Adaptive Range guided Multi-view Depth Estimation with Normal Ranking Loss

Yikang Ding[1⋆] Zhenyang Li[1∗] Dihe Huang[1] Kai Zhang[1]
Zhiheng Li[1⋆⋆] Wensen Feng[2∗∗]

[1]Tsinghua University [2]Huawei Company

Abstract. Deep learning algorithms for Multi-view Stereo (MVS) have surpassed traditional MVS methods in recent years, due to enhanced reconstruction quality and runtime. Deep-learning based methods, on the other side, continue to generate overly smoothed depths, resulting in poor reconstruction. In this paper, we aim to Boost Depth Estimation (BDE) for MVS and present an approach, termed as **BDE-MVSNet**, for reconstructing high-quality point clouds with precise depth prediction. We present a non-linear strategy that derives an adaptive depth range (ADR) from the estimated probability, motivated by distinctive differences in estimated probability between foreground and background pixels. ADR also tends to decrease fuzzy boundaries via upsampling low-resolution depth maps between stages. Additionally, we provide a novel structure-guided normal ranking (SGNR) loss that imposes geometrical consistency in boundary areas by using the surface normal vector. Extensive experiments on *DTU* dataset, *Tanks and Temples* benchmark, and *BlendedMVS* dataset demonstrate that our method outperforms known methods and achieves state-of-the-art performance.

1 Introduction

Multi-view Stereo (MVS) is the process of reconstructing the dense 3D geometry of an observed scene using posed images and camera parameters. It is a key problem in computer vision, with applications to various domains such as augmented and virtual reality, robotics, and 3D modeling. Although MVS has been studied for several decades, computing a high-quality 3D reconstruction in the presence of occlusions, low-textured regions, and blur remains a challenge [1]. Convolutional Neural Networks (CNNs) were adopted by MVS methods as an alternative for hand-crafted matching metrics and regularization schemes [2,3,4,5,6] following the success of deep learning in many fields of computer vision, offering a significant improvement to the completeness of the reconstructed model and the runtime required for generating it [7,8,9,10,11,12,13]. The basic learning-based MVS approach [7] begins with the extraction of deep features using a 2D CNN. The corresponding cost maps, each generated by considering

[⋆] Equal contribution.
[⋆⋆] Corresponding author (zhhli@mail.tsinghua.edu.cn, fengwensen@huawei.com).

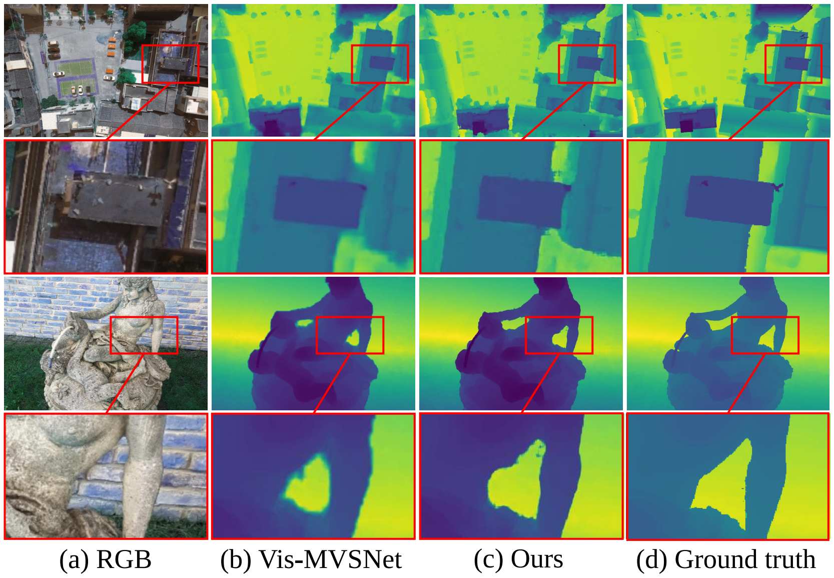

Fig. 1. Comparison of depth maps. (a) RGB images. (b)-(d) Depth maps predicted by Vis-MVSNet [9] and our approach, and the respective ground truth depth. Our method predicts much more accurate depth compared to Vis-MVSNet [9].The resolution of input images is 640 × 512.

the variance of warped activation maps, are then stacked to form a cost volume over different depth hypotheses. Before applying Softmax and regressing the depth of the reference image, a 3D CNN is applied to finalize the cost volume normalization. Recent state-of-the-art MVS approaches have recommended various optimizations, such as using coarse-to-fine processing of activation maps [8,9,10,11,12], since runtime and memory expand cubically with spatial resolution and have a complexity of $O(n^3)$. The above mentioned methods using depth range estimation in pixels [10,11] and imposing pixel-wise visibility constraints [9]. Despite these recent advancements, learning-based MVS methods still produce overly smoothed depth and imprecise boundaries, which are harmful to the reconstruction results.

In this work, we propose a MVS method, named BDE-MVSNet, which aims at boosting depth estimation, especially on the boundaries and the background areas in multi-view stereo tasks. The uncertainty associated with depth estimation in the foreground and background/boundary regions motivated our method. After a single estimating step, we observe that the depth at foreground pixels is assigned with high probability, whereas the depth at background pixels is commonly assigned with low probability. Thus, we propose an adaptive depth range (ADR) method that computes the depth range per pixel in a non-linear way from its probability. As a consequence, depth ranges for boundary and background pixels are kept broad, while foreground depth is sampled precisely, producing improved accuracy. We take inspiration from Monocular Depth Estimation (MDE) approaches [14,15,16] and present a novel structure-guided normal ranking (SGNR) loss, which promotes geometrical consistency using the surface normal vector, to improve depth regression. With the help of ADR strategy and SGNR loss, we estimate the depth with only two stages while processing relative high-resolution images. Without the requirement to up-sample the low-resolution depth maps, this method provides an alternative to three-stage cascade approaches like Vis-MVSNet [9], UCSNet [10] and CasMVSNet [8]. We employ Vis-MVSNet [9] as our baseline and implement our method upon it. Fig. 1 shows depth images predicted by Vis-MVSNet [9] and our method. *BDE-MVSNet* is able to produce much more accurate depth and sharper boundaries, even in challenging scenarios. We evaluate BDE-MVSNet on commonly benchmarked MVS datasets, namely DTU [17], Tanks and Temples [18] and BlendedMVS [19]. Extensive experiments show BDE-MVSNet achieves state-of-the-art performances.

To summarize, the following are our main contributions.

- We propose BDE-MVSNet which can predict accurate depth and reconstruct high-quality point clouds.
- We introduce ADR strategy to derive a per-pixel depth interval in a non-linear manner, which helps predict accurate depth in only two stages.
- We propose SGNR loss and show it can help in predicting sharp boundary and in decreasing tailing errors in the reconstructed point cloud.
- We qualify the performance of our method on multiple MVS datasets and show it achieves state-of-the-art performance.

2 Related Work

2.1 Learning-based MVS

The accuracy and efficiency of learning-based methods are directly affected by the number of depth hypotheses, the interval from which they are sampled and the spatial resolution of the activation maps, which determine the dimensions of the regularized cost volume [8]. Recent state-of-the-art learning-based MVS methods have suggested different strategies for extending the basic learned MVS paradigm (MVSNet [7]) and optimizing the aforementioned trade-offs. Different recurrent models were suggested for regularizing the cost volume in a sequential manner [20,21,22]. However, while decreasing memory consumption, sequential processing does not scale well with spatial resolution. Cascade methods [8,10,9,11,12,13] proposed instead to form a feature pyramid and regress depth maps in a coarse-to-fine manner. Typically, at the coarsest stage, depth is regressed as in the MVSNet paradigm. As the resolution increases, the number of depth hypotheses is decreased, resulting in improved efficiency. The depth range at a given stage is often centered around the depth estimated at the previous coarser stage resulting in more accurate depth ranges to sample from [8,10,9,11].

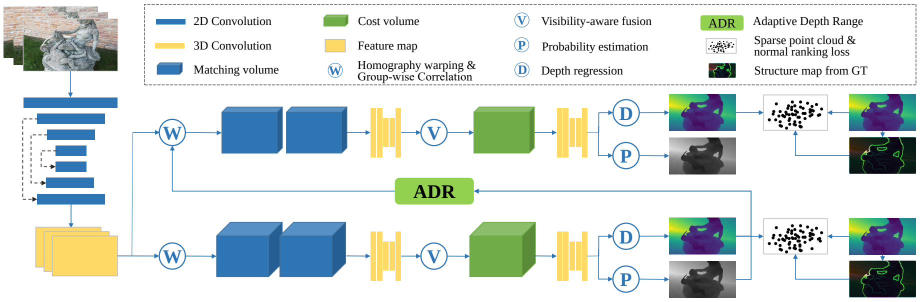

Fig. 2. Overview of BDE-MVSNet. We use a visibility-aware architecture, similar to [9], but apply only two processing stages using the same resolution feature maps. Our ADR strategy determines the per-pixel depth range at the second stage based on the estimated probability. We further apply our SGNR loss to enforce geometrical consistency.

One drawback of coarse-to-fine methods is the need to up-sample depth maps from stage to stage, which can yield fuzzy boundaries.

2.2 MVS Loss Functions

Learning-based MVS methods are typically optimized to minimize the L1 loss between the predicted depth map and ground truth (GT) depth [7,8,12,21,10] or cross-entropy loss between the predicted probability volume and the GT probability, which is generated by one-hot encoding [22,20,21]. A recent extension to this formulation is introduced by Vis-MVSNet [9], where the per-pixel visibility is explicitly addressed by computing the cost volume per pair of reference and source images before constructing and regularizing the joint cost volume. The common L1 loss is also extended to minimize the depth maps estimated in a pairwise and joint manner. Other extensions are also proposed to improve geometrical consistency through constraints on the surface normal vector [23,24]. Besides multi-view depth estimation, MDE approaches [14,15,25,26,27] also offered novel strategies for improving depth prediction. For example, ranking loss [14,15] and structure-guided point sampling [16] were shown to improve depth accuracy. In this work, we extend propose a novel SGNR loss based on MDE strategies and surface normal vector constraints.

3 Method

Given a collection of images and camera parameters (intrinsic matrices K, rotation matrices R and translation vector t), our proposed BDE-MVSNet aims to predict a dense depth map $d \in \mathbb{R}^{h \times w}$ for each reference image $I_0 \in \mathbb{R}^{h \times w \times 3}$, using respective m source images $\{I_i\}_{i=1}^{m}$ with highest co-visibility.

3.1 Network Architecture

The main architecture of our proposed BDE-MVSNet is shown in Fig. 2. It first extracts deep feature maps through a 2D U-Net from both reference and source images. Then, the feature maps feed into two branches, which refer to the two visibility aware-stages [9] with the same spatial resolution to construct and regularize 3D cost volumes. Between the two steps, our ADR strategy is used to calculate a per-pixel depth interval depending on the probability of the previous depth estimate. We train our model with a novel loss, whose components are the Vis-MVSNet loss and our proposed SGNR loss, which computes the structure map from the ground truth depth map and samples pair-wise points to calculate the normal ranking loss.

Feature Extraction We use a 2D U-Net to extract deep features from reference I_0 and source $\{I_i\}_{i=1}^{m}$ and process the finest-resolution feature maps, of size $\frac{h}{2} \times \frac{w}{2} \times 32$.

Cost Volume Construction Following [7,9], we construct pair-wise and joint cost volumes using differentiable homography warping and group-wise correlation. The warping process from I_i to I_0 can be described as:

$$H_{i,j} = K_i R_i (I - \frac{(t_0 - t_i) a_0^T}{d_j}) R_0^T K_0^{-1}, \tag{1}$$

where $H_{i,j}$ refers to the homography matrix at depth d_j and a_0 denotes the principle axis of the reference image. We first compute pairwise cost volumes and respective probability volumes [9] using group-wise correlation [28] and 3D CNN, which will then be fused to construct the final cost volume. Then, we apply a depth-wise 3D CNN with shape $1 \times 1 \times 1$ and a Softmax function to compute the probability volume $P \in \mathbb{R}^{N \times \frac{h}{2} \times \frac{w}{2}}$. The final depth with its probability map can be obtained from P using regression or winner-take-all. The generation of cost volume is identical for both stages, e.g. uses same spatial resolution and same sampling strategy. The main difference between the two stages lies in the prior depth range derivation. For the first stage, we use a fixed prior depth range for all pixels as in previous methods. While for the second stage, we update the depth range for each pixel using our ADR strategy (Section 3.2).

Depth Regression Given the probability volume P, we regress the predicted depth $\overline{d}(p)$ of each pixel p in the reference image by taking the probability-weighted mean of all N hypotheses:

$$\overline{d}(p) = \sum_{j=1}^{N} d_j \cdot P(p, j). \tag{2}$$

Depth Fusion At inference time, we apply our model to regress the depth maps of all images and then filter and fuse them to reconstruct the 3D point cloud as in [8], based on photometric and geometric consistency.

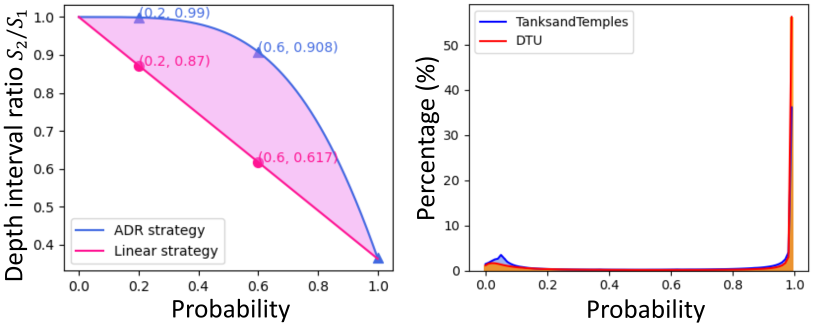

Fig. 3. Illustration of ADR. The left column shows comparison of ADR and linear strategy [11,10]. The right column shows the ditribution of the first stage's probability on DTU validation set [17].

3.2 Adaptive Depth Range

At the end of the first stage, the network produces a probability volume P and an estimated depth map $\overline{d}$. Given these two outputs, we can obtain an estimated probability $\overline{P}$ for each pixel, as in MVSNet [7]. The probability $\overline{P}(p)$ at pixel p provides an measurement for the uncertainty. We leverage this information for adapting the depth range at the next stage. We propose a non-linear strategy motivated by distinctive differences in pixel uncertainty between foreground and background areas. As shown in Fig. 3, when analyzing the distribution of estimated probability, we find foreground pixels are typically assigned with a probability of $0.9 \sim 0.999$, while pixels at boundary and background regions present a probability of $0.3 \sim 0.7$ and < 0.3 respectively. For depth estimations with high uncertainty, we would like to sample from a relatively wide range to decrease error rate. On the other hand, when the estimation is made with high certainty, a narrow range can help in achieving improved accuracy. Following this intuition, we propose our novel Adaptive Depth Range (ADR) method.

Given a depth range $[d_{min,i}, d_{max,i}]$, a fixed depth interval δ and a fixed number of depth hypotheses N_i for stage i, we can obtain the j-th depth hypothesis $d_{j,i}$ at the i-th stage,

$$\begin{cases} d_{j,i} = d_{min,i} + (j-1) \cdot \delta \\ d_{max,i} = d_{min,i} + N_i \cdot \delta \end{cases}, \tag{3}$$

We simply take a fixed $[d_{min,i}, d_{max,i}]$ for all pixels in the first stage ($i = 1$). While for the second stage, our ADR strategy computes a pixel-specific depth range $[\overline{d}_1(p) - ADR(p), \overline{d}_1(p) + ADR(p)]$ based on the estimated depth $\overline{d}_1(p)$ and probability $\overline{P}_1(p)$, where $ADR(p)$ can be written as,

$$ADR(p) = \frac{1}{2} \cdot N_2 \cdot s(p) \cdot \delta. \tag{4}$$

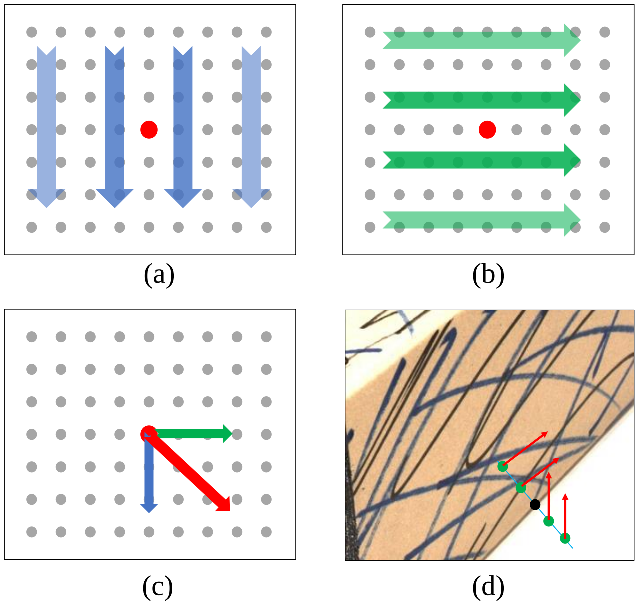

Fig. 4. Illustration of our structure-guided sampling scheme. (a)-(b): Sobel operator in y and x direction. (c): Gradient of center pixel (red point). (d): Four sampled points (in green) around a center point (in black) at a boundary area.

$s(p)$ is a pixel-specific scaling factor which can be obtained by,

$$s(p) = \cos(k \cdot \overline{P}_1^2(p)), \tag{5}$$

where k is a hyper-parameter (we set it to 1.2) which controls the range scaling. We illustrate the differences between the linear strategy employed by [11,10] and our ADR strategy in Fig. 3. ADR tries to maintain a large depth range for low-probability pixels and strictly narrow the depth range for high-probability pixels, which helps to assign proper depth range and predict accurate depth in fewer stages.

3.3 Structure-Guided Normal Ranking Loss

Inspired by recent advancements in MDE, we propose to modify the *pairwise ranking loss* [14] for MVS, and suggest imposing geometric consistency by updating the ordinal label using the *surface normal*. Our proposed SGNR loss mainly consists of two steps: structure-guided sampling and pairwise normal ranking loss.

Structure-Guided Sampling In order to accommodate for the typical MVS setting (dataset), where the background is usually masked in both RGB and depth images, we propose a modified version of four-point sampling scheme [16] to sample the point pairs $S = \{(p_{i,0}, p_{i,1})\}_{i=1}^{N}$. Given a reference image I, we first convert it to a gray-scale image I^* and use the *Sobel* operator to get the gradient maps G_x, G_y and gradient magnitude G (Fig. 4(a)-(c)). Before we get the final edge map, we compute a solid-region mask to avoid sampling points in masked background regions. We execute this operation by reducing 16 pixels along the orthogonal line crossing the background-mask to get the solid-region mask $M_{solid} \in \mathbb{R}^{h \times w}$. Then the final edge map E can be obtained by applying a solid-region mask to avoid sampling points in masked background regions by

$$E = I^*[G(p) \geq \alpha \cdot max(G)]_{p \in M_{solid}}, \tag{6}$$

where α is a threshold which controls the density of E, and we set it to 0.05 in our experiments.

For each edge point $e = (x, y) \in E$, we sample four points $\{(x_k, y_k)\}_{k=0}^{3}$ by

$$\begin{cases} x_k = x + \delta_k G_x(e)/G(e) \\ y_k = y + \delta_k G_y(e)/G(e) \end{cases}, \tag{7}$$

where we sample $\delta_0 < \delta_1 < 0 < \delta_2 < \delta_3$ within a small distance range β from the edge point. Given four sampled points p_0, p_1, p_2, p_3, we form three pairs of points for pairwise ranking: $[(p_0, p_1), (p_1, p_2), (p_2, p_3)]$. Similar to [16], we also sample some points using a random sampling scheme in order to preserve global structures. Specifically, we sample $3n$ pairs points through the four-point sampling scheme and n pairs through the random sampling scheme, where n is set to 1000 in our experiments.

Pairwise Normal Ranking Loss The pair-wise ranking loss performs well in the majority of cases, however it loses geometry consistency in plane or boundary regions. Take a look at the box-like object in Fig. 4(d), where a sample of four points is chosen from the area around an edge point at the object's boundary. Contrary to the [14], we believe it is irrational to suppose nearby points have the same depth estimation while ignoring their crucial geometrical constraints. In our work, we consider the inner point pair (p_0, p_1) and outer point pair (p_2, p_3) have the same surface normal vector, due to the pairs are relatively close in spatial. To better leverage the geometrical constraints, we propose to incorporate the surface normal vector in the loss formulation. After we have the GT depth map D^* with its corresponding camera intrinsic matrix K and the sampled points set S from our 4-point sampling scheme, we first select eight neighboring points $\{(x_{n,i}, y_{n,i})\}_{i=0}^{7}$ for each point p in S. In order to calculate the surface normal $n_p \in \mathbb{R}^{3\times1}$of p, we unproject the center point p and its neighboring points into 3D space as below,

$$P^* = K^{-1} D^*(p) p \tag{8}$$

Then, we compute the surface normal vector using three points for eight times as below,

$$\overrightarrow{N_{p,i}^*} = \overrightarrow{P^* P_{i,0}^*} \times \overrightarrow{P^* P_{i,1}^*}, \tag{9}$$

where $\{(P^*_{i,0}, P^*_{i,1})\}_{i=0}^{7}$ indicates two neighboring points of center point P^*. The final surface normal of p in GT depth map is presented as,

$$\overrightarrow{N^*_p} = \frac{1}{8}\sum_{i=0}^{7}\overrightarrow{N^*_{p,i}} \tag{10}$$

We can compute the surface normal map N^* and N with respect to GT depth map D^* and predicted depth map D in this way. After that, we calculate the surface normal vectors at each point and assign an ordinal label to each pair of points $(p_0, p_1) \in S$ based on their *cosine similarity*,

$$l = \begin{cases} 1, & |\cos(n^*_0, n^*_1)| \le \frac{1}{1+\tau} \\ 0, & otherwise \end{cases}, \tag{11}$$

where n^*_0 and n^*_1 are the GT depth-derived surface normal vectors corresponding to p_0 and p_1, and τ is a tolerance threshold. The *SGNR* loss for (p_0, p_1) is then given by,

$$\phi_{sgnr}(p_0, p_1) = \begin{cases} \log(1 + \exp(-\tan(\frac{|n_0,n_1|+\epsilon}{2}))), & l = 1 \\ (\tan(\frac{|n_0,n_1|+\epsilon}{2}))^2, & l = 0 \end{cases}, \tag{12}$$

where n_0 and n_1 are the surface normal vectors computed using the predicted depth map for p_0 and p_1 respectively, $|n_0, n_1|$ refers to the angle between n_0 and n_1, ϵ is a perturbation which we set to 1×10^{-4}.

In a word, the SGNR loss enforces the model to predict similar normal vectors for pairs with similar surface normals and dissimilar them otherwise. Our total SGNR loss is given by,

$$L_{sgnr} = \frac{1}{|S|}\sum_{(p_0,p_1)\in S}\phi_{sgnr}(p_0, p_1). \tag{13}$$

where $|S|$ refers to the number of point pairs.

3.4 Total Loss Function

The total loss function of BDE-MVSNet is composed of SGNR loss and the losses in Vis-MVSNet: the pair-wise L1 loss L_1^{pair}, the pair-wise joint loss L^{joint} and the L1 loss of the final depth map L^{final} for each stage. The final loss formulation is thus given by,

$$L_{total} = \sum_{k=1}^{2}\lambda_k[L^{final}_{1,k} + \frac{1}{m}\sum_{i=1}^{m}(L^{pair}_{1,k,i} + L^{joint}_{k,i}) + \alpha L_{sgnr,k}], \tag{14}$$

where λ_k is the weight for k-th stage, m is the number of source images and α is the weight of L_{sgnr}, which we set to be 0.5.

Table 1. Accuracy, completeness and overall scores of different MVS methods on the DTU test set. The resolution of input images is 864 × 1152.

Method	Acc.(mm) ↓	Comp.(mm) ↓	Overall(mm) ↓
Gipuma [30]	**0.283**	0.873	0.758
COLMAP [31]	0.400	0.664	0.532
MVSNet [7]	0.396	0.527	0.462
Vis-MVSNet [9]	0.369	0.361	0.365
PatchmatchNet [12]	0.427	**0.277**	0.352
CVP-MVSNet [32]	0.296	0.406	0.351
CasMVSNet [8]	0.346	0.351	0.348
UCSNet [10]	0.338	0.346	0.344
DDR-Net [11]	0.339	0.320	0.329
BDE-MVSNet	0.338	0.302	**0.320**

4 Experiments

4.1 Datasets

We evaluate our method using three commonly benchmarked MVS datasets: DTU [17], BlendedMVS [19] and Tanks and Temples [18]. DTU is an indoor-scene dataset, consisting of 124 scenes, scanned from 49 or 64 views under 7 different lighting conditions. BlendedMVS is a large-scale dataset that contains 17,000 MVS training samples covering a variety of 113 scenes. Tanks and Temples contains multiple realistic scenes. All the settings of evaluation datasets follow Vis-MVSNet [9].

4.2 Implementation Details

Our method is implemented with PyTorch [29] and trained on eight NVIDIA Tesla V100 cards. We optimize our network with Adam using a batch size of 16 and an initial learning rate of 0.001. We train for $160K$ iterations and decrease the learning rate by half at the $100K$, $120K$ and $140K$ iterations. During training, we use an image resolution of o 640 × 512 and set the number of source images to 3. For depth sampling strategy, we set $D_{max} = 128$ and the initial depth interval $\delta = 1$. We set $N_{d,1}, N_{d,2} = 32, 16$ for the number of depth hypotheses for stage 1 and 2 respectively. The ADR factor k is set to 1.2 and the weight of loss λ_k is set to 1.0 for all experiments.

4.3 Evaluation on DTU

Our proposed method benchmarked on the DTU [17] evaluation set. DTU dataset is divided into training, validation and evaluation sets. We train our model on the DTU training set and test on the evaluation set. We set the depth range to $[425mm, 905mm]$ and use 5 source views per image at a resolution of 1152×864 for depth estimation. As shown in Table 1, the Gipuma [30] and Patchmatch-Net [12] methods achieve the best accuracy and completeness respectively, while

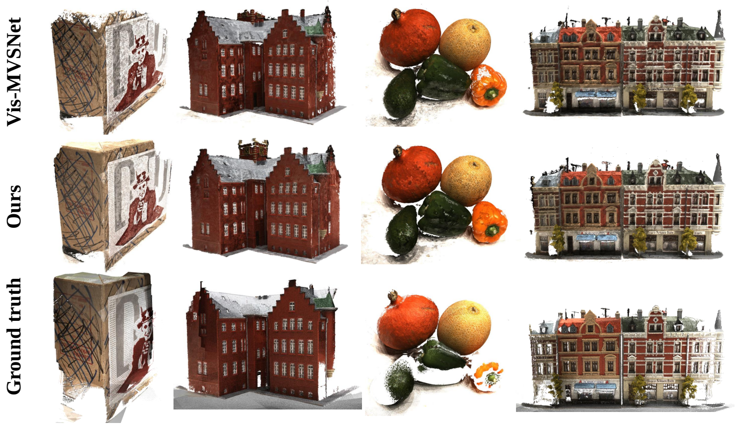

Fig. 5. Comparison of reconstructed point clouds on DTU validation set [17] between Vis-MVSNet and ours.

BDE-MVSNet outperforms other learning-based methods and traditional methods in terms of overall performance. Some qualitative results are shown in Fig. 5, compared with baseline method [9], BDE-MVSNet is able to reconstruct better point cloud results.

Table 2. The F-score of MVS methods on the Tanks and Temples intermediate test set (higher is better). The best method is highlighted in bold for each scene. The resolution of input images is 1920 × 1080.

Method	Pub.	Mean	Family	Francis	Horse	L.H.	M60	Panther	P.G.	Train
MVSNet [7]	ECCV18'	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
R-MVSNet [20]	CVPR19'	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
CVP-MVSNet [32]	CVPR20'	54.03	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54
UCSNet [10]	CVPR20'	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89
DDR-Net [11]	Arxiv20'	54.91	76.18	53.36	43.43	55.20	55.57	52.28	56.04	47.17
CasMVSNet [8]	CVPR20'	56.42	76.36	58.45	46.20	55.53	56.11	54.02	58.17	46.56
D2HC-RMVSNet [21]	ECCV20'	59.20	74.69	56.04	49.42	60.08	59.81	**59.61**	60.04	53.92
Vis-MVSNet [9]	BMVC20'	60.03	77.40	60.23	47.07	63.44	62.21	57.28	**60.54**	52.07
AA-RMVSNet [33]	ICCV21'	61.51	77.77	59.53	51.53	64.02	64.05	59.47	60.85	54.90
EPP-MVSNet [34]	ICCV21'	61.68	77.86	60.54	52.96	62.33	61.69	60.34	62.44	55.30
BDE-MVSNet (Ours)	-	**62.30**	**79.71**	**67.33**	**49.52**	**64.68**	**62.43**	58.28	58.15	**58.32**

4.4 Evaluation on Tanks and Temples

We evaluate our method on the Tanks and Temples dataset's intermediate set, [18]. To train BDE-MVSNet, we utilize the training set of the BlendedMVS dataset [19]. We set the number of source views to 7 and tested on images with a resolution of 1920 × 1080. All other hyper-parameters are set to the same values as in the training stage. The Table 2 reports the F-score of our method

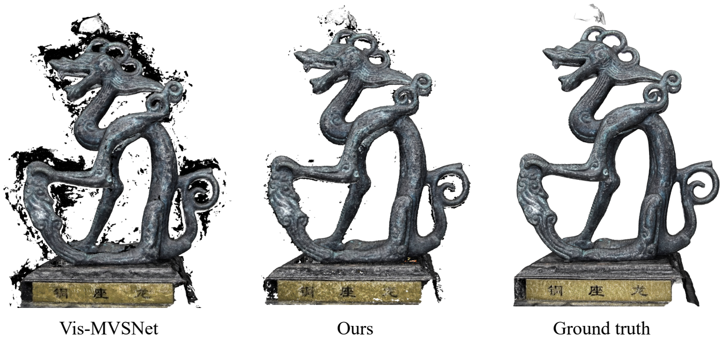

Fig. 6. 3D model of a challenging scene from BlendedMVS [19], reconstructed by Vis-MVSNet [9] and ours.

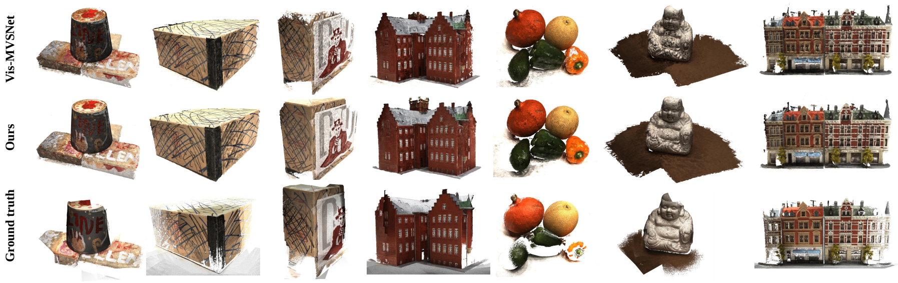

Fig. 7. More 3D model result of challenging scenes from BlendedMVS [19], reconstructed by Vis-MVSNet [9] and ours.

as well as other state-of-the-art learning-based MVS algorithms. BDE-MVSNet outperforms existing approaches in almost every scene due to superior depth prediction.

4.5 Evaluation on BlendedMVS

While MVS benchmarking typically concentrates only on the final output (the point cloud), we also report the quality of the predicted depth maps.

We follow the depth evaluation protocol of BlendedMVS [19] and report three metrics: the mean absolute error between the predicted and the ground truth depth maps, denoted as end point error (EPE), and the proportion in % of pixels with an error > 1 and > 3 in the scaled depth maps, denoted as e_1 and e_3, respectively.

We set the number of source images $m = 5$ with a resolution of 640×512 and set all other hyper-parameters as in the training phase. We train our model on the BlendedMVS dataset and evaluate using its validation set (Table 4). Our approach outperforms other state-of-the-art methods in terms of depth quality

Table 3. The effect of depth range update strategy on the quality of the depth estimation. Linear strategy is used in [11,10]. The resolution of input images is 640 × 512.

Depth Range Strategy	EPE ↓	e_1 ↓	e_3 ↓
None	1.35	17.82	6.55
Linear	1.24	16.81	5.67
ADR (ours)	**1.09**	**15.63**	**5.61**

Table 4. Accuracy (*EPE*, *e*1, *e*3), memory and runtime for depth map estimation results obtained with different MVS methods and ours. Results are reported for the BlendedMVS [19] validation set, using image resolution of 640 × 512.

Method	EPE ↓	e_1 ↓	e_3 ↓	Memory	Runtime
MVSNet [7]	1.49	21.98	8.32	5.50G	1.18s
CasMVSNet [8]	1.43	19.01	9.77	2.71G	0.44s
CVP-MVSNet [32]	1.90	19.73	10.24	-	-
DDR-Net [11]	1.41	18.08	8.32	-	-
Vis-MVSNet [9]	1.47	18.47	7.59	1.85G	0.56s
Ours	**1.06**	**15.14**	**5.13**	1.81G	0.47s

by a large margin. A qualitative comparison on a challenging scene between Vis-MVSNet and ours is shown in Fig. 6. Thanks to the predicted accurate depth, BDE-MVSNet is able to reduce the tailing error in the final point cloud result.

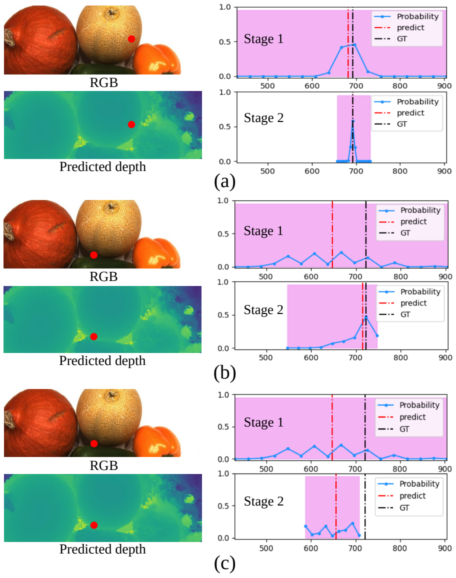

Fig. 8. Comparison between proposed ADR strategy with linear strategy [11,10] in 2 examples.

4.6 Runtime and Memory Analysis

We measure the runtime and memory cost of depth estimation using our method and several state-of-the-art methods on a Tesla V100 GPU as shown in Table 4. Our method achieves an improved memory and runtime cost compared to VisMVSNet [9], MVSNet [7]. Also provides a better runtime-memory trade-off compared to CasMVSNet.

4.7 Ablation Study

We carry out ablation experiments in order to evaluate the contribution of our proposed ADR strategy and SGNR loss. As shown in Table 5, ADR is the main contributor for the observed reduction in *EPE*, e_1 and e_3

Table 5. Ablation study on the BlendedMVS validation set. The resolution of input images is 640 × 512.

ADR	SGNR	EPE ↓	e_1 ↓	e_3 ↓
-	-	1.35	17.82	6.55
-	✓	1.30	16.98	5.99
✓	-	1.09	15.63	5.61
✓	✓	**1.06**	**15.14**	**5.13**

values. When employing it together with the SGNR loss we achieve a further improvement in depth accuracy.

We further compare the linear strategy employed by [11,10] to our proposed ADR strategy. The Table 3 reports the EPE, e_1 and e_3 of our model, when trained without updating the depth range between stages and when doing so either with a linear strategy or with ADR. While the linear depth update improves performance, ADR achieves better depth accuracy.

As shown in Fig.8. On the left, we show the RGB images and the corresponding predicted depth maps. On the right, we show the probability details of a pixel (red point in images) with depth intervals (pink). (a): For a point with high probability, ADR and linear strategy both narrow the depth range and predict accurate depth in the second stage. (b): For a point with medium probability at the boundary edge, ADR keeps the depth interval large and predicts accurate depth. (c): For the same point with (b), linear strategy narrows the depth interval and gets the wrong depth.

5 Conclusion

In this paper, we present BDE-MVSNet to boost depth estimation for multi-view stereo. We propose a non-linear method for deriving per-pixel depth range and a novel structure-guided normal ranking loss. Together these optimizations yield more accurate depth prediction and 3D reconstruction. To the best of our knowledge, our work is the first to directly tackle boundary prediction and can be used to improve the performance of learning-based MVS methods. As a result, our proposed BDE-MVSNet achieves state-of-the-art performance on multiple datasets.

References

1. Zhu, Q., Min, C., Wei, Z., Chen, Y., Wang, G.: Deep learning for multi-view stereo via plane Sweep: A survey (2021) 1
2. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on pattern analysis and machine intelligence **30** (2007) 328–341 1
3. Campbell, N.D., Vogiatzis, G., Hernández, C., Cipolla, R.: Using multiple hypotheses to improve depth-maps for multi-view stereo. In: European Conference on Computer Vision, Springer (2008) 766–779 1
4. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE transactions on pattern analysis and machine intelligence **32** (2009) 1362–1376 1
5. Tola, E., Strecha, C., Fua, P.: Efficient large-scale multi-view stereo for ultra high-resolution image sets. Machine Vision and Applications **23** (2012) 903–920 1
6. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 873–881 1
7. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: MVSNet: Depth inference for unstructured multi-view stereo. In: European Conference on Computer Vision. (2018) 767–783 1, 3, 4, 5, 6, 10, 11, 13
8. 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: IEEE Conference on Computer Vision and Pattern Recognition. (2020) 2492–2501 1, 2, 3, 4, 5, 10, 11, 13
9. Zhang, J., Yao, Y., Li, S., Luo, Z., Fang, T.: Visibility-aware multi-view stereo network. British Machine Vision Conference (2020) 1, 2, 3, 4, 5, 10, 11, 12, 13
10. Cheng, S., Xu, Z., Zhu, S., Li, Z., Li, L.E., Ramamoorthi, R., Su, H.: Deep stereo using adaptive thin volume representation with uncertainty awareness. In: IEEE Conference on Computer Vision and Pattern Recognition. (2020) 2521–2531 1, 2, 3, 4, 6, 7, 10, 11, 13, 14
11. Yi, P., Tang, S., Yao, J.: DDR-Net: Learning multi-stage multi-view stereo with dynamic depth range (2021) 1, 2, 3, 6, 7, 10, 11, 13, 14
12. Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M.: PatchmatchNet: Learned multi-view patchmatch stereo. In: IEEE Conference on Computer Vision and Pattern Recognition. (2021) 14194–14203 1, 2, 3, 4, 10
13. Ding, Y., Yuan, W., Zhu, Q., Zhang, H., Liu, X., Wang, Y., Liu, X.: Transmvsnet: Global context-aware multi-view stereo network with transformers. arXiv preprint arXiv:2111.14600 (2021) 1, 3
14. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: Advances in Neural Information Processing Systems. Volume 29. (2016) 3, 4, 7, 8
15. Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., Li, R., Luo, Z.: Monocular relative depth perception with web stereo data supervision. In: IEEE Conference on Computer Vision and Pattern Recognition. (2018) 311–320 3, 4
16. Xian, K., Zhang, J., Wang, O., Mai, L., Lin, Z., Cao, Z.: Structure-guided ranking loss for single image depth prediction. In: IEEE Conference on Computer Vision and Pattern Recognition. (2020) 608–617 3, 4, 8
17. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanaes, H.: Large scale multi-view stereopsis evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2014) 406–413 3, 6, 10, 11

18. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and Temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics **36** (2017) 1–13 3, 10, 11
19. Yao, Y., Luo, Z., Li, S., Zhang, J., Quan, L.: BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In: IEEE Conference on Computer Vision and Pattern Recognition. (2020) 1787–1796 3, 10, 11, 12, 13
20. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent MVSNet for high-resolution multi-view stereo depth inference. In: IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5525–5534 3, 4, 11
21. Yan, J., Wei, Z., Yi, H., Ding, M., Zhang, R., Chen, Y., Wang, G., Tai, Y.W.: Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In: Computer Vision – European Conference on Computer Vision. (2020) 674–689 3, 4, 11
22. Wei, Z., Zhu, Q., Min, C., Chen, Y., Wang, G.: AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network. IEEE International Conference on Computer Vision (2021) 3, 4
23. Yin, W., Liu, Y., Shen, C.: Virtual Normal: Enforcing geometric constraints for accurate and robust depth prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1–1 4
24. Long, X., Liu, L., Theobalt, C., Wang, W.: Occlusion-aware depth estimation with adaptive normal constraints. In: Computer Vision – European Conference on Computer Vision. (2020) 640–657 4
25. Chakrabarti, A., Shao, J., Shakhnarovich, G.: Depth from a single image by harmonizing overcomplete local network predictions. In: Advances in Neural Information Processing Systems. Volume 29. (2016) 4
26. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: IEEE International Conference on Computer Vision. (2015) 2650–2658 4
27. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2002–2011 4
28. Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: IEEE Conference on Computer Vision and Pattern Recognition. (2019) 3268–3277 5
29. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. Volume 32. (2019) 10
30. Galliani, S., Lasinger, K., Schindler, K.: Massively parallel multiview stereopsis by surface normal diffusion. In: IEEE International Conference on Computer Vision. (2015) 873–881 10
31. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision, Springer (2016) 501–518 10
32. Yang, J., Mao, W., Alvarez, J.M., Liu, M.: Cost volume pyramid based depth inference for multi-view stereo. In: IEEE Conference on Computer Vision and Pattern Recognition. (2020) 4877–4886 10, 11, 13
33. Wei, Z., Zhu, Q., Min, C., Chen, Y., Wang, G.: Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 6187–6196 11

34. Ma, X., Gong, Y., Wang, Q., Huang, J., Chen, L., Yu, F.: Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 5732–5740 11