iMVS: Improving MVS Networks by Learning Depth Discontinuities

Nail Ibrahimli, Hugo Ledoux, Julian Kooij, Liangliang Nan
Delft University of Technology
{n.ibrahimli, h.ledoux, j.f.p.kooij, liangliang.nan}@tudelft.nl

Abstract

Existing learning-based multi-view stereo (MVS) techniques are effective in terms of completeness in reconstruction. We further improve these techniques by learning depth continuities. Our idea is to jointly estimate the depth and boundary maps. To this end, we introduce learning-based MVS strategies to improve the quality of depth maps via mixture density and depth discontinuity learning. We validate our idea and demonstrate that our strategies can be easily integrated into existing learning-based MVS pipelines where the reconstruction depends on high-quality depth map estimation. We also introduce a bimodal depth representation and a novel spatial regularization approach to the MVS networks. Extensive experiments on various datasets show that our method sets a new state of the art in terms of completeness and overall reconstruction quality. Experiments also demonstrate that the presented model and strategies have good generalization capabilities. The source code will be available soon.

1. Introduction

Multi-view stereo (MVS) techniques have been widely used to obtain dense 3D reconstruction from images. Traditional MVS techniques extract dense correspondences from multiple calibrated views and generate a dense 3D representation (i.e., point cloud or dense triangle mesh) of the scene. These methods rely on image correspondences in the RGB space, which are sensitive to textureless and non-Lambertian surfaces, and lighting variations.

Recent development in deep learning allows the use of learned feature maps instead of directly working on RGB images to build more robust MVS pipelines [6, 7, 14, 19, 25, 36, 39, 41, 42, 43, 45]. By learning feature maps about the objects in the scene, learning-based MVS methods have demonstrated better completeness than traditional methods in reconstructing man-made objects with low texture and non-Lambertian surfaces. The key idea of the most recent learning-based MVS methods is to learn the depth map from input images via variations of the 3D cost volume regularization [14, 42] or differentiable PatchMatch Stereo [10, 36].

Earlier works have shown that range images have small spatial variations except for sparse object discontinuities [16], and that higher-quality depth maps can be obtained through spatial regularization [2, 3]. These works inspire us to learn the geometric edges of the scene objects, instead of the photometric edges. Unlike photometric edges, geometric edges indicate where the object discontinuities truly lie (see Fig. 1). The second-order depth variations (i.e., variations of the changes in depth) are small in smooth surface regions and have large values at boundaries. This motivates us to estimate depth discontinuities by learning to detect the boundaries such that the second-order depth variation can be penalized only in the pixels that lie across non-boundary regions.

In the context of depth estimation, transitions between boundaries with depth discontinuities usually cause smoothing noise that is often referred to as the “bleeding artifacts” [35, 49]. This kind of noise can be alleviated by post-processing filters, which often damage the completeness of the reconstruction. Our insight into the “bleeding artifacts” issue is that it is due to building MVS pipelines on pixel-wise depth estimation, especially alongside wrongly inherited smooth surface assumptions around the true boundaries. As illustrated in Fig. 1, it can be ambiguous to determine if a pixel relates to a boundary with large depth discontinuities. We, therefore, learn to model the depth as univariate bimodal distribution instead of a single value estimate. In this work, we estimate bimodal depth
distribution instead of the single value estimate to reduce noise and improve accuracy without compromising completeness. We also jointly estimate the geometric boundary maps to get depth maps with smooth surfaces.

Our ideas can be easily integrated into existing learning-based MVS pipelines where a depth map is estimated. Extensive experiments on various benchmark datasets have demonstrated that our method advances the state of the art in terms of completeness and overall reconstruction quality. Besides, our method has high generalization capabilities, which have been validated by training our model on one dataset and testing it on other datasets. In summary, the contributions of this work are:

- a novel multi-task learning architecture for joint estimation of depth maps and object boundary maps.
- a bimodal depth representation that represents depth as a distribution, which can capture uncertainties in depth discontinuities.
- several new loss functions for depth discontinuity-based spatial regularization, which helps to learn discontinuities in depth to regularize the depth maps.

2. Related work

As learning-based MVS networks are built on top of photogrammetry-based MVS algorithms and developed from two-view methods, we review photogrammetry-based MVS algorithms, learning-based two view methods, and the recent development in learning-based MVS networks.

2.1. Photogrammetry-based MVS

Multi-View stereo methods purely built upon photogrammetry and multi-view geometry theory are usually referred to as traditional multi-view stereo methods. Janai et al. [18] showed that the taxonomy of the traditional multi-view stereo methods can be divided into four classes based on their representations of the scene and output. These scene representations are depth maps, point clouds, volumetric representations, and mesh or surfaces.

Volumetric representations use either discrete occupancy function [22], or levelset alike signed distance functions [11], which limits them to small scale reconstruction. The most common mesh-based approaches run variations of the marching cubes algorithm [24] on top of a signed distance function based on a volumetric surface representation [9].

The seminal point cloud-based method PMVS [12] has shown that starting with an initial sparse set of point features it is possible to create an initial set of patches and densify them by iterative greedy expansion and photo-geometric filtering. These methods usually demand a uniformly sampled sparse set of points across the image domain to be able to create point clouds with better completeness.

Depth map-based approaches usually first try to estimate a 2.5D depth map for each view. By using multi-view fusion pipelines [9, 46], these depth maps are consolidated into a single geometric model. Although the plane sweeping algorithm [8] has high memory consumption, it was the most commonly used technique for depth map estimation. To use plane sweeping stereo for a large dynamic range of outdoor videos, Pollefeys et al. [29] took advantage of GPS and inertia measurements to place the reconstructed models in geo-registered coordinates. Using random initialization and propagation techniques, the PatchMatch based MVS algorithms [13, 32] were able to estimate the depth map of each view with low memory consumption. In this work, we use a differentiable PatchMatch-based module to achieve a similar goal.

2.2. Learning-based two view methods

Learning-based two-view methods have introduced the initial building blocks for two-view stereo matching and depth estimation, which were later adapted for multi-view settings. The most common building blocks for learning-based depth map estimation pipelines are feature extraction and depth estimation from the feature space. Shared weight-based feature extraction was introduced by Zhbontar et al. [47], and later improved by using cost volume regularization for depth map extraction [5, 20, 40]. To reduce memory demand of the cost volume, Duggal et al. [10] introduced differentiable PatchMatch Stereo for two-view depth map estimation. These approaches were later adapted for multi-view settings via differentiable homography [5, 7, 14, 36, 42].

Tosi et al. [35] showed that it is possible to improve the quality of the learning-based two-view stereo networks by integrating an MLP-based bimodal mixture density network. In their work, they improved the accuracy of stereo matching networks [5, 40] that were used as a backbone to their mixture density head. Inspired by their work, we use the estimation of mixture density network parameters differently, i.e., we use it as an internal structure of the depth refinement network and input it with RGB-depth pairs instead of rectified image pairs. Our network learns depth maps alongside boundary maps. We use the same backbone to jointly estimate depth density parameters and boundary maps. In our pipeline, we use a 2D CNN-based U-Net [30] architecture to estimate the bimodal depth density parameters of each pixel in a discrete space.

2.3. Learning-based MVS

State-of-the-art learning-based MVS approaches adapt the photogrammetry-based MVS algorithms by implementing them as a set of differentiable operations defined in
the feature space. MVSNet [42] introduced good quality 3D reconstruction by regularizing the cost volume that was computed using differentiable homography on feature maps of the reference and source images. Its network architecture is similar to learning-based two-view stereo matching architecture GCNet [20]. Both MVSNet [42] and GCNet [20] regularize cost volume using a 3D CNN-based U-Net network. The cost volume itself has a very high demand for memory. To circumvent this problem, R-MVSNet uses GRUs [43] to regularize the cost volume sequentially. Follow-up works [14, 41], used feature pyramids and cost volume pyramids to learn in a coarse-to-fine manner instead of constructing a cost volume at a fixed resolution. To fully avoid construction of feature cost volume, Wang et al. [36] introduced a learning-based Multi-View PatchMatch Stereo pipeline. Variations of PatchMatch Stereo are seen as suitable options to work with high-resolution images since both traditional and learning-based Multi-View Patchmatch Stereo avoids the memory demands of Plane Sweep Stereo or feature cost volume regularization.

The recent work of PatchMatchNet [36] showed state-of-the-art results in terms of reconstruction completeness, which is used as a baseline in this work. We use differentiable PatchMatch-based MVS as part of the internal structure of our pipeline. To improve reconstruction quality, we estimate the geometric boundaries of the scene objects where depth discontinuities lie, and we present a method to regularize the depth map with an estimated boundary map. To our knowledge, our work is the first that uses mixture density networks in a learning-based multi-view stereo pipeline.

3. Method

In contrast to existing MVS approaches with depth map representations, in which the depth of each pixel is expressed as a single value, we introduce a bimodal depth representation that represents depth as distribution. Our depth map is thus not a common grid of per-pixel scalars, but per-pixel mixture density parameters. The motivation of this module is to implicitly integrate uncertainty notion into our pipeline, which can model and rectify the “over-smoothing” noise gathered in intra-object transitions, foreground-background transitions, and partial occlusions. This module also aims to learn depth discontinuities for spatial regularization of the depth map.

The overview of our proposed network architecture is shown in the Fig. 2. Our network has three parts, namely, feature extraction, coarse-to-fine PatchMatch Stereo, and depth discontinuity learning, detailed as follows.

3.1. Feature extraction

We follow the common practice of using feature pyramids to learn features in multiple scales [7, 14, 41], which also allows us to build our algorithm in a coarse-to-fine regression manner. We adapted the FPN [23] with residual connections between encoder and decoder, and use three layers of decoder outputs as our extracted features, with each next level having half the resolution of the previous one and the finest level having half dimension (width and height) of the original image. In Fig. 2, the red blocks show three scales of features fed to the coarse-to-fine PatchMatch Stereo module.

3.2. Coarse-to-fine PatchMatch Stereo

Following PatchmatchNet [36] that demonstrates good reconstruction completeness and low memory demands, we regress three levels of initial depth maps in a coarse-to-fine manner. We randomly initialize the depth values at the coarsest level, and at a finer level, we initialize the depth values with the outputs of the coarser levels. Following the initialization step, we run an iterative feedback loop between the propagation and evaluation steps. We propagate our estimates with good scoring values to the neighboring pixels. In the evaluation step, we use candidate depth values for differentiable homography warping and matching cost computation.

3.3. Depth discontinuity learning

The output of coarse-to-fine PatchMatch Stereo is a conventional depth map of half the resolution (half width and half height) of the original input. Hui et al. [17] showed that a low-resolution depth map can be progressively up-sampled with the guidance of the associated high-resolution color image. We follow this idea to bring the depth to the same resolution as the color image.

Unlike other learning-based networks [42, 44] that use residual network [15] to refine depth maps, we refine depth maps via learning mixture density parameters and geometric edge maps. Based on the fact that depth maps have piecewise smoothness and that they can be improved by spatial regularization to smooth regions as shown in earlier works [2, 3, 16], we propose to refine depth-map quality by learning depth discontinuities.

Previous methods based on pixel-wise single value estimates implicitly balance the depth estimation error between nearby foreground and background pixels for boundary points. This causes over-smoothing and bleeding artifacts across intra-object transitions, foreground-background transitions, and partial occlusions of the objects. To address these problems, our refinement network regresses the parameters of a bimodal distribution. We used the bimodal Laplacian distribution since it has a sharper shape modality than Gaussian and due to the fact that it optimizes over $L_1$ distance instead of $L_2$ distance between the groundtruth and estimated mean, to be robust against outliers. The bimodal
Laplacian density distribution can be written as

\[ p(x; \theta) = \frac{\alpha}{2\sigma_1} \exp\left(-\frac{|x-\mu_1|}{\sigma_1}\right) + \frac{1-\alpha}{2\sigma_2} \exp\left(-\frac{|x-\mu_2|}{\sigma_2}\right) \]

where \( \alpha \) is the mixture weight that can be seen as the likeness of each mode. Later in our work (see Sec. 4), we observe that the network learns to assign different \( \alpha \) values to different scene parts, and in most cases it is binary classifying foreground and background pixels. \( \mu_1 \) and \( \mu_2 \) are the two depth estimates of the corresponding modes. \( \sigma_1 \) and \( \sigma_2 \) are the two depth variance measures of each depth value. We also treat \( \frac{\alpha}{\sigma_1} \) and \( \frac{1-\alpha}{\sigma_2} \) as responsibility scores, which aims to determine which mode is responsible for the depth of a given pixel.

### 3.4. Depth fusion

For custom data without camera information, we run a standard Structure from Motion (SfM) pipeline [31] to obtain calibrated views. Our proposed architecture outputs the mixture density parameters, which can be used to compute pixel-wise depth values. We treat \( \frac{\alpha}{\sigma_1} \) and \( \frac{1-\alpha}{\sigma_2} \) as responsibility scores where \( \alpha, \sigma_1, \sigma_2 \) are estimated mixture parameters. These responsibility scores determine which mode is responsible for the depth of a given pixel. We thus create our final depth map by using the mean \( \mu \) of each pixel’s responsible mode. Following traditional MVS methods [13, 32] and learning-based MVS methods [36, 42] that run depth map fusion pipeline to integrate the 2.5D depth maps into a point cloud representation, we use the same fusion pipeline that does photometric and geometric consistency checks as in traditional method [28] before accepting a point. After photo-geometric filtering, we obtain a photogeometrically consistent depth map as shown in Fig. 2.

**3.5. Loss function**

Our loss function has four terms: Depth-groundtruth loss, Edge-depth loss, Smoothness loss, and Bimodal depth loss, each defined with a specific purpose.

- **Depth-groundtruth loss.** This loss term measures the difference in depth maps between prediction and the groundtruth. It is defined as the mean absolute error (MAE) of the estimated depth map, i.e., \( L_1 \) distance between the estimated depth and ground-truth depth across all stages of the PatchMatch Stereo and the final reconstructed depth,

\[
L_{gt} = \sum_{k=0}^{3} \left( \frac{1}{N_k} L_1(D_k, \hat{D}_k) \right),
\]

where \( k \in \{0, 1, 2, 3\} \) denotes the scale index of the coarse-to-fine PatchMatch stereo that estimates initial low resolution depth maps, with 0 representing the finest input and output resolution, and from 3 to 1 the coarser-to-finer scales of the PatchMatch stereo output. \( \hat{D}_k \) and \( D_k \) represent the ground-truth depth map and estimated depth map at resolution level \( k \), respectively. \( N_k \) represents a number of pixels in each scale.

- **Edge-depth loss.** Geometric edges or boundaries are expected where there are depth discontinuities in the depth map. Thus, the edge-depth loss term measures how much the estimated edges agree with the second-order depth variations (i.e., depth discontinuities). It is defined as the mean squared error (MSE) \( L_2 \) distance between the estimated edges and groundtruth changes of variations in depth,

\[
L_{ed} = \frac{1}{N} L_2(E, \phi(\Delta \hat{D}, \tau)),
\]

where \( \phi \) is the function that takes Laplacian of the depth and threshold value \( \tau \) to return the mask image where the Laplacian response of the depth map is higher than the \( \tau \). \( N \) is...
denotes the number of pixels. With this term, we explicitly inform the network that we are expecting geometric edges or boundaries at the pixels where there exist depth discontinuities. We calculate depth discontinuities using the Laplacian operator that is the second-order depth change.

**Smoothness loss.** Except for the geometric edges and boundaries with depth discontinuities, real-world objects typically demonstrate piecewise smoothing surfaces. Thus, we would like to encourage local smoothness for the regions without depth discontinuities. We achieve this by introducing an edge-aware smoothness loss term to penalize second-order depth variations in non-boundary regions,

\[
L_{sm} = \frac{1}{N} \sum_{i \in \Omega} \omega(E_i) |\Delta D_i|,
\]

where \( E_i \) will have a value close to 1 for boundaries and close to 0 for non-boundary pixels. \( \omega \) is a weight function that plays a role of a switch, which returns a value close to 0 for boundaries and close to 1 for non-boundary pixels. Thus, second-order depth change in non-boundary regions contributes to our smoothness loss. \( \beta \) is a tunable hyper-parameter that controls the sharpness of change in the \( \omega \) function. \( N \) denotes the number of the pixels in the image space \( \Omega \). To the best of our knowledge, this is the first time depth discontinuities are explicitly learned and used for spatial regularization in multi-view stereo networks.

**Bimodal loss.** We are interested in the estimation of parameters by maximizing the probability mixture density function given in Eq. (1). Specifically, maximizing the probability mixture density function will give us good candidates \( \mu_1 \) and \( \mu_2 \) for the depth map, and better strategy parameters \( \sigma_1 \), \( \sigma_2 \), and \( \alpha \) to choose between these bimodal depth means. By computing the negative-log-likelihood of the distribution, we can simply convert the probability maximization problem to a minimization problem. Our bimodal loss term is defined as

\[
L_{bi} = \frac{1}{N} \sum_{i \in \Omega} - \log(p(\hat{D}_i; \theta, i)),
\]

where \( \hat{D}_i \) represent the groundtruth depth measured at pixel \( i \), and \( \theta \) is the parameter of the bimodal distribution introduced in Eq. (1). This loss term tries to optimize the parameters of bimodal mixture distribution to better describe the groundtruth data.

**Total loss.** We simply use the weighted sum of the aforementioned loss terms

\[
L_{total} = L_{gt} + \lambda_1 L_{ed} + \lambda_2 L_{sm} + \lambda_3 L_{bi}
\]

as a training criterion for our network to optimize the parameters via backpropagation. \( \lambda_1 \), \( \lambda_2 \), and \( \lambda_3 \) are hyper-parameters empirically set based on our experiments on the validation set.

4. Experiments and Evaluation

We have tested and evaluated our method on multiple datasets: both small baseline dataset DTU [1] and large baseline dataset “Tanks and Temples” [21] and ETH3D [33]. We used the same model to quantitatively evaluate the generalization capabilities of our method and to compare it with other methods. All the metric results of the other methods were collected from the corresponding papers, and the 3D point clouds of other papers were reconstructed using the code and pre-trained models provided by the authors. For a fair comparison, all methods were trained using the same dataset.

Fig. 3 demonstrates some of our visual results (i.e., point clouds) and the visual comparison with the baseline method PatchmatchNet [36]. We can observe that our results show better completeness and are cleaner, especially for the indoor scenes with low textures.

4.1. Evaluation on DTU dataset

The DTU dataset [1] is a benchmark with 124 scenes captured by a structured-light sensor under seven different lighting conditions. It has been widely used for developing learning-based MVS methods and evaluating their performance in terms of completeness and accuracy. Following this benchmark, we recorded the accuracy, completeness, and overall performance metrics of our method, and compared them to other methods. The accuracy is measured as the mean error distance between the closest points from the reconstruction to the structured-light reference. completeness is measured as the mean error distance between the closest points from the reference to the reconstruction, and overall is the algebraic mean of accuracy and completeness. Therefore having lower metric scores is better for this benchmark.

The result on the DTU dataset is reported in Tab. 1. From the result, we can see that traditional photogrammetry-based methods generally have better accuracy, while learning-based methods have better completeness and overall performance. Furthermore, it also reveals that the completeness gap between learning-based and photogrammetry-based methods are bigger than their gap in accuracy, which motivated us to use a coarse-to-fine PatchMatch Stereo to build our initial depth estimation block, to reduce the accuracy gap while still improving completeness. Compared with the state-of-the-art leaning-based MVS methods, our method demonstrates better completeness and overall scores. This reveals that learning depth discontinuities is an effective means to improve both reconstruction accuracy and completeness.

4.2. Evaluation on “Tanks and Temples” dataset

“Tanks and Temples” is a real-world large-scale dataset consisting of both indoor and outdoor scenes [21]. It has
two parts: an intermediate set consisting of images of sculptures, large vehicles, and house-scale buildings (taken from the exterior), and an advanced set consisting of images of large indoor scenes and large outdoor scenes with complex geometric layouts and repetitive structures. This benchmark has three metrics, namely, recall, precision, and F-score. Recall and precision represent the completeness and accuracy of the reconstruction, respectively, both measured in percentage (%). The F-score combines precision and recall, and it is defined as the harmonic mean of a model’s precision and recall.

In our experiments, we used our model trained using the DTU dataset with 14 epochs with all the proposed loss terms. We compared the results against those from our baseline method PatchmatchNet [36]. For both methods, we ran the same depth map fusion algorithm with the same threshold value to not gain any advantage in the evaluation process. As can be seen from the statistics reported in Tab. 3, our results on the intermediate set have better performance on all evaluation metrics. On the advanced set, our results demonstrate better accuracy and F-score, and the results from PatchmatchNet have slightly better completeness. For
4.3. Evaluation on ETH3D dataset

The ETH3D benchmark [33] consists of high-resolution images of scenes with sparse scene coverage, high viewpoint variation, and camera parameter information. The quantitative evaluation of our method and the comparison with PatchmatchNet [36] on the ETH3D dataset [33] are detailed in Tab. 2. Both methods have used the same fusion pipeline. Our method demonstrates better accuracy and F-score, while PatchmatchNet has better completeness.

4.4. Ablation study

We have conducted an ablation study to understand and analyze the contributions of the aforementioned loss terms of our architecture. Since the edge-depth loss and the smoothness loss terms together strive for edge-aware smoothness, we do not separate them in our experiments. We retrieve the last two metrics from the validation set while tuning our hyper-parameters. “Depth map error” represents the accuracy of the estimated depth map, calculated using mean absolute error (MAE) between the estimated depth map and groundtruth. “Error > 8 mm” represents the percentage of points in the depth map having the higher error than 8 mm.

The results are detailed in Tab. 4. We can see that using all lost terms improves the depth map quality on the validation set. For testing, we observe from Tab. 4 that our point clouds have better completeness and overall metrics with bimodal and depth ground-truth loss while having edge-aware smoothness term results in better accuracy. Our network also improves both accuracy and completeness if we compare it against the baseline.

4.5. Effectiveness of depth discontinuity learning

From the above experiments and evaluation, our method demonstrates superior reconstruction quality in terms of completeness and overall quality, which benefits from our depth discontinuity learning. To understand the role of depth discontinuity learning in reconstruction, we visualize the learned depth discontinuities (denoted as edge maps) for a few randomly picked examples in Fig. 4, and compare them with the edge maps predicted using the seminal learning-based edge detection method HED [38]. We can see that by learning depth discontinuities, our network can retrieve edges where the true depth discontinuities lie. Thus, as a key component for learning-based MVS pipelines, our discontinuity-aware depth learning is more robust to photometrical changes, shadows, and small variations in depth.

To reveal how our depth discontinuity learning works, we also demonstrated the \( \alpha \) map of each example in Fig. 4 (d), where \( \alpha \) is the mixture weight in the bimodal Laplacian density distribution (see Eq. (1)). It is surprisingly interesting to observe that our network tries to learn to differentiate foreground and background. As shown in Fig. 4 (d), \( \alpha \) expresses itself as binary classification for foreground and background pixels.

4.6. Limitations

Although our method has good completeness and an overall score (see Tab. 1), it has still not reached the accuracy level of traditional photogrammetry-based algorithms.

| Method         | Accuracy (mm) | Completeness (mm) | Overall (mm) |
|----------------|---------------|-------------------|--------------|
| Ours (\( L_1, 4 \)) | 0.283         | 0.873             | 0.578        |
| Ours (\( L_{1,2,3,4} \)) | 0.399         | 0.280             | 0.339        |

Table 1. Quantitative comparison with photogrammetry-based and learning-based MVS methods, on the DTU dataset [1]. Two different settings (with different loss functions) of our method were tested. \( L_1 \): depth-groundtruth loss; \( L_2 \): edge-depth loss; \( L_3 \): smoothness loss; \( L_4 \): bimodal loss. Please note that the metrics are error-based and thus the smaller the better.

| Method         | Accuracy (%) | Completeness (%) | F-score |
|----------------|--------------|------------------|---------|
| PatchmatchNet  | 64.81        | 65.43            | 64.21   |
| Ours           | 64.96        | 65.21            | 64.37   |

Table 2. Quantitative evaluation of our method and comparison with PatchmatchNet [36] on the ETH3D training set [33]. Following the benchmark, the accuracy and completeness measures are quantified using the percentage of points below a 2 cm error margin (the higher the better).
### Table 3. Evaluation and comparison with PatchmatchNet [36] on the “Tanks and Temples” dataset [21].

| Methods          | Intermediate set | Advanced set |  
|------------------|------------------|--------------|
|                  | Precision (%)    | Recall (%)   | F-score     | Precision (%)    | Recall (%)   | F-score     |
| PatchmatchNet    | 43.64            | 69.38        | 53.15       | 27.27            | 41.66        | 32.31       |
| Ours             | **45.12**        | **69.69**    | **54.30**   | **28.31**        | **41.06**    | **32.80**   |

### Table 4. Ablation study on the point clouds and depth maps from the DTU dataset [1].

| Methods          | Point clouds (testing) | Depth maps (validation) |  
|------------------|------------------------|-------------------------|
|                  | Accuracy (mm) | Completeness (mm) | Overall (mm) | Depth map error (mm) | Error ratio (%; on points with error > 8 mm) |
| PatchmatchNet    | 0.427         | 0.277             | 0.352        | 7.33         | 11.68       |
| Architecture + $L_1$ | 0.412      | 0.273             | 0.342        | 5.41         | 9.07        |
| Architecture + $L_{1,2,3}$ | 0.412     | 0.270             | 0.341        | 5.44         | 8.96        |
| Architecture + $L_{1,4}$      | 0.405      | **0.267**         | **0.336**    | 5.47         | 9.01        |
| Architecture + $L_{1,2,3,4}$  | **0.399**  | 0.280             | 0.339        | **5.31**     | **8.91**    |

### Figure 4. Edges maps of a few randomly chosen examples. For each example, the images from left to right are: (a) color image; (b) edge map predicted by HED [38]; (c) our learned edge map; (d) the $\alpha$ map. Our learned edge maps better capture the depth discontinuities, regardless of the photometric changes. It is also interesting to observe that our $\alpha$ maps distinguish between foreground and background.

such as Gipuma [13], which is a common weakness of recently developed learning-based MVS methods. There is usually a trade-off between accuracy and completeness since increasing completeness also means increasing the potential source of the noise. It is also worth noting that in this work we have used the same fusion pipeline as in other papers [36, 42]. This fusion pipeline with our model depth estimates tends to create excessively dense point clouds (especially when the input is a large number of high-resolution images), which often need to be downsampled in down-stream applications such as surface reconstruction. Increasing the accuracy and keeping the completeness by compromising point density is certainly favorable.

### 5. Conclusion

We have presented a strategy for improving MVS networks by learning depth discontinuities. Our experiments have shown that learning depth maps implicitly as a mixture distribution can improve reconstruction quality. Learning depth discontinuities and integrating them to the network as a means of prior knowledge for piecewise smoothness regularization improves the accuracy alongside keeping the good completeness of the final reconstruction. On large datasets with good view coverage, we have noticed that our network excels to generate more points that successfully pass the geometric and photometric filters of the fusion algorithm, which contributes to better completeness and overall better geometry. As future work, we plan to investigate the foreground and background transitions and learn these transitions via the means of probabilistic framework and multi-modal distribution.
References

[1] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016.

[2] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *Proceedings 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Cat. No.98CB36231), pages 648–655, 1998. doi: 10.1109/CVPR.1998.698673.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[4] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In D. Forsyth, P. Torr, and A. Zisserman, editors, *Computer Vision – ECCV 2008*, pages 766–779, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-88682-2.

[5] J.-R. Chang and Y.-S. Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[6] R. Chen, S. Han, J. Xu, and H. Su. Point-based multi-view stereo network. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[7] S. Cheng, Z. Xu, S. Zhu, Z. Li, L. E. Li, R. Ramamoorthi, and H. 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.

[8] R. T. Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 358–363. IEEE, 1996.

[9] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312. IEEE, 1996.

[10] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun. Deep-pruner: Learning efficient stereo matching via differentiable patchmatch. In *ICCV*, 2019.

[11] O. Faugeras and R. Keriven. *Variational principles, surface evolution, PDE's, level set methods and the stereo problem*. IEEE, 2002.

[12] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *IEEE TPAMI*, 32(8):1362–1376, 2010.

[13] S. Galliani, K. Lasinger, and K. Schindler. Massively parallel multiview stereopsis by surface normal diffusion. June 2015.

[14] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. 2019.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] J. Huang, A. Lee, and D. Mumford. Statistics of range images. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, volume 1, pages 324–331 vol.1, 2000. doi: 10.1109/CVPR.2000.855836.

[17] T.-W. Hui, C. C. Loy, and X. Tang. Depth map supersolution by deep multi-scale guidance. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[18] J. Janai, F. Güney, A. Behl, A. Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020.

[19] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang. SurfaceNet: an end-to-end 3d neural network for multiview stereopsis. In *ICCV*, 2017.

[20] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

[21] A. Knäpitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[22] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International journal of computer vision*, 38(3):199–218, 2000.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[24] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[25] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10451–10460, 2019. doi: 10.1109/ICCV.2019.01055.

[26] K. Luo, T. Guan, L. Ju, Y. Wang, Z. Chen, and Y. Luo. Attention-aware multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
[27] X. Ma, Y. Gong, Q. Wang, J. Huang, L. Chen, and F. Yu. Ep-p-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5732–5740, 2021. 7

[28] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nistér, and M. Pollefeys. Real-time visibility-based fusion of depth maps. In 2007 IEEE 11th International Conference on Computer Vision, pages 1–8. IEEE. 2007. 4

[29] M. Pollefeys, D. Nistér, J.-M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S.-J. Kim, P. Merrell. et al. Detailed real-time urban 3d reconstruction from video. International Journal of Computer Vision, 78 (2):143–167, 2008. 2

[30] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. arxiv 2015. arXiv preprint arXiv:1505.04597, 2019. 2

[31] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 4

[32] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In European Conference on Computer Vision (ECCV), 2016. 2, 4

[33] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. A multi-view stereo benchmark with high-resolution images and multicamera videos. In Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 5, 7

[34] E. Tola, C. Strecha, and P. Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. 23(5):903–920, 2012. 7

[35] F. Tosi, Y. Liao, C. Schmitt, and A. Geiger. Smd-nets: Stereo mixture density networks. In Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 1, 2

[36] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[37] Z. Wei, Q. Zhu, C. Min, Y. Chen, and G. Wang. Aa-rmsvnet: Adaptive aggregation recurrent multi-view stereo network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6187–6196, 2021. 7

[38] S. Xie and Z. Tu. Holistically-nested edge detection. In Proceedings of the IEEE international conference on computer vision, pages 1395–1403, 2015. 7, 8

[39] Q. Xu and W. Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. CoRR, abs/1912.11746, 2019. URL http://arxiv.org/abs/1912.11746. 1, 7

[40] G. Yang, J. Manela, M. Happold, and D. Ramanan. Hierarchical deep stereo matching on high-resolution images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5515–5524, 2019. 2

[41] J. Yang, W. Mao, J. M. Alvarez, and M. Liu. Cost volume pyramid based depth inference for multi-view stereo. In The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. 1, 3, 7

[42] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan. Mvsnet: Depth inference for unstructured multi-view stereo. European Conference on Computer Vision (ECCV), 2018. 1, 2, 3, 4, 7, 8

[43] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. Computer Vision and Pattern Recognition (CVPR), 2019. 1, 3, 7

[44] A. Yu, W. Guo, B. Liu, X. Chen, X. Wang, X. Cao, and B. Jiang. Attention aware cost volume pyramid based multi-view stereo network for 3d reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing, 175:448–460, 2021.

[45] Z. Yu and S. Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In CVPR, 2020. 1, 7

[46] C. Zach, T. Pock, and H. Bischof. A globally optimal algorithm for robust tv-l1 range image integration. In 2007 IEEE 11th International Conference on Computer Vision, pages 1–8. IEEE. 2007. 2

[47] J. Zbontar, Y. LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. J. Mach. Learn. Res., 17(1):2287–2318, 2016. 2

[48] J. Zhang, Y. Yao, S. Li, Z. Luo, and T. Fang. Visibility-aware multi-view stereo network. British Machine Vision Conference (BMVC), 2020. 7

[49] S. Zhu, G. Brazil, and X. Liu. The edge of depth: Explicit constraints between segmentation and depth. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13116–13125, 2020. 1