Code available at:

https://github.com/mirmix/ddlmvs

Article

DDL-MVS: Depth Discontinuity Learning for Multi-View Stereo Networks

Nail Ibrahimli, Hugo Ledoux, Julian F. P. Kooij, and Liangliang Nan

Delft University of Technology {n.ibrahimli, h.ledoux, j.f.p.kooij, liangliang.nan}@tudelft.nl

Abstract: We propose an enhancement module called depth discontinuity learning (DDL) for learning-based multi-view stereo (MVS) methods. Traditional methods are known for their accuracy but struggle with completeness. While recent learning-based methods have improved completeness at the cost of accuracy, our DDL approach aims to improve accuracy while retaining completeness in the reconstruction process. To achieve this, we introduce the joint estimation of depth and boundary maps, where the boundary maps are explicitly utilized for further refinement of the depth maps. We validate our idea by integrating it into an existing learning-based MVS pipeline where the reconstruction depends on high-quality depth map estimation. Extensive experiments on various datasets, namely DTU, ETH3D, "Tanks and Temples" and BlendedMVS, show that our method improves reconstruction quality compared to our baseline, Patchmatchnet. Our ablation study demonstrates that incorporating the proposed DDL significantly reduces the depth map error, for instance by more than 30% on the DTU dataset, and leads to improved depth map quality in both smooth and boundary regions. Additionally, our qualitative analysis has shown that the reconstructed point cloud exhibits enhanced quality without any significant compromise on completeness. Finally, the experiments reveal that our proposed model and strategies exhibit strong generalization capabilities across the various datasets.

Keywords: Multi-view Stereo, 3D Reconstruction, Depth Map Refinement, Depth Boundary Estimation

Citation: Ibrahimli, N.; Ledoux, H.; Kooij, J.F.P.; Nan, L. DDL-MVS: Depth Discontinuity Learning for Multi-View Stereo Networks. *Remote Sens.* **2022**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

1. Introduction

Multi-view stereo (MVS) techniques have been widely used to obtain dense 3D reconstruction from images. MVS allows aerial images to be converted into accurate 3D models, which provide a more comprehensive representation of the large scene. This 3D information can be used for various applications, such as digital surface modelling [1], landform analysis [2], and urban planning [3]. It provides valuable insights into the shape, structure, and topography of the scene, enabling better understanding and interpretation of remote sensing data.

Traditional MVS techniques [4–6] 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 developments in deep learning allow the use of learned feature maps instead of directly working on RGB images to build more robust MVS pipelines [7–17]. 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. Recent learning-based MVS methods learn to reconstruct the depth map from input images by regularizing the 3D cost volume [7,13] or by Patchmatch-based iterative optimization [17,18]. Still, depth estimation remains challenging, and depth discontinuities at transitions between object boundaries are usually erroneous [19,20]. While this kind of error can be alleviated by post-processing filters, it often reduces the completeness of the reconstruction.

In MVS pipelines, it is common for a single depth value to be estimated per pixel, accompanied by a smooth surface assumption. This spatial regularization technique results in higher-quality depth maps, as shown in previous studies such as [21,22], which in turn improves the completeness of the reconstructed 3D model. However, a limitation of this approach is that it tends to oversmooth the true depth continuities at object boundaries, as pointed out in recent works such as [20,23]. Furthermore,

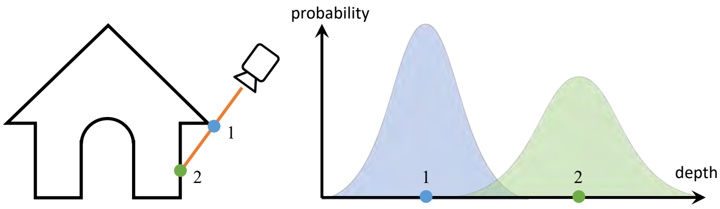


Figure 1. We propose to estimate depth as a bimodal univariate distribution. Using this depth representation, we improve multi-view depth reconstruction, especially across geometric boundaries.

as illustrated in Fig.1, pixels near depth discontinuities can pose ambiguity in determining which side of the depth boundary they belong to.

These findings motivate us to pursue two complementary objectives. First, we aim to explicitly detect the geometric edges, instead of relying solely on photometric edges that capture color and texture changes [21,22]. Geometric edges more accurately indicate the true locations of object boundaries than photometric edges (see Fig. 1). We propose to estimate geometric boundary maps jointly with the depth maps, such that smooth depth surfaces can be enforced while considering the local geometry. Second, as shown in Fig. 1, we propose to estimate per-pixel depth as a univariate bimodal distribution rather than as a single depth value. This allows us to explicitly represent the depth ambiguity and avoids over-smoothing the depth discontinuities. We integrate both objectives into a multi-task learning architecture to improve the depth accuracy while avoiding the completeness trade-off of previous approaches.

To confirm the validity of our idea, we integrate it into the existing learning-based Multi-View Stereo (MVS) pipeline. Extensive experiments that we ran on various benchmark datasets (see Sect. 4) demonstrate that our method obtains better results.[1] Moreover, 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 to multi-view stereo networks are: (1) a novel multi-task learning architecture for joint estimation of depth maps and object boundary maps for learning-based multi-view stereo pipelines; (2) a bimodal depth representation that represents depth as a distribution learned from multi-view images; (3) a general loss formulation for depth discontinuity-based spatial regularization, which helps to learn discontinuities in depth and to regularize the depth maps.

The structure of this article is as follows: In Section 2, we will review the existing literature and contrast the most related works to our approach. Section 3 will delve into our methodology and outline the MVS pipeline we use to test our approach. Section 4 will present and discuss the experimental results obtained from our study. Finally, in Section 5, we will conclude the paper by summarizing our main findings.

2. Related work

As learning-based MVS networks are inspired by 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. [24] 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 [25] or levelset alike signed distance functions [26], which limits them to small-scale reconstruction. The most common mesh-

[1] The code is available at https://github.com/mirmix/ddlmvs

based approaches run variations of the marching cubes algorithm [27] on top of a signed distance function based on a volumetric surface representation [28].

The seminal point cloud-based method by Furukawa et.al. [4] 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 [28,29], these depth maps are consolidated into a single geometric model. Although the plane sweeping algorithm [30] 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. [31] 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 [5,32] 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 [33], and later improved by using cost volume regularization for depth map extraction [34–36]. To reduce memory demand of the cost volume, Duggal et al. [18] introduced differentiable PatchMatch Stereo (PMS) for two-view depth map estimation. These approaches were later adapted for multi-view settings via differentiable homography [7,12,13,17,35].

EdgeStereo [37] uses a pre-trained sub-network for detecting the edges, and the edge cues are then fed into the disparity branch to improve the disparity map. Tosi et al. [20] 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 [35,36] that were used as a backbone to their mixture density head. Inspired by these works, we also represent depth as bimodal distribution, and we jointly estimate depth maps and object boundary maps in the multi-view stereo setting using a novel multi-task learning architecture. Our pipeline does not involve any parallel (sub)networks and learns directly from multi-view images to estimate edge-depth pairs jointly.

The continuous disparity network [23] aims to regress the multi-modal depth by jointly estimating both probability and offset volume by minimizing a Wasserstein distance between the ground truth and the distribution estimated from the volumes. The offset volume aims to obtain continuous disparity estimations. Our method avoids regressing the offset values and instead, directly estimates bimodal distribution parameters.

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 [7] 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 the learning-based two-view stereo matching architecture GCNet [34]. Both MVSNet [7] and GCNet [34] regularize cost volume using a 3D CNN-based U-Net. The cost volume itself has a very high demand for memory. To circumvent this problem, R-MVSNet uses GRUs [8] to regularize the cost volume sequentially. Follow-up works [13,16], 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 the construction of feature cost volume, Wang et. al. [17] introduced a learning-based Multi-View PMS pipeline. Variations of PMS are seen as suitable options to work with high-resolution images since both traditional and learning-based Multi-View PMS avoids the memory demands of Plane Sweep Stereo or feature cost volume regularization.

In contrast to two-view or multi-view Plane Sweep stereo [28,29] and cost volume regularization methods [7,34], to reduce memory consumption our pipeline estimates depth maps by fully avoiding the cost volume creation and usage of 3D CNN networks. For this, we are leveraging differentiable PatchMatch-based Multi-View Stereo as part of the internal structure of our pipeline [5,17]. The recent work of PatchMatchNet [17] showed state-of-the-art results in terms of reconstruction completeness, which is used as a baseline in this work.

To enhance the quality of scene reconstruction, our proposed method focuses on estimating the geometric boundaries of objects in the scene where depth discontinuities occur. We introduce a technique to regularize the depth map by incorporating an estimated boundary map. Our approach distinguishes itself from DEF-MVSNet [38] in terms of how edge information is represented and modeled. While DEF-MVSNet primarily focuses on determining flow directions as pixel offsets, our method explicitly learns and smooths the edge map by defining each pixel as a bimodal distribution. This distinction contributes to the unique characteristics of our approach.

Similarly, our method deviates from BDE-MVSNet [39], which also aims to find flow directions for edge pixels using gradient information. Instead, we explicitly learn the boundary map, placing emphasis on regularizing smoothness in regions that are not classified as boundaries. In comparison to ElasticMVS [40] which proposes an elastic part representation for encoding physically connected part segmentations, our approach focuses solely on explicitly learning the boundary map. By utilizing the boundary map for regularization, our objective is to enhance smoothness rather than encode physically connected part segmentations and capture surface connectedness and boundaries within the image.

During the development of our method, we also explored some depth derivative-based loss functions, similar to those utilized in previous works [37,41]. However, we did not observe significant improvements when employing these loss functions. Therefore, we adopted a different approach by explicitly learning the boundary map to regulate smoothness in regions that are not classified as boundaries.

In comparison to two-view stereo matching pipeline SMD-Nets [20], we employ a mixture density network as an internal structure for depth refinement, inputting it with RGB-Depth pairs instead of rectified left-right image pairs. Unlike previous methods, we learn the depth and boundary map simultaneously, utilizing the same backbone architecture for estimating the density parameters and boundary map in parallel. In comparison with previous methods, our pipeline utilizes a 2D CNN-based U-Net architecture [42] to estimate the bimodal depth density parameters for each pixel in discrete space.

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, our approach takes advantage of 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 the uncertainty notion into our pipeline, which enables us to learn depth discontinuities for spatial regularization of the depth map and to further alleviate the noise gathered in intra-object transitions, foreground-background transitions, and partial occlusions.

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

3.1. Feature extraction

We employ a widely used technique in the Computer Vision field known as feature pyramid learning [12,13,16], which enables us to build our algorithm in a coarse-to-fine regression fashion. We adopted the Feature Pyramid Networks [43] with residual connections between encoder and decoder, and use three layers of decoder outputs as our extracted features. Each subsequent level has half the resolution of the level before it, and the finest level has half the width and height of the

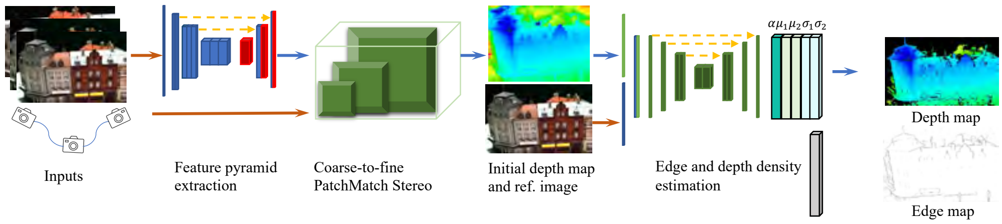


Figure 2. An overview of the proposed multi-view depth discontinuity learning network that outputs depth and edge information for each pixel. The brown arrows represent input feed and the blue arrows represent pipeline flow. We first extract multi-scale features from color images with FPN [43] alike auto-encoder. Then we feed extracted features and camera parameters to the coarse-to-fine PMS module to extract the initial depth map. Using the initial depth map and RGB pair, our network learns bimodal depth parameters and geometric edge maps. We use mixture parameters and photo-geometric filtering to compute our final depth map. The edge map visualized here is negated edge map (for a clear view).

original image. In Fig. 2, the red blocks show three scales of features fed to the coarse-to-fine PMS module.

3.2. Coarse-to-fine PMS

Being agnostic to the backbone our method is independent of the underlying rough depth estimation method. Both cost volume regularization and the PatchMatch-based approach can be used for depth estimation. We follow PatchmatchNet [17] that demonstrates good reconstruction completeness and low memory demands. Our pipeline 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 PMS is a conventional depth map of half the resolution (half width and half height) of the original input. Hui et al. [44] showed that a low-resolution depth map can be progressively upsampled with the guidance of the associated high-resolution color image. This idea inspires the proposed framework's attempt to match the resolution of the color and depth images.

Contrary to existing learning-based networks [7,45], which revise depth maps using residual networks, we refine depth maps via learning mixture density parameters and geometric edge maps. In contrast to SMD-Nets [20], which employ corrected image pairs as input, we use RGB-depth pairs as the input to the depth refinement network and convolutional mixed density networks as the internal structure. To the best of our knowledge, our work is the first learning-based MVS method that explicitly learns depth discontinuity maps (aka geometric edge maps) to simultaneously refine the quality and improve the smoothness of the depth maps.

In our pipeline, we use a 2D CNN-based U-Net [42] architecture to estimate the bimodal depth density parameters of each pixel in a discrete space. During the development of our pipeline, we experimented with different network variations to learn separate boundary maps and mixture density parameters, including two parallel network streams and single encoder and multiple decoder architectures. However, we found that using multiple subnetworks increased the number of parameters and GPU memory demands without leading to any substantial improvement in results. Therefore, we chose to use a single encoder and decoder architecture for our proposed pipeline. 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 [21,22,46], 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. Our refinement network regresses the parameters of a bimodal distribution. We use the bimodal Laplacian distribution, which was inspired by Tosi et al. [20] work. During development, we observed that the Laplacian distribution [47] had slightly better results than the Gaussian. The Laplacian distribution has a sharper shape modality than Gaussian. It optimizes over $\mathcal{L}_1$ distance instead of $\mathcal{L}_2$ distance between the groundtruth and estimated mean. This makes it more robust against outliers. The bimodal Laplacian density distribution can be written as

$$\theta = \{\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2\} \tag{1}$$

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

where α is the mixture weight that can be seen as the likeliness of each mode. Later in our work (see Sec. 4), we observe that the network learns to assign different α values to different scene parts, and in most cases it is binary classifying foreground and background pixels. μ_1 and μ_2 are the two depth estimates of the corresponding modes. σ_1 and σ_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 the responsible mode for the depth of a given pixel.

Besides extending bimodal depth estimation to the multi-view case, our proposed convolutional mixture density network also shows that with a single stream compact discontinuity learning network architecture, it is possible to achieve three goals: (1) Upsampling; (2) Refining; (3) Multi-task learning.

3.4. 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., $\mathcal{L}_1$ distance between the estimated depth and ground-truth depth across all stages of the PMS and the final reconstructed depth,

$$L_{gt} = \sum_{k=0}^{3}[\frac{1}{N_k}\mathcal{L}_1(D_k, \hat{D}_k)], \tag{2}$$

where $k \in \{0, 1, 2, 3\}$ denotes the scale index of the coarse-to-fine PMS 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 PMS output. $\hat{D}_k$ and D_k represent the ground-truth depth map and estimated depth map at resolution level k, respectively. The DTU dataset [48] contains masks that identify pixels with valid ground truth depth information. N_k represents the 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) ($\mathcal{L}_2$ distance) between the estimated edge E and groundtruth changes of variations in depth $\hat{D}$,

$$L_{ed} = \frac{1}{N}\mathcal{L}_2(E, \phi(\Delta\hat{D}, \tau)), \tag{3}$$

where ϕ is the function that takes Laplacian of the depth and threshold value τ to return the mask image where the Laplacian response [49] of the depth map is higher than the τ. The DTU dataset [48] contains masks that identify pixels with valid ground truth depth information. The variable N represents the count of masked pixels that have corresponding ground truth labels for depth. 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, which 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,

$$\begin{aligned} L_{sm} &= \frac{1}{N}\sum_{i\in\Omega}\omega(E_i)|\Delta D_i| \\ \omega(E_i) &= \exp(-\beta E_i) \end{aligned}, \tag{4}$$

where E_i will have an estimated value close to 1 for boundaries and close to 0 for non-boundary pixels. ω 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. β is a tunable hyper-parameter that controls the sharpness of change in the ω function. N denotes the number of pixels in the image space Ω with a valid grountruth depth. 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 adopt a common approach of minimizing the negative-log likelihood of the distribution to increase the likelihood of true depth. Tosi et al. [20] have demonstrated that this loss term for bimodal depth in the two-view stereo setting can produce inspiring results. Our bimodal loss term is defined as

$$L_{bi} = \frac{1}{N}\sum_{i\in\Omega} -\log(p(\hat{D}_i;\theta,i)), \tag{5}$$

where $\hat{D}_i$ represents the groundtruth depth measured at pixel i, and θ is the parameter of the bimodal distribution introduced in Eq. 1. The distribution p can be computed using the Eq. 1. N denotes the number of the pixels in the image space Ω with a valid grountruth depth.

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

$$L_{total} = L_{gt} + \lambda_1 L_{ed} + \lambda_2 L_{sm} + \lambda_3 L_{bi} \tag{6}$$

as a training criterion for our network to optimize the parameters via backpropagation. $\lambda_1 = 4$, $\lambda_2 = 1.25$, and $\lambda_3 = 0.5$ are hyper-parameters empirically set based on our experiments on the validation set.

4. Experiments and Evaluation

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.

4.1. Datasets

We have tested and evaluated our method on multiple datasets: the small baseline dataset DTU [48], and the large baseline datasets "Tanks and Temples" [50], ETH3D [51] and Blended-MVS [52].

The DTU dataset [48] is a benchmark with 120 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. All the learning-based methods in Tab. 1 are trained on the same 79 scans, validated on the same 18 scans, and evaluated on the remaining 22 scans. [2]

"Tanks and Temples" is a real-world large-scale dataset consisting of both indoor and outdoor scenes [50]. It has two parts: an intermediate set consisting of images of sculptures, large vehicles,

[2] Validation set: scans 3, 5, 17, 21, 28, 35, 37, 38, 40, 43, 56, 59, 66, 67, 82, 86, 106, 117.
Evaluation set: scans 1, 4, 9, 10, 11, 12, 13, 15, 23, 24, 29, 32, 33, 34, 48, 49, 62, 75, 77, 110, 114, 118.

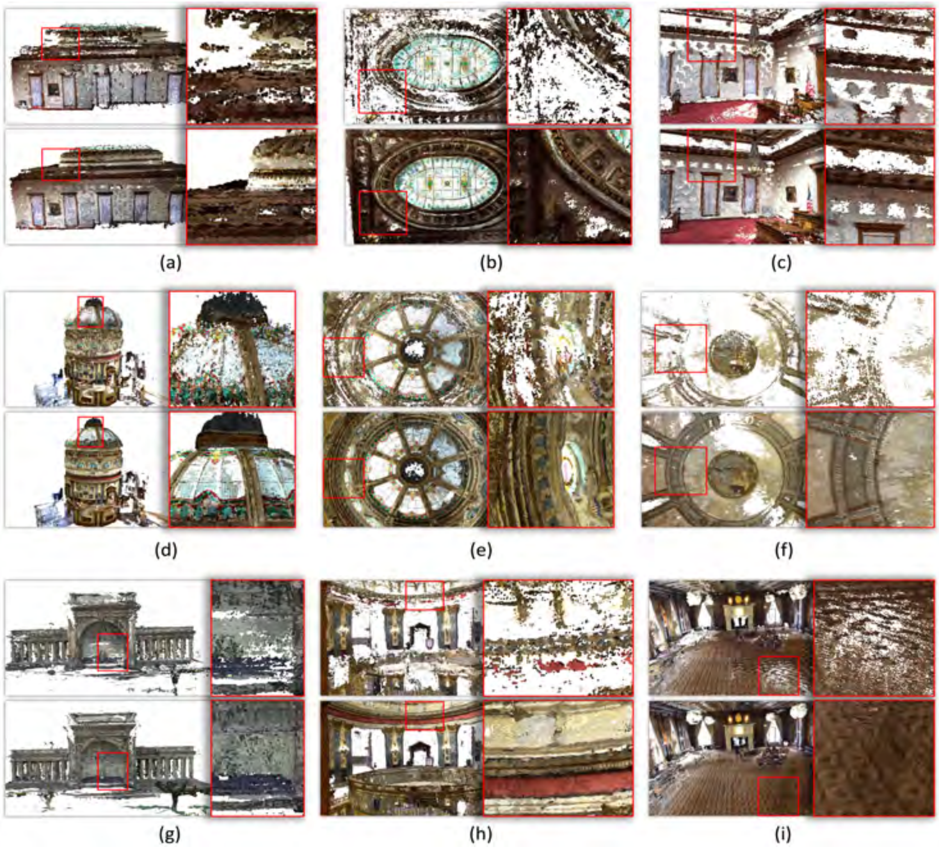


Figure 3. Comparison between our method and the baseline method PatchmatchNet [17] on a set of scenes from the Tank and Temples dataset [50]. For each scene, the top row shows the results from PatchmatchNet, and the bottom row shows the results from our method. A zoomed view of the marked image region is shown on the right of each result.

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.

The ETH3D dataset [51] is a collection of calibrated images, containing various indoor and outdoor environments, including urban scenes, garages, and rooms. The dataset provides ground truth camera poses, 3D point cloud geometry, and images for each scene, making it suitable for tasks such as camera pose estimation and 3D reconstruction.

BlendedMVS [52] is a large-scale MVS dataset for generalized multi-view stereo networks. The dataset contains samples covering a variety of scenes, including architecture, sculptures, aerial images, and small objects.

4.2. Evaluation on DTU dataset

In this section, we present our findings based on the DTU benchmark [48], where we evaluated the performance of our method using the *accuracy*, *completeness*, and *overall* metrics. The *accuracy* metric measures the mean error distance between the closest points in the reconstruction and the reference based on structured light. The *completeness* metric quantifies the mean error distance between the closest points in the reference and the reconstruction. The *overall* metric is the algebraic mean of *accuracy* and *completeness*. Lower scores indicate better performance in this benchmark.

The result on the DTU dataset is reported in Tab. 1. For a fair comparison, all techniques were trained on the same dataset and employed the same validation and train split. 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 is bigger than their gap in *accuracy*, which motivated us to use a coarse-to-fine PMS to build our initial depth

Method	Accuracy $(mm)\downarrow$	Completeness $(mm)\downarrow$	Overall $(mm)\downarrow$
Traditional photogrammetry-based			
Camp [55]	0.835	0.554	0.695
Furu [4]	0.613	0.941	0.777
Tola [6]	0.342	1.190	0.766
Gipuma [5]	**0.283**	0.873	0.578
Learning-based			
SurfaceNet [9]	0.450	1.040	0.745
MVSNet [7]	0.396	0.527	0.462
R-MVSNet [8]	0.383	0.452	0.417
CIDER [15]	0.417	0.437	0.427
P-MVSNet [14]	0.406	0.434	0.420
Point-MVSNet [10]	0.342	0.411	0.376
AttMVS [56]	0.383	0.329	0.356
Fast-MVSNet [11]	0.336	0.403	0.370
Vis-MVSNet [57]	0.369	0.361	0.365
CasMVSNet [13]	0.325	0.385	0.355
UCS-Net [12]	0.338	0.349	0.344
EPP-MVSNet [58]	0.413	0.296	0.355
CVP-MVSNet [16]	0.296	0.406	0.351
AA-RMVSNet [59]	0.376	0.339	0.357
DEF-MVSNET [38]	0.402	0.375	0.388
ElasticMVS [40]	0.374	0.325	0.349
MG-MVSNET [41]	0.358	0.338	0.348
BDE-MVSNet [39]	0.338	0.302	0.320
UniMVSNet [53]	0.352	0.278	0.315
TransMVSNet [54]	0.321	0.289	**0.305**
PatchmatchNet [17]	0.427	0.277	0.352
PatchmatchNet + Ours ($L_{1,4}$)	0.405	**0.267**	0.336
PatchmatchNet + 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 [48]. 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.

estimation block, to reduce the *accuracy* gap while still improving completeness. During a similar development phase, several methods such as UniMVSNet [53] and TransMVSNet [54] have been published, showcasing superior performance compared to our proposed approach. However, despite these advancements, we maintain a strong belief in the effectiveness of our Depth Discontinuity Learning (DDL) module in enhancing the baseline performance of PatchmatchNet. This reveals that learning depth discontinuities is an effective means to improve both reconstruction accuracy and completeness.

4.3. Evaluation on "Tanks and Temples" dataset

In this section, we present our findings based on the " Tanks and Temples" dataset [50]. 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 [17]. 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

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 [17] on the ETH3D training set [51]. 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).

	Intermediate set			Advanced set		
Methods	P (%) ↑	R (%) ↑	F-score ↑	P (%) ↑	R (%) ↑	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 3. Evaluation and comparison with PatchmatchNet [17] on the “Tanks and Temples" dataset [50].

	Point clouds (testing)			Depth maps (validation)	
Methods	Acc. (mm) ↓	Comp. (mm) ↓	Overall (mm) ↓	Depth map (mm) ↓	Error ratio (%; error $>$ 8 mm) ↓
PatchmatchNet [17]	0.427	0.277	0.352	7.09	11.58
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	**5.28**	**8.79**

Table 4. Ablation study on the point clouds and depth maps from the DTU dataset [48]. L_1: depth-groundtruth loss; L_2: edge-depth loss; L_3: smoothness loss; L_4: bimodal loss. Note that L_2 and L_3 cannot be separated because they together work for edge-aware smoothness.

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. As depicted in Fig. 3, our approach improves baseline [17] in accurately capturing the overall geometry and exhibits improved completeness in smooth regions. This is substantiated by both qualitative and quantitative results, which demonstrate that our approach outperforms the baseline in terms of overall reconstruction quality.

4.4. Evaluation on ETH3D dataset

In this section, we present our findings based on the ETH3D benchmark [51]. The ETH3D benchmark [51] 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 [17] on the ETH3D dataset [51] 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.

4.5. Ablation study

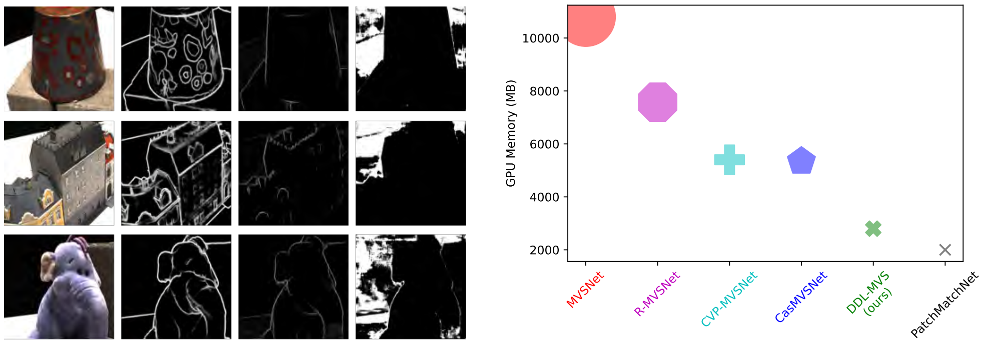
We have conducted an ablation study to understand and analyze the contributions of the aforementioned loss terms of our architecture. The results are detailed in Tab. 4. Since the edge-depth loss and the smoothness loss terms together strive for edge-aware smoothness, we do not separate them in our experiments. We retrieve the last two metrics from the validation set while tuning our hyper-parameters. The “Depth map” 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 a higher error than 8 mm.

From Tab. 4, we can see that using all lost terms improves the depth map quality on the validation set. For testing, we observe that our point clouds have better completeness and overall

	Boundary and Smooth region		Depth maps
Methods	Boundary region (mm) ↓	Smooth region (mm) ↓	Whole depth map (mm) ↓
PatchmatchNet [17]	22.05	6.66	7.09
Ours	**19.86**	**4.84**	**5.28**

Table 5. Evaluation of depth map errors in boundary and smooth regions using the DTU dataset [48].

metrics with bimodal and depth ground-truth loss while having edge-aware smoothness term results in better accuracy. Our network also improves the arithmetic mean of accuracy and completeness if we compare it against the baseline.



(a) Edge maps **(b)** GPU memory consumption

Figure 4. Edges maps and GPU Memory consumption. (a) Edges maps of a few randomly chosen examples. For each example, the images from left to right are the color image, the edge map predicted by HED [60], our learned edge map, and the α map, respectively. We can see that our learned edge maps better capture the depth discontinuities, regardless of the photometric changes. It is also interesting to observe that our α maps distinguish between foreground and background. (b) Comparison of GPU memory demands with existing learning-based MVS networks on DTU dataset with image size 1152×864.

4.6. Effect 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 (a), and compare them with the edge maps predicted using the seminal learning-based edge detection method HED [60]. 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. In the earlier stage of the development of DDL-MVS, we tried to feed the network with HED [60] output and jointly refine the depth and edge maps similar to EdgeStereo [37]. It turned out that even after refinement, the edges were too sensitive to photometric changes, leading to higher depth errors.

To reveal how our depth discontinuity learning contributes to depth estimation, we demonstrate the α map of each example in the last column of Fig. 4 (a), where α 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, for which the α values express a binary classification for foreground and background pixels.

Our suggested framework enhances the quality of depth maps for both smooth and boundary regions, as demonstrated quantitatively in Tab. 5. We have computed mean absolute error (MAE) between the estimated depth and the groundtruth. In contrast to the rest of the pixels, which correspond to a smooth area, boundary pixels are those pixels where the laplacian of the groundtruth depth is greater than 5.

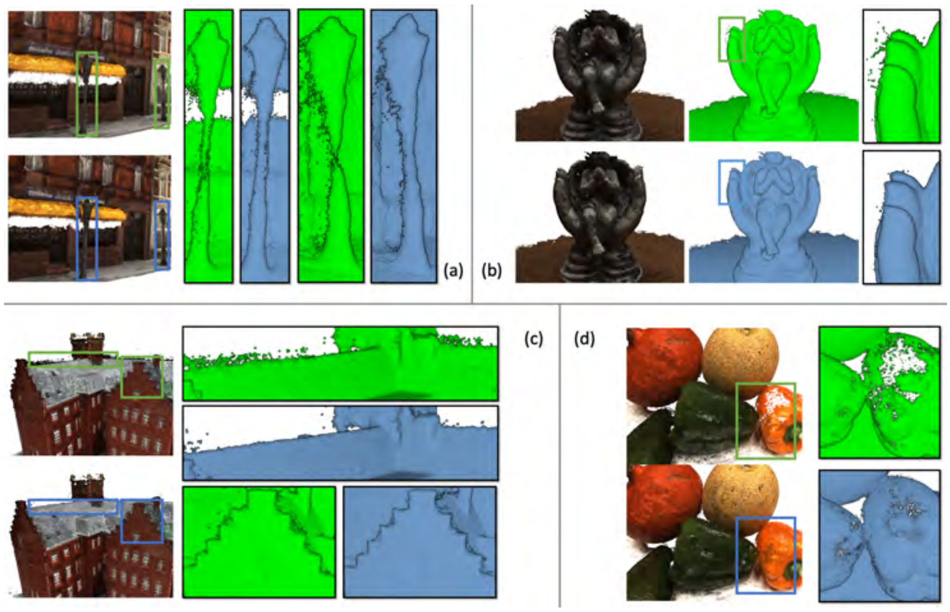


Figure 5. Comparison between our method and the baseline method PatchmatchNet [17] on a set of scenes from the DTU dataset [48]. For each scene, the colored image at the top shows with green boxes the results from PatchmatchNet, and the colored image at the bottom with blue boxes shows the results from our method. A zoomed view of each marked image region is shown on the right of each result. The blue images depict our results, while the green images depict the baseline results.

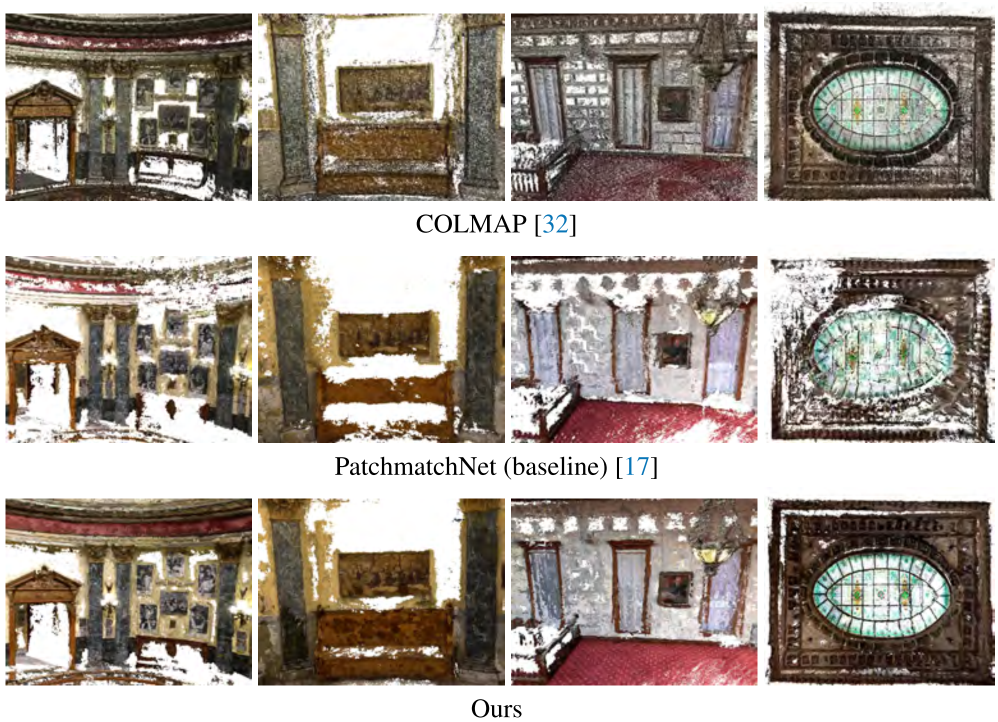


Figure 6. Comparisons with the state-of-the-art traditional photogrammetry-based MVS method COLMAP [32] and learning-based MVS method PatchmatchNet [17].

The proposed approach also enhances the quality of the point clouds as demonstrated qualitatively in Fig. 5, from which we can see that thin structures and smooth regions are captured more completely, and the boundary regions have a lower amount of noise.

Figure 6 presents a comparison of our proposed learning-based MVS method with two methods, namely COLMAP [32] and PatchmatchNet [17] (our baseline). COLMAP is a state-of-the-art traditional photogrammetry-based method. We visualized the outcomes of the methods on four different scene parts from the "Tanks and Temples" [50] dataset, and the last column of the figure shows the exterior of the *courtroom*'s top, with the lower part of the point cloud clipped to better reveal the ceiling's completeness and accuracy.

To ensure a fair comparison, we provided COLMAP with the ground-truth camera parameters. Our experiments demonstrate that our proposed method generates denser and more complete point clouds than the traditional photogrammetry-based method. However, the traditional method achieves better accuracy, partially due to its sparsity. Our method's results exhibit the highest completeness and are cleaner than the other methods. Additionally, our method outperforms PatchmatchNet [17] in terms of reconstruction accuracy. Please refer to the supplementary video for more visual comparisons.

4.7. Generalization to aerial images

To further evaluate the generalization capabilities of our proposed methods, we conducted experiments using aerial images. Aerial images are commonly used in remote sensing applications for tasks such as large-scale 3D reconstruction. For our experiments, we utilized the BlendedMVS dataset [52], which consists of aerial images with a low resolution of 768×576 images.

Figure 7 demonstrates some example images from the BlendedMVS dataset, illustrating the qualitative results obtained from our proposed methods. These results show that our method can generate 3D reconstructions from aerial images, even with low-resolution input images. This indicates the potential of our approach for remote sensing applications that require large-scale 3D reconstructions from aerial imagery.

On a single RTX 2080, the time needed for depth inference per image is 90 *ms* when using 5 neighboring views, and increases to 110 *ms* when using 7 neighboring views. As an illustration of the running time, the bottom left building example in Fig. 7 is comprised of 77 images. It takes 79.993 seconds to generate a point cloud from the calibrated views with the default 5 neighboring views.

Figure 7. Experiments with BlendedMVS dataset. Qualitative results of our proposed methods for aerial image-based 3D reconstruction are visualized here.

4.8. Memory consumption and running times

In Fig. 4, we report our comparison of GPU memory demands with existing learning-based MVS networks on the DTU dataset [48], from which we can see that the memory demand of our network is much lower than most of the existing networks. In the DTU dataset with the default parameters and the 5-view case, the average depth inference time for our model is 345*ms*. This is comparable to the performance of PatchmatchNet [17], which took $300ms$. We used a GPU of NVIDIA GeForce RTX 2080 for the experiments.

4.9. Limitations

Although our method has good completeness and a good overall score (see Tab. 1), it has still not reached the accuracy level of traditional photogrammetry-based algorithms such as Gipuma [5], which is a common weakness in recently developed learning-based MVS methods with high completeness score. In this paper, our goal is to improve the accuracy of the reconstruction process while simultaneously maintaining a high level of completeness. Although the accuracy of our proposed network is not among the highest compared to some traditional state-of-the-art methods, we would like to emphasize that currently, learning-based approaches struggle to achieve a state-of-the-art accuracy result while maintaining a high completeness score. This is due to the trade-off between accuracy and completeness in the depth map fusion process, which is a key component of the reconstruction pipeline. Such a trade-off implies that increasing completeness leads to an increasing potential source of noise. Although using bimodality helps to reduce the noise, we observe that our work, like other traditional and learning-based algorithms, contains noise, especially in sparsely viewed regions that may need further research. It is also worth noting that in this work we have used the same fusion pipeline as in other papers [7,17].

5. Conclusion

We have presented a strategy for improving the baseline MVS network by learning depth discontinuities. The proposed depth discontinuity learning module has demonstrated superior performance compared to the baseline [17]. The results of our ablation study, as shown in Tab. 4, highlight the significant reduction in depth map error achieved by incorporating the proposed DDL module, reducing the error by more than 30%. Experimental findings presented in Tab. 5 demonstrate the enhanced quality of our approach in terms of depth map accuracy in smooth and boundary regions. Moreover, our visual results shown in Fig. 3 and Fig. 5 revealed that the reconstructed point cloud obtained from our approach exhibits improved accuracy in capturing object and scene details compared to the baseline model while maintaining completeness.

The results of Fig. 5 and Tab. 5 further reinforce the superiority of our method, with better qualitative and quantitative results in both smooth and boundary regions in the DTU [48] dataset. These results indicate that our method has strong generalization capabilities and the ability to produce high-quality depth maps with improved accuracy and precision. Furthermore, our experimental results demonstrate the potential of our method for remote sensing applications, such as large-scale point cloud reconstruction from aerial images.

Our experiments have demonstrated that learning depth maps as a mixture distribution and integrating depth discontinuities into the network as prior knowledge for piecewise smoothness regularization leads to improved reconstruction quality, with enhanced accuracy and overall quality of the final reconstruction.

Author Contributions: Conceptualization N.I. and L.N.; implementation, software and methodology, N.I.; original draft preparation, N.I. and L.N.; review and editing N.I., L.N., J.K., H.L.; supervision L.N., J.K., H.L.;

Funding: This work was supported by the 3D Urban Understanding Lab funded by the TU Delft AI Initiative.

Data Availability Statement: Openly available public datasets have been utilized and cited in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lemaire, C. Aspects of the DSM production with high resolution images. *ISPRS* **2008**, *37*, 1143–1146.
2. Peppa, M.V.; Mills, J.P.; Moore, P.; Miller, P.E.; Chambers, J.E. Automated co-registration and calibration in SfM photogrammetry for landslide change detection. *Earth Surf. Process. Landf.* **2019**, *44*, 287–303.
3. Nguatem, W.; Mayer, H. Modeling urban scenes from pointclouds. In Proceedings of the ICCV. IEEE, 2017, pp. 3857–3866.
4. Furukawa, Y.; Ponce, J. Accurate, dense, and robust multi-view stereopsis. *IEEE TPAMI* **2010**, *32*, 1362–1376.
5. Galliani, S.; Lasinger, K.; Schindler, K. Massively parallel multiview stereopsis by surface normal diffusion. In Proceedings of the ICCV. IEEE, 2015, pp. 873–881.

6. Tola, E.; Strecha, C.; Fua, P. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine vision and applications* **2012**, *23*, 903–920.
7. Yao, Y.; Luo, Z.; Li, S.; Fang, T.; Quan, L. MVSNet: Depth inference for unstructured multi-view stereo. In Proceedings of the ECCV. Springer, 2018, pp. 767–783.
8. Yao, Y.; Luo, Z.; Li, S.; Shen, T.; Fang, T.; Quan, L. Recurrent MVSNet for high-resolution multi-view stereo depth inference. In Proceedings of the CVPR. IEEE, 2019, pp. 5525–5534.
9. Ji, M.; Gall, J.; Zheng, H.; Liu, Y.; Fang, L. SurfaceNet: An end-to-end 3D neural network for multiview stereopsis. In Proceedings of the ICCV. IEEE, 2017, pp. 2307–2315.
10. Chen, R.; Han, S.; Xu, J.; Su, H. Point-based multi-view stereo network. In Proceedings of the ICCV, 2019, pp. 1538–1547.
11. Yu, Z.; Gao, S. Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and Gauss-Newton refinement. In Proceedings of the CVPR. IEEE, 2020, pp. 1949–1958.
12. 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 Proceedings of the CVPR. IEEE, 2020, pp. 2524–2534.
13. Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; Tan, P. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In Proceedings of the CVPR. IEEE, 2020, pp. 2495–2504.
14. Luo, K.; Guan, T.; Ju, L.; Huang, H.; Luo, Y. P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo. In Proceedings of the ICCV. IEEE, 2019, pp. 10451–10460.
15. Xu, Q.; Tao, W. Learning inverse depth regression for multi-view stereo with correlation cost volume. In Proceedings of the AAAI, 2020, Vol. 34, pp. 12508–12515.
16. Yang, J.; Mao, W.; Alvarez, J.M.; Liu, M. Cost volume pyramid based depth inference for multi-view stereo. In Proceedings of the CVPR. IEEE, 2020, pp. 4877–4886.
17. Wang, F.; Galliani, S.; Vogel, C.; Speciale, P.; Pollefeys, M. Patchmatchnet: Learned multi-view patchmatch stereo. In Proceedings of the CVPR. IEEE, 2021, pp. 14194–14203.
18. Duggal, S.; Wang, S.; Ma, W.C.; Hu, R.; Urtasun, R. DeepPruner: Learning efficient stereo matching via differentiable patchmatch. In Proceedings of the ICCV. IEEE, 2019, pp. 4384–4393.
19. Zhu, S.; Brazil, G.; Liu, X. The edge of depth: Explicit constraints between segmentation and depth. In Proceedings of the CVPR. IEEE, 2020, pp. 13116–13125.
20. Tosi, F.; Liao, Y.; Schmitt, C.; Geiger, A. SMD-Nets: Stereo mixture density networks. In Proceedings of the CVPR. IEEE, 2021, pp. 8942–8952.
21. Boykov, Y.; Veksler, O.; Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE TPAMI* **2001**, *23*, 1222–1239.
22. Boykov, Y.; Veksler, O.; Zabih, R. Markov random fields with efficient approximations. In Proceedings of the CVPR. IEEE, 1998, pp. 648–655.
23. Garg, D.; Wang, Y.; Hariharan, B.; Campbell, M.; Weinberger, K.Q.; Chao, W.L. Wasserstein distances for stereo disparity estimation. In Proceedings of the NeurIPS, 2020, Vol. 33, pp. 22517–22529.
24. Janai, J.; Güney, F.; Behl, A.; Geiger, A.; et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision* **2020**, *12*, 1–308.
25. Kutulakos, K.N.; Seitz, S.M. A theory of shape by space carving. *IJCV* **2000**, *38*, 199–218.
26. Faugeras, O.; Keriven, R. *Variational principles, surface evolution, PDE's, level set methods and the stereo problem*; IEEE, 2002.
27. Lorensen, W.E.; Cline, H.E. Marching cubes: A high resolution 3D surface construction algorithm. *ACM siggraph computer graphics* **1987**, *21*, 163–169.
28. Curless, B.; Levoy, M. A volumetric method for building complex models from range images. In Proceedings of the Computer graphics and interactive techniques, 1996, pp. 303–312.
29. Zach, C.; Pock, T.; Bischof, H. A globally optimal algorithm for robust tv-l 1 range image integration. In Proceedings of the ICCV. IEEE, 2007, pp. 1–8.
30. Collins, R.T. A space-sweep approach to true multi-image matching. In Proceedings of the CVPR. IEEE, 1996, pp. 358–363.
31. Pollefeys, M.; Nistér, D.; Frahm, J.M.; Akbarzadeh, A.; Mordohai, P.; Clipp, B.; Engels, C.; Gallup, D.; Kim, S.J.; Merrell, P.; et al. Detailed real-time urban 3d reconstruction from video. *IJCV* **2008**, *78*, 143–167.
32. Schönberger, J.L.; Zheng, E.; Pollefeys, M.; Frahm, J.M. Pixelwise view selection for unstructured multi-view stereo. In Proceedings of the ECCV. Springer, 2016.
33. Zbontar, J.; LeCun, Y.; et al. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.* **2016**, *17*, 2287–2318.
34. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; Bry, A. End-to-end learning of geometry and context for deep stereo regression. In Proceedings of the ICCV. IEEE, 2017, pp. 66–75.
35. Chang, J.R.; Chen, Y.S. Pyramid stereo matching network. In Proceedings of the CVPR. IEEE, 2018, pp. 5410–5418.
36. Yang, G.; Manela, J.; Happold, M.; Ramanan, D. Hierarchical deep stereo matching on high-resolution images. In Proceedings of the CVPR. IEEE, 2019, pp. 5515–5524.
37. Song, X.; Zhao, X.; Hu, H.; Fang, L. Edgestereo: A context integrated residual pyramid network for stereo matching. In Proceedings of the ACCV. Springer, 2018.

38. Lin, K.; Li, L.; Zhang, J.; Zheng, X.; Wu, S. High-Resolution Multi-View Stereo with Dynamic Depth Edge Flow. In Proceedings of the ICME. IEEE, 2021, pp. 1–6.
39. Ding, Y.; Li, Z.; Huang, D.; Zhang, K.; Li, Z.; Feng, W. Adaptive Range guided Multi-view Depth Estimation with Normal Ranking Loss. In Proceedings of the ACCV, 2022, pp. 1892–1908.
40. Zhang, J.; Tang, R.; Cao, Z.; Xiao, J.; Huang, R.; Fang, L. ElasticMVS: Learning elastic part representation for self-supervised multi-view stereopsis. *NeurIPS* **2022**, *35*, 23510–23523.
41. Zhang, X.; Yang, F.; Chang, M.; Qin, X. MG-MVSNet: Multiple Granularities Feature Fusion Network for Multi-View Stereo. *Neurocomputing* **2023**.
42. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention. Springer, 2015, pp. 234–241.
43. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the CVPR. IEEE, 2017, pp. 2117–2125.
44. Hui, T.W.; Loy, C.C.; Tang, X. Depth map super-resolution by deep multi-scale guidance. In Proceedings of the ECCV. Springer, 2016.
45. Yu, A.; Guo, W.; Liu, B.; Chen, X.; Wang, X.; Cao, X.; Jiang, B. Attention aware cost volume pyramid based multi-view stereo network for 3D reconstruction. *ISPRS J. of Photogrammetry and Remote Sensing* **2021**, *175*, 448–460.
46. Huang, J.; Lee, A.; Mumford, D. Statistics of range images. In Proceedings of the CVPR. IEEE, 2000, Vol. 1, pp. 324–331 vol.1.
47. Laplace, P.S. Laplace distribution. *Encyclopedia of Mathematics* **1801**. Original publication in 1801, available in English translation.
48. Aanæs, H.; Jensen, R.R.; Vogiatzis, G.; Tola, E.; Dahl, A.B. Large-scale data for multiple-view stereopsis. *IJCV* **2016**, pp. 1–16.
49. Laplace, P.S. Laplace operator. *Encyclopedia of Mathematics* **1820**. Original publication in 1820, available in English translation.
50. Knapitsch, A.; Park, J.; Zhou, Q.Y.; Koltun, V. Tanks and Temples: Benchmarking large-scale scene reconstruction. *ACM TOG* **2017**, *36*.
51. Schöps, T.; Schönberger, J.L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; Geiger, A. A Multi-View Stereo Benchmark with high-resolution images and multi-camera videos. In Proceedings of the CVPR. IEEE, 2017.
52. Yao, Y.; Luo, Z.; Li, S.; Zhang, J.; Ren, Y.; Zhou, L.; Fang, T.; Quan, L. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In Proceedings of the CVPR. IEEE, 2020, pp. 1790–1799.
53. Peng, R.; Wang, R.; Wang, Z.; Lai, Y.; Wang, R. Rethinking depth estimation for multi-view stereo: A unified representation. In Proceedings of the CVPR. IEEE, 2022, pp. 8645–8654.
54. Wang, X.; Zhu, Z.; Huang, G.; Qin, F.; Ye, Y.; He, Y.; Chi, X.; Wang, X. MVSTER: epipolar transformer for efficient multi-view stereo. In Proceedings of the ECCV. Springer, 2022, pp. 573–591.
55. Campbell, N.D.F.; Vogiatzis, G.; Hernández, C.; Cipolla, R. Using Multiple Hypotheses to Improve Depth-Maps for Multi-View Stereo. In Proceedings of the ECCV; Forsyth, D.; Torr, P.; Zisserman, A., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp. 766–779.
56. Luo, K.; Guan, T.; Ju, L.; Wang, Y.; Chen, Z.; Luo, Y. Attention-aware multi-view stereo. In Proceedings of the CVPR. IEEE, 2020.
57. Zhang, J.; Yao, Y.; Li, S.; Luo, Z.; Fang, T. Visibility-aware multi-view stereo network. In Proceedings of the BMVC, 2020.
58. 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 ICCV. IEEE, 2021, pp. 5732–5740.
59. Wei, Z.; Zhu, Q.; Min, C.; Chen, Y.; Wang, G. AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network. In Proceedings of the ICCV. IEEE, 2021, pp. 6187–6196.
60. Xie, S.; Tu, Z. Holistically-nested edge detection. In Proceedings of the ICCV. IEEE, 2015, pp. 1395–1403.