSNet (Peng et al. 2022) in terms of the estimated depth maps on Scan33 and Scan11, qualitative comparisons are visualized in Figure 4. As illustrated, our method can recover the depth of thin structures and object boundaries well, and the edge of our depth map is able to align with the image better compared with other methods.

Results on Tanks and Temples To validate the generalization ability of EPNet, we test our method on Tanks and Temples benchmark without any fine-tuning on it, results are given in Table 2. The number of views is 11 and the resolution of images is 1920 × 1024. For a fair comparison, we respectively use the model trained on DTU only, and the model trained on DTU and fine-tuned on BlendedMVS to test on Tanks and Temples. As illustrated, our method achieves excellent performance for both models trained on DTU only and fine-tuned on BlendedMVS. Specially, our DTU-only trained model outperforms all learning-based methods

	Method	Intermediate (↑)			Advanced (↑)		
		Acc.	Comp.	F_1	Acc.	Comp.	F_1
Tra.	COLMAP	43.16	44.48	42.14	33.65	23.96	27.24
	ACMM	49.19	70.85	57.27	35.63	34.90	34.02
DTU	MVSNet	40.23	49.70	43.48	-	-	-
	NP-CVP-MVSNet	45.82	80.12	57.53	-	-	-
	CasMVSNet	47.62	74.01	56.84	29.68	35.24	31.12
	PatchmatchNet	43.64	69.37	53.15	27.27	41.66	32.31
	IterMVS	46.82	73.50	56.22	28.04	42.60	33.24
	PVSNet	53.71	63.88	56.88	29.43	41.17	33.46
	Effi-MVS	47.53	71.58	56.88	32.23	41.90	34.39
	EPNet	53.26	71.60	60.46	30.75	44.12	35.80
BlendedMVS	NP-CVP-MVSNet	47.05	**84.69**	59.64	-	-	-
	Vis-MVSNet	54.44	70.48	60.03	30.16	41.42	33.78
	IterMVS	47.53	74.69	56.94	28.70	44.19	34.17
	PVSNet	48.20	78.63	59.11	32.93	41.41	35.51
	EPP-MVSNet	53.09	75.58	61.68	**40.09**	34.63	35.72
	TransMVSNet	55.14	76.73	63.52	33.84	44.29	37.00
	UGNet	56.40	72.93	63.12	38.36	37.86	37.12
	GBi-Net	54.48	71.25	61.42	30.58	48.83	37.32
	MVSTER	51.17	77.50	60.92	33.23	45.90	37.53
	UniMVSNet	**57.54**	73.82	**64.36**	33.76	47.22	38.96
	EPNet	57.01	72.57	63.68	34.26	**50.54**	**40.52**

Table 2: Quantitative results on Tanks and Temples dataset using percentage metric (%) (higher is better). Methods are separated into three categories (from top to bottom): traditional, trained on DTU, and trained or fine-tuned on BlendedMVS.

that also are trained on DTU only and traditional methods (Schönberger et al. 2016; Xu and Tao 2019) on both *advanced* subset and *intermediate* subset. For methods trained or fine-tuned on BlendedMVS, our method obtains comparable performance to state-of-the-art methods. It is worth noting that our method under this condition achieves the best result on the *advanced* subset among all published works. Moreover, our method performs much better on two subsets in terms of F_1 score compared with state-of-the-art efficient methods PatchmatchNet (Wang et al. 2021), IterMVS (Wang et al. 2022a), Effi-MVS (Wang, Li, and Dai 2022), GBi-Net (Mi, Di, and Xu 2022) and MVSTER (Wang et al. 2022b), which further verifies the superiority of our method.

Results on ETH3D We further verify the generalization ability of our method on more challenging ETH3D high-res benchmark without any fine-tuning on it, quantitative results of point clouds on ETH3D high-res benchmark are shown in Table 3. The number of views is 7 for the model trained on DTU and 10 for the model fine-tuned on BlendedMVS, the resolution of images is 2432 × 1600. Obviously, our method is superior to both traditional methods and learning-based methods. For the model solely trained on DTU, our method performs better than all learning-based methods listed in the Table 3. We obtain much better performance on both training set and test set when the model is fine-tuned on BlendedMVS, which not only performs better than state-of-the-art traditional method ACMM (Xu and Tao 2019) but also state-of-the-art learning-based methods. This demonstrates wonderful generalization ability of the EPNet.

	Method	Training set (↑) Acc.	Comp.	F_1	Test set (↑) Acc.	Comp.	F_1
Tra.	COLMAP	**91.85**	55.13	67.66	**91.97**	62.98	73.01
	ACMM	90.67	70.42	78.86	90.65	74.34	80.78
DTU	PVSNet	67.84	69.66	67.48	66.41	80.05	72.08
	PatchmatchNet	65.43	64.81	64.21	69.71	77.46	73.12
	IterMVS	73.62	61.87	66.36	76.91	72.65	74.29
	EPNet	71.90	66.17	68.08	72.87	78.80	75.46
BlendedMVS	GBi-Net	73.17	69.21	70.78	82.02	75.65	78.40
	MVSTER	68.08	76.92	72.06	77.09	82.47	79.01
	IterMVS	79.79	66.08	71.69	84.73	76.49	80.06
	UGNet	79.62	67.57	72.78	82.78	79.87	80.83
	PVSNet	83.00	71.76	76.57	81.55	83.97	82.62
	EPP-MVSNet	82.76	67.58	74.00	85.47	81.79	83.40
	Vis-MVSNet	83.32	65.53	72.77	86.86	80.92	83.46
	EPNet	79.36	**79.28**	**79.08**	80.37	**87.84**	**83.72**

Table 3: Comparisons of reconstructed point clouds on ETH3D using percentage metric (%) at threshold $2cm$ (higher is better). Methods are separated into three categories (from top to bottom): traditional, trained on DTU, and trained or fine-tuned on BlendedMVS.

Model	CPC	HEPR	DTU (↓) Acc.	Comp.	Overall	ETH3D (↑) F_1
Model-A			0.360	0.300	0.330	63.84
Model-B	✓		0.358	0.295	0.327	64.33
EPNet	✓	✓	0.299	0.323	0.311	68.08

Table 4: Ablation results on DTU evaluation set by using the distance metric [*mm*] (lower is better) and ETH3D training set by using F_1 score (%) at threshold $2cm$ (higher is better).

Ablation Study

The ablation study is performed to validate the effectiveness of each component in the proposed method. All experiments in this part are conducted on the model trained with the DTU training set, and tested on the DTU evaluation set with resolution 1600×1152 and number of views $N = 5$ and on the ETH3D training set with resolution 2432×1600 and number of views $N = 7$. Due to that the CPC loss is able to transfer the error between the GT and the estimated depth map in the depth domain to the image domain, the error of detailed regions can be magnified to be optimized better. This is beneficial for predicting more accurate depth maps. As shown in Table 4, Model-B achieves better results compared to Model-A in terms of the reconstruction *Acc.* and *overall* quality on DTU and better generalization ability on ETH3D. Moreover, the performance of full model is greatly improved when the HEPR is introduced. This is because the HEPR module not only can help to refine the erroneous areas estimated by the multi-scale depth estimation module but also can achieve edge-preserving upsampling to enable the edge of the estimated depth map to align with the context of the image better. As a result, the issue of blur artifacts can be alleviated which greatly improves the quality of depth maps and contributes to more accurate point clouds. The qualitative results shown in Figure 5 further validate the effectiveness of the CPC loss and HEPR module.

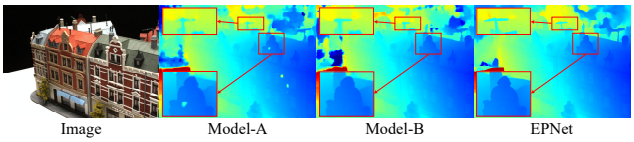

Figure 5: Qualitative comparisons of estimated depth maps for ablation study in terms of Scan15 in DTU dataset.

Model	Hypotheses	DTU (↓) Overall	Mem.	Time	ETH3D (↑) F_1
Model-C*	48,32,8	0.325	8065	0.7272	59.77
Model-D	32,16,8	0.339	2907	0.1792	59.22
Model-A	32,16,8,8,8	0.330	2843	0.1920	63.84

Table 5: Experiments results on DTU evaluation set by using the distance metric [*mm*] (lower is better), memory consumption (MB) and run-time (s), and on ETH3D training set by using F_1 score (%) at threshold $2cm$ (higher is better). * denotes the resolution of final output depth maps is $H \times W$, otherwise it is $1/2H \times 1/2W$. Hypotheses give the number of depth hypotheses for each stage. $\underline{Underline}$ denotes the resolution of depth maps output by these stages is the same.

Memory and Run-time Comparison

Table 5 presents experimental results of several variants of Model-A. It can be observed from the experimental results of Model-C* and Model-D that estimating depth maps with resolution from $1/8H \times 1/8W$ to $1/2H \times 1/2W$ rather than from $1/4H \times 1/4W$ to $H \times W$ can greatly reduce the memory and run-time consumption with just a little performance degradation. Comparing Model-D with Model-A, it can be seen that stacking stages at the same resolution can increase the number of depth hypotheses without adding extra memory consumption and with a small increase in run-time, which is helpful for improving performance. Moreover, combing Model-A with HEPR and CPC to get the EPNet, it consumes about 2753 *MB* and $0.2298s$ to infer a depth map on DTU. We further compare the memory and run-time consumption with several top-performing learning-based methods, the results are reported in Figure 1 (a). As shown in Figure 1, compared with the most efficient methods PatchmatchNet (Wang et al. 2021), IterMVS (Wang et al. 2022a) and MVSTER (Wang et al. 2022b), the memory and run-time overhead of our method are comparable, but EPNet has much more powerful generalization ability.

Conclusion

We present an efficient edge-preserving multi-view stereo network in this paper. The proposed HEPR module can help the EPNet achieve edge-preserving depth upsampling and correct erroneous regions in the depth maps estimated by the MSDE module of EPNet. Moreover, to enhance gradient flows of detailed regions, a CPC loss is proposed, so that the network can be optimized better. In addition, by carefully designing the structure of the multi-scale depth estimation module, our method can estimate depth maps efficiently. Extensive experiments show that EPNet achieves state-of-the-art performance on various datasets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62176096 and 61991412.

References

Aanæs, H.; Jensen, R. R.; Vogiatzis, G.; Tola, E.; and Dahl, A. B. 2016. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2): 153–168.

Cheng, S.; Xu, Z.; Zhu, S.; Li, Z.; Li, L. E.; Ramamoorthi, R.; and Su, H. 2020. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2524–2534.

Ding, Y.; Yuan, W.; Zhu, Q.; Zhang, H.; Liu, X.; Wang, Y.; and Liu, X. 2022a. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8585–8594.

Ding, Y.; Zhu, Q.; Liu, X.; Yuan, W.; Zhang, H.; and Zhang, C. 2022b. KD-MVS: Knowledge Distillation Based Self-supervised Learning for Multi-view Stereo. In *European Conference on Computer Vision*, 630–646. Springer.

Galliani, S.; Lasinger, K.; and Schindler, K. 2015. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, 873–881.

Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; and Tan, P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2495–2504.

Hui, T.-W.; Loy, C. C.; and Tang, X. 2016. Depth map super-resolution by deep multi-scale guidance. In *European conference on computer vision*, 353–369. Springer.

Ji, M.; Gall, J.; Zheng, H.; Liu, Y.; and Fang, L. 2017. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, 2307–2315.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Ma, X.; Gong, Y.; Wang, Q.; Huang, J.; Chen, L.; and Yu, F. 2021. EPP-MVSNet: Epipolar-Assembling Based Depth Prediction for Multi-View Stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5732–5740.

Mi, Z.; Di, C.; and Xu, D. 2022. Generalized Binary Search Network for Highly-Efficient Multi-View Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12991–13000.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 8026–8037.

Peng, R.; Wang, R.; Wang, Z.; Lai, Y.; and Wang, R. 2022. Rethinking Depth Estimation for Multi-View Stereo: A Unified Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8645–8654.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Schönberger, J. L.; Zheng, E.; Frahm, J.-M.; and Pollefeys, M. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, 501–518. Springer.

Schöps, T.; Schönberger, J. L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; and Geiger, A. 2017. A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2538–2547.

Su, W.; Xu, Q.; and Tao, W. 2022. Uncertainty Guided Multi-View Stereo Network for Depth Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7796–7808.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, 402–419. Springer.

Tosi, F.; Liao, Y.; Schmitt, C.; and Geiger, A. 2021. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8942–8952.

Wang, F.; Galliani, S.; Vogel, C.; and Pollefeys, M. 2022a. IterMVS: Iterative Probability Estimation for Efficient Multi-View Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8606–8615.

Wang, F.; Galliani, S.; Vogel, C.; Speciale, P.; and Pollefeys, M. 2021. PatchmatchNet: Learned Multi-View Patchmatch Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14194–14203.

Wang, S.; Li, B.; and Dai, Y. 2022. Efficient Multi-View Stereo by Iterative Dynamic Cost Volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8655–8664.

Wang, X.; Zhu, Z.; Huang, G.; Qin, F.; Ye, Y.; He, Y.; Chi, X.; and Wang, X. 2022b. MVSTER: epipolar transformer for efficient multi-view stereo. In *European Conference on Computer Vision*, 573–591. Springer.

Wei, Z.; Zhu, Q.; Min, C.; Chen, Y.; and Wang, G. 2021. AA-RMVSNet: Adaptive Aggregation Recurrent Multi-view Stereo Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6187–6196.

Xu, H.; Zhou, Z.; Wang, Y.; Kang, W.; Sun, B.; Li, H.; and Qiao, Y. 2021. Digging into uncertainty in self-supervised multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6078–6087.

Xu, Q.; Su, W.; Qi, Y.; Tao, W.; and Pollefeys, M. 2022. Learning Inverse Depth Regression for Pixelwise Visibility-Aware Multi-View Stereo Networks. *International Journal of Computer Vision*, 130(8): 2040–2059.

Xu, Q.; and Tao, W. 2019. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5483–5492.

Xu, Q.; and Tao, W. 2020. Learning Inverse Depth Regression for Multi-View Stereo with Correlation Cost Volume. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12508–12515.

Yan, J.; Wei, Z.; Yi, H.; Ding, M.; Zhang, R.; Chen, Y.; Wang, G.; and Tai, Y.-W. 2020. Dense Hybrid Recurrent Multi-view Stereo Net with Dynamic Consistency Checking. In *Proceedings of the European Conference on Computer Vision*.

Yang, J.; Alvarez, J. M.; and Liu, M. 2022. Non-parametric Depth Distribution Modelling based Depth Inference for Multi-view Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8626–8634.

Yang, J.; Mao, W.; Alvarez, J. M.; and Liu, M. 2020. Cost Volume Pyramid Based Depth Inference for Multi-View Stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4877–4886.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, 767–783.

Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; and Quan, L. 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5525–5534.

Yao, Y.; Luo, Z.; Li, S.; Zhang, J.; Ren, Y.; Zhou, L.; Fang, T.; and Quan, L. 2020. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1790–1799.

Zhang, J.; Yao, Y.; Li, S.; Luo, Z.; and Fang, T. 2020. Visibility-aware Multi-view Stereo Network. In *Proceedings of the British Machine Vision Conference (BMVC)*.