[Supplementary Material] Uncertainty-Aware Deep Multi-View Photometric Stereo

Berk Kaya[1] Suryansh Kumar[1*] Carlos Oliveira[1] Vittorio Ferrari[2] Luc Van Gool[1,3]
ETH Zürich[1], Google Research[2], KU Leuven[3]

Abstract

Our supplementary material is organized as follows: First, for completeness, we provide more details on the network pipeline to assist notations, symbols, and equation clarification. Next, we extend the experimental evaluations supplied in the main paper. Specifically, we supply an additional statistical comparison with the TSDF fusion method. As highlighted in the main paper, the global metric may not be a true reflection of the recovered surface topology; therefore, we additionally compare the mesh quality of the recovered surfaces with the baselines to demonstrate the superiority of our approach. ***Our supplementary also includes a video that demonstrates the visual results. We highly recommend the reader to check our video.***

A. Further Clarification

Although we present a dense description of our proposed method and experimental evaluations, we want to provide more details for completeness and further clarification. To that end, we define our evaluation metrics with explicit mathematical formulations. We also reiterate the PatchMatch based deep-MVS network pipeline by clarifying the notations, symbols, and equations.

A.1. Evaluation Metrics

Our quantitative analysis is based on Chamfer-L_1 distance, precision and $\mathcal{F}$-score on the reconstructed and ground-truth point sets: $\mathcal{R}, \mathcal{G} \subset \mathbb{R}^3$. For a single reconstructed point $r \in \mathcal{R}$, distance to the ground-truth is defined as follows:

$$d_{r\rightarrow\mathcal{G}} = \min_{g\in\mathcal{G}} \|r - g\| \,. \tag{1}$$

The individual distance measures are accumulated to define Chamfer-L_1 distance and $\mathcal{F}$-score as follows:

$$CD = \frac{1}{2|\mathcal{X}_1|}\sum_{x\in\mathcal{X}_1} d_{x\rightarrow\mathcal{X}_2} + \frac{1}{2|\mathcal{X}_2|}\sum_{x\in\mathcal{X}_2} d_{x\rightarrow\mathcal{X}_1}, \tag{2}$$

$$\mathcal{F}(\tau) = \frac{2P(\tau)R(\tau)}{P(\tau)+R(\tau)}, \tag{3}$$

where

$$P(\tau) = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}}[d_{r\rightarrow\mathcal{G}} < \tau], \tag{4}$$

$$R(\tau) = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}[d_{g\rightarrow\mathcal{R}} < \tau], \tag{5}$$

stand for precision and recall measures respectively. Here, [.] is the Iverson bracket and τ is the distance threshold.

A.2. Network Design

This section provides a detailed description of our PatchMatch based deep-MVS network [59] which we use to estimate the per-pixel depth. We start by explaining why we prefer this framework in our approach. Then, we provide an in-depth description of the network pipeline for better understanding.

• **Why PatchMatch based deep-MVS network?** MVS aims for the reconstruction of the dense 3D geometry given a collection of images with known camera parameters. Traditional approaches generally rely on hand-crafted features to find the correspondences among different views and perform triangulation to reconstruct the scene [12]. Despite the practical applicability of the traditional MVS, it is still fragile against illumination changes, occlusions, non-textured areas, and non-Lambertian surfaces. In recent years, many deep learning-based MVS methods have been proposed to overcome such challenges by utilizing the power of neural networks [7, 40, 67, 69]. However, common deep MVS methods require large GPU memory, provide inferior runtime performance, and therefore, are not applicable to full resolution scenes. To mitigate these shortcomings, we use a PatchMatch based deep MVS network in our approach [59].

Traditional PatchMatch puts forward a randomized and iterative algorithm to find approximate nearest-neighbor matches between image patches [2]. The extension of the algorithm is used in the scene space for better-performing MVS [3, 16, 66, 71]. The PatchMatch based MVS is fast, allows for sub-pixel precision, and handles foreshortening problems for large baseline stereo setups. Consequently,

*Corresponding Author (k.sur46@gmail.com)

our PatchMatch based deep MVS network demonstrates all these benefits with low memory requirements and fast runtime capabilities, making it an ideal choice for our problem.

• **Learning-based PatchMatch Network.** Similar to PatchMatch algorithm [2], the PatchMatch based deep-MVS network employs [2] via similar three steps (but in 3d scene space) as follows: ***(i)*** Initialization step: Generating depth hypotheses, ***(ii)*** Propagation step: Propagate the hypotheses to neighbors, and ***(iii)*** Evaluation step: Compute the similarity cost and search for best solution. We apply these steps on per-pixel multi-scale features that are hierarchically extracted from MVS images $\mathcal{Y}$ at M different resolution scales [37, 59]. This allows us to estimate depth in a coarse-to-fine manner. Before providing more details on these iterative steps, we reintroduce the notation for clarification. We denote the reference frame by $Y^r \in \mathbb{R}^{w \times h}$, coordinates of the i^{th} pixel by $\mathbf{y}_i$, frame r feature by Φ^r, and camera r intrinsic calibration matrix by $\mathbf{K}_r$. For each reference frame, we pick N_s source frames where $Y^s \in \mathbb{R}^{w \times h}$ denotes a source frame. $(\mathbf{R}_{r,s}, \mathbf{t}_{r,s})$ denotes the relative motion between frame r and s. We skip to add extra notation for stage number for simplicity of writing.

(i) Initialization. In the first iteration, we randomly sample per pixel $\mathcal{D}_f$ depth hypotheses in the pre-defined inverse depth range $[d_{min}, d_{max}]$. Our sampling strategy ensures that the inverse depth range interval sampled into $\mathcal{D}_f$ hypotheses is proper, and one hypothesis is covered at each interval. Once initialized, local perturbations are invoked in the subsequent iteration at each stage to diversify the hypotheses and make the method robust to front-to-parallel surface issues [3]. For local perturbation, per pixel, N_l^m hypotheses are generated at stage m in the normalized inverse depth range R_m.

(ii) Propagation. Let Φ^r denote the reference feature map, ϵ_j the fixed 2D offset for depth hypothesis j, and $\tilde{\epsilon}_j(\mathbf{y}_i)$ the learnable 2D offset for pixel i at coordinates $\mathbf{y}_i$. A 2D CNN is applied on Φ^r to learn the 2D offset for each pixel. The depth hypotheses $\mathbf{D}_p$ at pixel i is obtained as follows:

$$\mathbf{D}_p(\mathbf{y}_i) = \{\mathbf{D}(\mathbf{y}_i + \epsilon_j + \tilde{\epsilon}_j(\mathbf{y}_i))\}_{j=1}^{N_d^m} \tag{6}$$

where, N_d^m denotes the number of depth hypotheses at stage m and $\mathbf{D}$ denotes the depth map in the last iteration. The learnable offset idea based on features allows to gather the hypotheses from the same surface rather than in the fixed set of neighbors, hence it is faster and more accurate.

(iii) Evaluation. Let $\Phi^r(\mathbf{y}_i)$, $\Phi^s(\mathbf{y}_i^{s,j}) \in \mathbb{R}^C$ be the reference feature and the warped source feature maps of pixel i and depth hypothesis d_j, respectively. Here, C is the number of feature channels. We get $\mathbf{y}_{i,j}$ via warping as follows:

$$\mathbf{y}_i^{s,j} = \mathbf{K}_s\Big(\mathbf{R}_{r,s}\big(d_j(\mathbf{y}_i) \cdot \mathbf{K}_r^{-1}\mathbf{y}_i\big) + \mathbf{t}_{r,s}\Big) \tag{7}$$

Next $\Phi^s(\mathbf{y}_i^{s,j})$ is obtained using differentiable bi-linear interpolation. To get the matching cost, we must sum per pixel cost from all the views and the depth hypotheses. For that, the cost per depth hypothesis is computed using group-wise correlation and aggregated over the number of views with per-pixel visibility weight [55, 68]. If G denotes the number of groups into which the feature maps are divided along channel dimension, then g^{th} group similarity $\Delta_s^g \in \mathbb{R}$ for source view s is given by:

$$\Delta_s^g(\mathbf{y}_i, j) = \Lambda < \Phi_g^r(\mathbf{y}_i), \Phi_g^s(\mathbf{y}_i^{s,j}) > \tag{8}$$

Here, $\Lambda \in \mathbb{R}$ is the ratio of number of group to number of channels. Collecting the group similarity for all the pixels and over hypotheses gives $\Delta_s \in \mathbb{R}^{w \times h \times \mathcal{D} \times G}$. For vectorized usage, let $\Delta_s(\mathbf{y}_i, j) \in \mathbb{R}^G$ denote the respective group similarity vector. To incorporate the visibility information per pixel $\mathbf{w}_s(\mathbf{y}_i)$ in the source image Y^s, a network composed of 3D convolutional layer with $1 \times 1 \times 1$ kernels and sigmoid activation is used. This simple pixel-wise network takes the initial set of group similarity Δ_s to provide the visibility weight measure $\mathcal{W}_s \in \mathbb{R}^{w \times h \times \mathcal{D}}$ for a pixel in the range 0 to 1. Accordingly, the view weight is computed as $\mathbf{w}_s(\mathbf{y}_i) = \max(\{\mathcal{W}_s(\mathbf{y}_i, j)\}_{j=0}^{\mathcal{D}-1})$. Using the visibility weight, the weighted group similarity $\tilde{\Delta}(\mathbf{y}_i, j)$ for pixel i and j^{th} depth hypothesis is computed as:

$$\tilde{\Delta}(\mathbf{y}_i, j) = \big(\sum_{s=1}^{N_s} \mathbf{w}_s(\mathbf{y}_i)\big)^{-1}\big(\sum_{s=1}^{N_s} \mathbf{w}_s(\mathbf{y}_i)\Delta_s(\mathbf{y}_i, j)\big) \tag{9}$$

The weighted group similarity over all the pixels and hypotheses is computed as $\tilde{\Delta} \in \mathbb{R}^{w \times h \times \mathcal{D} \times G}$. To get the cost $\mathbf{J} \in \mathbb{R}^{w \times h \times \mathcal{D}}$ per pixel and depth hypothesis, a 3D convolution network with $1 \times 1 \times 1$ kernel is applied on $\tilde{\Delta}$.

For aggregating the matching cost, an adaptive propagation strategy is followed. Similar to the propagation strategy per pixel, an additional spatial offset $\tilde{\mathbf{y}}_i^t$ per pixel i is learnt based on the AANet [3, 65]. For a spatial window with N_w pixels, the spatial cost aggregation is computed as

$$\tilde{\mathbf{J}}(\mathbf{y}_i, j) = \big(\sum_{t=1}^{N_w} w_t \cdot \tilde{d}_t\big)^{-1}\big(\sum_{t=1}^{N_w} w_t \cdot \tilde{d}_t \cdot \mathbf{J}(\mathbf{y}_i + \mathbf{y}_i^t + \tilde{\mathbf{y}}_{\mathbf{i}}^t, j)\big) \tag{10}$$

$\mathbf{y}_i^t$ is the pixel coordinates within the window. d_t and w_t $\forall\ t \in [1, N_w]$ are the weights per pixel based on the depth hypotheses and feature similarity, respectively. Feature weight at a sampled location is based on the feature similarity between corresponding features in Φ^r and $\mathbf{y}_i$, which is computed via group-wise correlation [20]. Whereas, the depth weights are based on the absolute difference in the inverse depth between the sampled location and $\mathbf{y}_i$ using j^{th} hypotheses. To regress the depth per pixel, we apply softmax (σ) to $\tilde{\mathbf{J}}(\mathbf{y}, j)$ which gives the confidence measures $\mathcal{C}$ of the estimation.

$$\mathbf{D}(\mathbf{y}_i) = \sum_{j=0}^{\mathcal{D}-1} d_j(\mathbf{y}_i) \cdot \sigma(\tilde{\mathbf{J}}(\mathbf{y}_i, j)) \tag{11}$$

Further, an independent depth residual network based on Hui *et al*. work [26] is used to obtain the refined depth map $\mathbf{D}_{ref}$. It extracts the features Φ^D from $\mathbf{D}$, the Φ^I from Y^r, and upscale Φ^D to image size via deconvolution. Both of these features are concatenated and subsequently multiple 2D convolution layers are used to compute the depth residual. For more details on the PatchMatch based deep-MVS network, refer [59].

B. Additional Results

In this section, we extend the experimental results in the main paper by providing further statistical analysis and qualitative comparisons.

B.1. Comparison with Standard TSDF Fusion

In the main paper, we already provided the comparative results on two subjects. Table(1) provides $\mathcal{F}$-score and Chamfer-L_1 metric stats for the rest of the DiLiGenT-MV subjects. Clearly, the results show the superiority of our approach against the classical TSDF Fusion approach [9].

Method Type →	TSDF Fusion [9]		Ours	
Dataset↓ | **Metric** →	$\mathcal{F}$-score (↑)	Chamfer-L_1 (↓)	$\mathcal{F}$-score (↑)	Chamfer-L_1 (↓)
BEAR	0.129	4.624	**0.895**	**0.415**
BUDDHA	0.398	2.069	**0.922**	**0.455**
COW	0.192	3.392	**0.979**	**0.329**
POT2	0.056	6.100	**0.907**	**0.515**
READING	0.314	2.238	**0.970**	**0.355**
AVERAGE	0.218	3.685	**0.935**	**0.414**

Table 1. Comparison of the reconstruction quality with TSDF Fusion [9], which is a standard method of choice for robust 3D fusion (outlier removal). We use $\mathcal{F}$-score (higher is better) and Chamfer-L_1 (lower is better) metrics for statistical evaluation.

B.2. Quality of Reconstructed Surface Geometry

Extending the qualitative analysis in the main paper, we demonstrate the quality of the recovered meshes for DiLiGenT-MV objects. Fig.1-Fig.4 show the colored Wireframe model comparison of the object surface recovered using our approach, B-MVPS [36] and GT. The visualizations show that the distribution of the geometric primitives of B-MVPS [36] is **irregular** and **unevenly distributed**. Similarly, Fig.5-Fig.9 show the quality of the meshes compared to NeRF [43], R-MVPS [49], and B-MVPS [36]. Overall, it can be observed that our method provides surfaces which are superior in quality, regular, hence more useful for geometry processing applications.

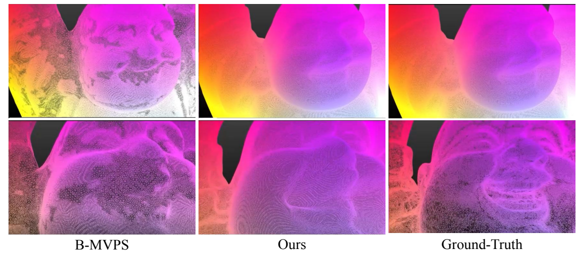

Figure 1. Colored Wireframe qualitative comparison with SOTA B-MVPS [36] on BUDDHA.

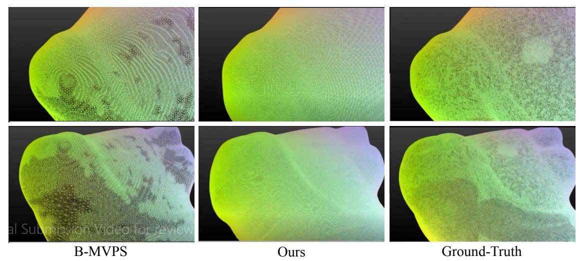

Figure 2. Colored Wireframe qualitative comparison with SOTA B-MVPS [36] on COW.

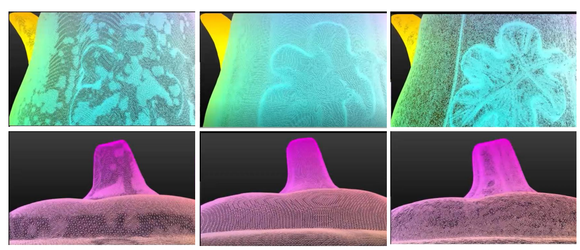

Figure 3. Colored Wireframe qualitative comparison with SOTA B-MVPS [36] on POT2.

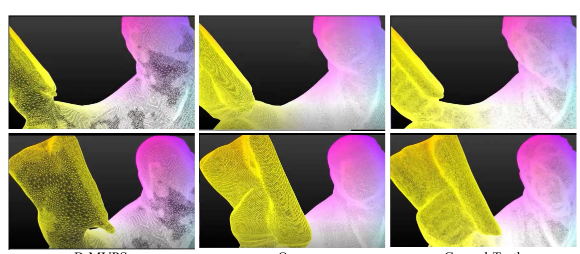

Figure 4. Colored Wireframe qualitative comparison with SOTA B-MVPS [36] on READING.

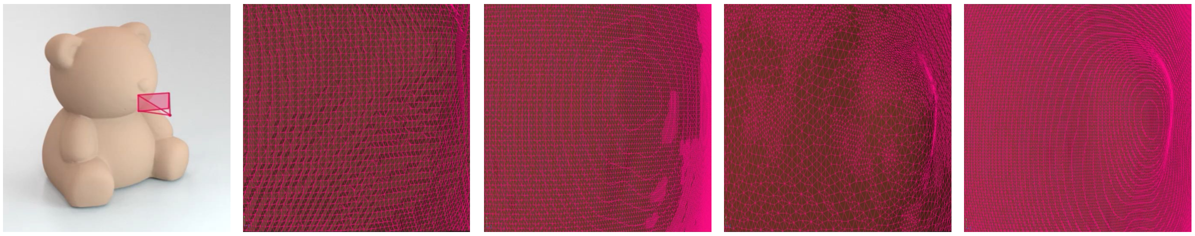

Figure 5. Qualitative mesh comparison on BEAR.

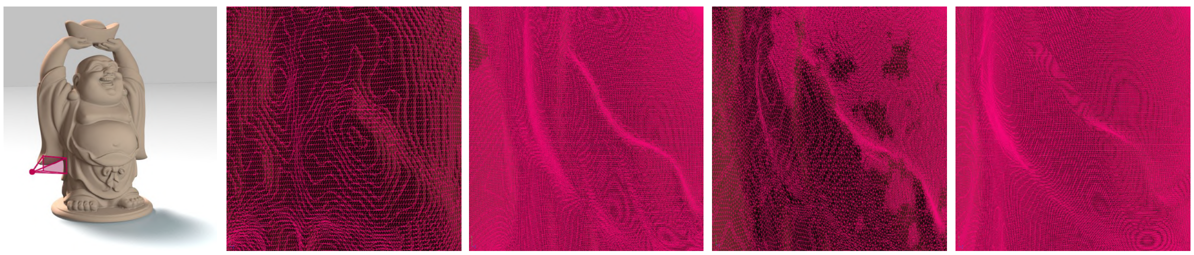

Figure 6. Qualitative mesh comparison on BUDDHA.

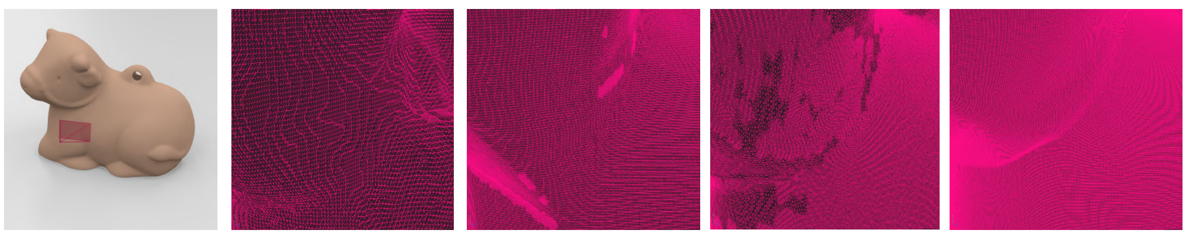

Figure 7. Qualitative mesh comparison on COW.

Figure 8. Qualitative mesh comparison on POT2.

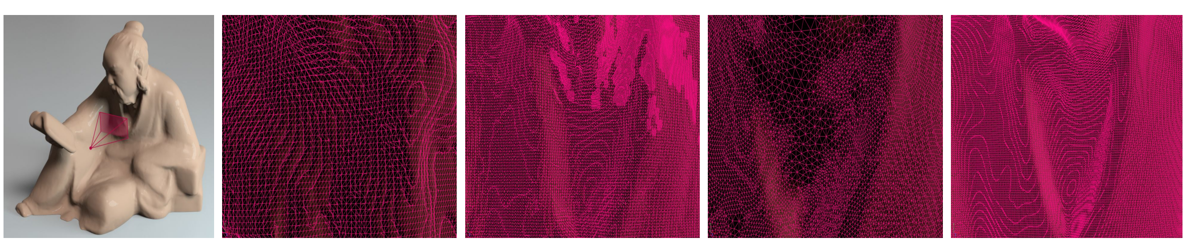

Figure 9. Qualitative mesh comparison on READING.

References

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

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

[3] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011.

[4] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008.

[5] Avishek Chatterjee and Venu Madhav Govindu. Photometric refinement of depth maps for multi-albedo objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 933–941, 2015.

[6] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. Self-calibrating deep photometric stereo networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8747, 2019.

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

[8] Michael G Crandall and Pierre-Louis Lions. Viscosity solutions of hamilton-jacobi equations. *Transactions of the American mathematical society*, 277(1):1–42, 1983.

[9] Brian Curless and Marc 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, 1996.

[10] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5868–5877, 2017.

[11] James Davis, Ravi Ramamoorthi, and Szymon Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–359. IEEE, 2003.

[12] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

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

[14] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*, 2015.

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

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

[17] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

[18] Jason Geng. Structured-light 3d surface imaging: a tutorial. *Advances in Optics and Photonics*, 3(2):128–160, 2011.

[19] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020.

[20] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019.

[21] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[22] Carlos Hernandez, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):548–554, 2008.

[23] B.K. Horn. *Shape Form Shading*. Artificial intelligence. MIT Press, 1989.

[24] Berthold KP Horn and Michael J Brooks. The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*, 33(2):174–208, 1986.

[25] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

[26] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *European conference on computer vision*, pages 353–369. Springer, 2016.

[27] Satoshi Ikehata. Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018.

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

[29] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1261, 2020.

[30] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Uncalibrated neural inverse rendering for photometric stereo of general surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3804–3814, 2021.

[31] Berk Kaya, Suryansh Kumar, Francesco Sarno, Vittorio Ferrari, and Luc Van Gool. Neural radiance fields approach to deep multi-view photometric stereo. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1965–1977, 2022.

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

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

[34] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4649–4657, 2017.

[35] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Superpixel soup: Monocular dense 3d reconstruction of a complex dynamic scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[36] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020.

[37] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[38] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. A differential volumetric approach to multi-view photometric stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1052–1061, 2019.

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

[40] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10452–10461, 2019.

[41] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[42] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Deep level sets: Implicit surface representations for 3d shape inference. *arXiv preprint arXiv:1901.06802*, 2019.

[43] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[44] Theo Moons, Luc Van Gool, and Maarten Vergauwen. *3D reconstruction from multiple images: principles*. Now Publishers Inc, 2009.

[45] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[46] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH 2005)*, 24(3):536–543, 2005.

[47] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. IEEE, 2011.

[48] Tae-Hyun Oh, Hyeongwoo Kim, Yu-Wing Tai, Jean-Charles Bazin, and In So Kweon. Partial sum minimization of singular values in rpca for low-level vision. In *Proceedings of the IEEE international conference on computer vision*, pages 145–152, 2013.

[49] Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Robust multiview photometric stereo using planar mesh parameterization. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1591–1604, 2016.

[50] Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Multiview photometric stereo using planar mesh parameterization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1161–1168, 2013.

[51] Nicholas G Polson and Vadim Sokolov. Deep learning: A bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304, 2017.

[52] Yvain Quéau, Jean-Denis Durou, and Jean-François Aujol. Variational methods for normal integration. *Journal of Mathematical Imaging and Vision*, 60(4):609–632, 2018.

[53] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. Deep photometric stereo network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 501–509, 2017.

[54] Francesco Sarno, Suryansh Kumar, Berk Kaya, Zhiwu Huang, Vittorio Ferrari, and Luc Van Gool. Neural architecture search for efficient uncalibrated deep photometric stereo. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 361–371, 2022.

[55] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[56] B. Shi, Z. Mo, Z. Wu, D. Duan, S. Yeung, and P. Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):271–284, 2019.

[57] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3707–3716, 2016.

[58] Tatsunori Taniai and Takanori Maehara. Neural inverse rendering for general reflectance photometric stereo. In *International Conference on Machine Learning (ICML)*, pages 4857–4866, 2018.

[59] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14203, 2021.

[60] Lawrence B Wolff. Polarization vision: a new sensory approach to image understanding. *Image and Vision computing*, 15(2):81–93, 1997.

[61] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):191139, 1980.

[62] Changchang Wu et al. Visualsfm: A visual structure from motion system. 2011.

[63] Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Asian Conference on Computer Vision*, pages 703–717. Springer, 2010.

[64] Wuyuan Xie, Miaohui Wang, Mingqiang Wei, Jianmin Jiang, and Jing Qin. Surface reconstruction from normals: A robust dgp-based discontinuity preservation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5328–5336, 2019.

[65] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020.

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

[67] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12508–12515, 2020.

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

[69] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[70] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020.

[71] Enliang Zheng, Enrique Dunn, Vladimir Jojic, and Jan-Michael Frahm. Patchmatch based joint view selection and depthmap estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1510–1517, 2014.

[72] Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3d reconstruction with rgb-d cameras. In *Computer graphics forum*, volume 37, pages 625–652. Wiley Online Library, 2018.