Constraining Depth Map Geometry for Multi-View Stereo: A Dual-Depth Approach with Saddle-shaped Depth Cells

Xinyi Ye[1] Weiyue Zhao[1] Tianqi Liu[1] Zihao Huang[1] Zhiguo Cao[1*] Xin Li[2]
[1]Key Laboratory of Image Processing and Intelligent Control, Ministry of Education; School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[2] Department of Computer Science, University of Albany, Albany NY 12222
{xinyiye, zhaoweiyue, tianqiliu, zihaohuang, zgcao}@hust.edu.cn
xli48@albany.edu

Abstract

Learning-based multi-view stereo (MVS) methods deal with predicting accurate depth maps to achieve an accurate and complete 3D representation. Despite the excellent performance, existing methods ignore the fact that a suitable depth geometry is also critical in MVS. In this paper, we demonstrate that different depth geometries have significant performance gaps, even using the same depth prediction error. Therefore, we introduce an ideal depth geometry composed of ***Saddle-Shaped Cell****s, whose predicted depth map oscillates upward and downward around the ground-truth surface, rather than maintaining a continuous and smooth depth plane. To achieve it, we develop a coarse-to-fine framework called Dual-MVSNet (DMVSNet), which can produce an oscillating depth plane. Technically, we predict two depth values for each pixel (****Dual-Depth****), and propose a novel loss function and a checkerboard-shaped selecting strategy to constrain the predicted depth geometry. Compared to existing methods,DMVSNet achieves a high rank on the DTU benchmark and obtains the top performance on challenging scenes of Tanks and Temples, demonstrating its strong performance and generalization ability. Our method also points to a new research direction for considering depth geometry in MVS.*

1. Introduction

Multi-view stereo (MVS) is a fundamental technique that bridges the gap between 2-D photograph clues and 3-D spatial information. It takes multiple 2-D RGB observations, as well as their respective camera parameters, to reconstruct the 3-D representation of a scene. There are many applications for MVS that range from autopilot [12] to virtual reality [10]. Although traditional MVS methods have achieved significant performance, many learning-based methods [39, 17, 16, 24, 9] have demonstrated their superior ability to tackle low-texture and repetitive pattern regions for a more accurate and complete reconstruction.

Generally, reconstructing a scene using learning-based multiview stereo (MVS) techniques involves two phases: depth prediction and fusion rendering. Learning-based methods primarily focus on optimizing the depth prediction process to provide accurate depth maps for subsequent fusion rendering. Therefore, the learning-based MVS reconstruction task can be seen as a depth prediction task. Recent works [9, 32, 34] have improved the accuracy of depth prediction by enhancing feature matching and cost regularization. Additionally, techniques such as deformable convolutions [34, 9] and attention mechanisms [9, 32] have been utilized to obtain accurate depth maps. However, an interesting phenomenon has been observed: *Depth maps with smaller estimation errors might not achieve better 3-D reconstruction quality after fusion rendering.* Are there other factors limiting the accuracy of 3-D reconstruction? After a thorough investigation into the fusion process, we found that the depth geometry is an important factor that has been overlooked in MVS. Different depth geometries suffer from significant performance gaps, even for the same depth estimation error case. Thus it is worth considering *what constitutes a good depth geometry*.

To address the question, as shown in Fig. 1, we introduced two ideal depth geometries representing two extreme cases: geometry composed of one-sided cells vs. geometry composed of a saddle-shaped cell. The former has depth planes on the same side as the ground-truth surface, while the latter oscillates back and forth on both sides of the ground-truth surface. To evaluate the impact of these depth geometries on 3D reconstruction, we artificially controlled cells of the predicted depth planes while ensuring the same absolute error in depth prediction. Interestingly, we found

*Corresponding author
https://github.com/DIVE128/DMVSNet

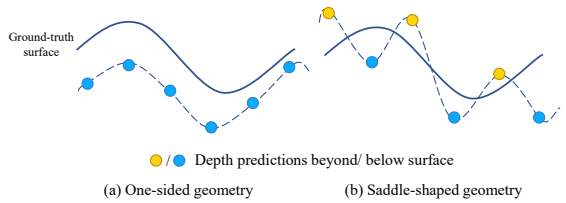


Figure 1. **Brief view of depth geometries.** There are 1-D views of the one-sided geometry and the saddle-shaped geometry.

that saddle-shaped cells significantly improved 3D reconstruction performance compared to one-sided cells. We attribute this performance improvement to the depth interpolation operation during the fusion rendering phase, which is highly sensitive to the depth cell. Saddle-shaped cells can minimize the expected interpolation depth error and thus lead to enhanced 3D reconstruction performance. And we propose a novel method to obtain saddle-shaped depth cells.

The depth geometry with saddle-shaped cells requires the prediction of depth values in an oscillating pattern. We propose a new method called Dual-Depth, which predicts two depth values for each pixel. To achieve an oscillating depth geometry, we first constrain the prediction error of the two depth values separately to ensure that they are both as close as possible to the ground-truth depth. Then, we jointly optimize these two predicted depths by constraining the interval of them. A novel checkerboard selection strategy is also proposed to combine the two depth values to obtain the final depth map. By embedding the above Dual-Depth method into a coarse-to-fine framework, we propose a novel MVS network, named Dual MVSNet (DMVSNet).

We conducted extensive experiments to demonstrate the suitability of the depth geometry with saddle-shaped cells for MVS and the effectiveness of our proposed method. Due to the Dual-Depth method, DMVSNet outperforms most methods in DTU and Intermediate Tanks and Temple. On Advanced Tanks and Temple, DMVSNet achieved SOTA performance, improving by 5.6%. These results highlight the importance of depth geometry for the MVS task and demonstrate that our proposed method is effective. Moreover, our approach offers a new direction for future research, where the depth geometry can be exploited to enhance the reconstruction performance.

Contribution. In this work, we introduced a new perspective for considering depth geometry in MVS. We proposed the depth geometry with saddle-shaped cells for the first time and demonstrated its importance for the MVS reconstruction task. Technically, we proposed the dual-depth method to achieve saddle-shaped cells and designed the corresponding network framework DMVSNet. With the help of the Dual-Depth method, DMVSNet achieves top performance on the DTU dataset and SOTA performance on the Tanks and Temple dataset.

2. Relative Work

Traditional MVS. There are four traditional MVS methods that can be classified based on their output: point-based [27, 2], volume pixels-based [29], mesh-based [13, 35], and depth maps-based [3, 14, 28, 27]. Among them, depth map-based methods break down the reconstruction task into two parts: depth prediction and fusion. Since depth prediction can be performed in parallel and requires only a subset of views, methods based on it are more flexible. After obtaining the estimated depth maps of all images, a fusion process is utilized to generate the 3-D point representation, which is the most commonly used in MVS methods and applications.
Learning-based MVS. Despite traditional MVS methods demonstrating their advantages, they rely on the hand-crafted similarity metric [24]. In contrast, pioneering works such as surefacenet [17] and MVSNet [39] leverage the power of neural networks to generalize potential patterns and learn the metric from the data. MVSNet [39] introduces a learning depth map-based pipeline and is widely applied in the following works. R-MVSNet [40] proposes recurrent structures on cost regularization for efficiency. The coarse-to-fine framework based on the pipeline is presented by [16, 38, 7]. The two essential components of depth prediction are feature matching and cost regularization. Improving the quality of feature representations can benefit both components, and recent research has focused on this aspect. Techniques such as deformable convolution [34, 9] and attention mechanisms [9, 4] have been used to obtain more precise depth maps, resulting in improved reconstruction quality. However, accurate depth maps are not the only determining factor. In this paper, we will show that previously neglected depth geometry is critical as well.
Depth prediction. In addition to multiview depth prediction, there are two other types: Monocular depth prediction [25, 33, 11, 36] and stereo depth prediction [18, 20, 5]. The former is often used in visual effects that do not require a highly accurate depth map because of its inherently ill-posed nature. The latter is more accurate due to the epipolar constraint and can be applied in motion-sensing games and autonomous driving. However, for applications that require a precise and complete depth map, such as 3-D reconstruction, the shading issues inherent in stereo depth prediction can be mitigated by using multiview depth prediction. In the context of MVS, although many studies concentrate on the quality of depth maps, to our knowledge, there is no one that emphasizes the importance of the geometry of the estimated depth, which is the concern of this paper.

3. Motivation

3.1. Estimated bias and interpolated bias

Given a reference image $\boldsymbol{I}_1 \in \mathbb{R}^{3\times H\times W}$ and its source images $\{\boldsymbol{I}_i\}_{i=2}^{N}$, as well as their respective camera intrinsics

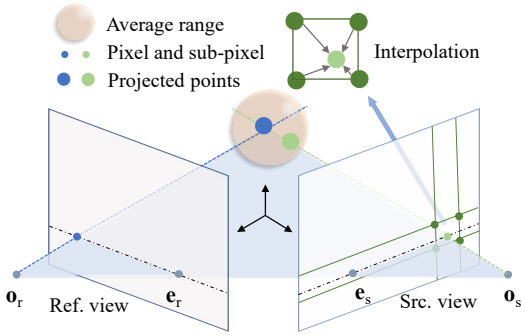


Figure 2. **Fusion process.** Projecting a pixel in the ref. view to the 3-D space and further projecting it to the sub-pixels in the others views. The sub-pixels are reprojected to the 3-D space with the interpolated depth of the depth estimations of their surrounding pixels. The final depth for the pixel in the ref. view to generate the 3-D point is calculated by an averaging among the projected and reprojected points within a range around the projected point.

Settings	Acc.↓	Comp.↓	Overall↓
CasMVSNet	0.366	0.324	0.345
w. one-sided	0.467 (−27.6%)	0.380 (−17.2%)	0.424 (−22.9%)
w. saddle-shapped	0.243 (+36.6%)	0.249 (+23.1%)	0.246 (+28.7%)

Table 1. Results in DTU with different scenarios of Fig. 3. The depth error for them is the same of 10.47mm.

and extrinsics estimated by image matching methods [26, 43, 42, 30, 44], MVS methods predict a depth map $\boldsymbol{D} \in \mathbb{R}^{H\times W}$ aligned with $\boldsymbol{I}_1$. The depth maps are then filtered and utilized to fuse 3-D cloud points with the given camera intrinsics $\{\boldsymbol{K}_i\}_{i=1}^N$ and extrinsics $\{\boldsymbol{T}_i\}_{i=1}^N$.

During the fusion process illustrated in Fig. 2, the pixels in the reference view are projected onto a 3D point in space using the estimated depth map. This 3D point is then reprojected onto sub-pixels in other views using their respective camera parameters, and the corresponding depth maps are used to obtain new 3D points. The final 3D reconstruction result is determined by the depth differences of pixels in the reference view and the estimated depths of subpixels in other views (*e.g.*, by averaging). Therefore, the accuracy of the 3D reconstruction result is affected not only by the accuracy of the estimated depth maps but also by the accuracy of the interpolated depths of subpixels. The subpixel depths are estimated by linearly interpolating the depths of neighboring pixels, and their accuracy is influenced by the estimation bias and depth cell[†]. Fig. 4 shows that the accuracy of the interpolated depth can vary under the same estimation bias and interpolation position due to different depth cells. Therefore, it is important to consider the impact of the depth geometry with different cells for MVS.

3.2. One-sided V.S. Saddle-shaped

To briefly illustrate the difference of depth cells, we present two hypothetical depth cells in Fig. 3: a) One-sided

[†]The concept of "Cell" in this paper is similar to that in HOG [8].

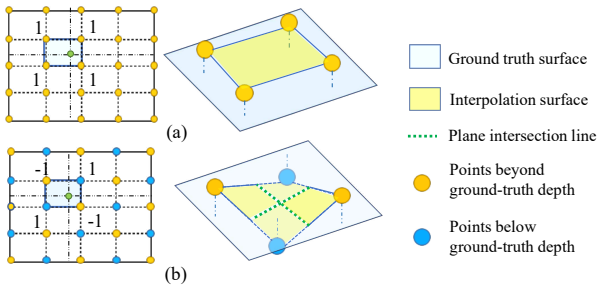


Figure 3. **Two kinds of depth cells with the same estimation bias.** (a)One-sided: all depths are on the same side (beyond or below) of the ground truth surface and there is no intersection line between them. (b)Saddle-shaped: depths of two adjacent pixels are not on the same side of the ground truth and there are two plane intersection lines on any four adjacent pixels. The average absolute estimated bias of (a) and (b) are all '1'. The expectations of absolute interpolated bias are '1' and '0.25' respectively.

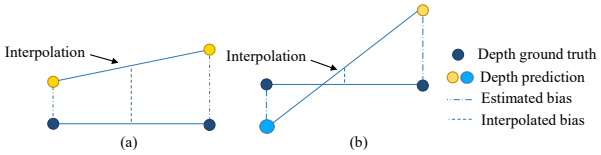


Figure 4. **Estimated bias and Interpolated bias.** Interpolated bias can have significant performance gaps because of the depth geometries (a) and (b), even under the same estimated bias.

Settings	Depth Err.↓	ACC.↓	Comp.↓	Overall↓
+attention	**9.15**	0.369	0.318	0.343
+dcn	9.76	**0.356**	**0.317**	**0.336**

Table 2. A counter-intuitive phenomenon. Despite better depth prediction performance, poorer 3D metrics were obtained.

cells; b) Saddle-shaped cells. We assume that the interpolated positions with the same absolute estimation bias of '1' are uniformly distributed. The spatial volume between the depth plane (yellow) and the ground truth plane (blue) can be considered as the expected absolute interpolation error. Mathematically, the expected absolute interpolation error for the "one-sided cell" is four times higher than that for the "saddle-shaped cell".

To quantitatively demonstrate the impact of depth geometry with different cells on the performance of 3D point reconstruction, we conducted a toy verification experiment. Under the assumption that the absolute estimation bias of each pixel is the same, we flipped the estimated depth values using the true depth, making them distributed according to the two cells shown in Fig. 3. The experimental results in Table 1 show that depth geometries with different cells have a significant impact on the quality of 3D point reconstruction, including accuracy and completeness, even with differences in accuracy exceeding 60% (the second and third rows in Table 1). This indicates that the depth geometry with saddle-shaped cells is a feasible approach to improve the performance of MVS.

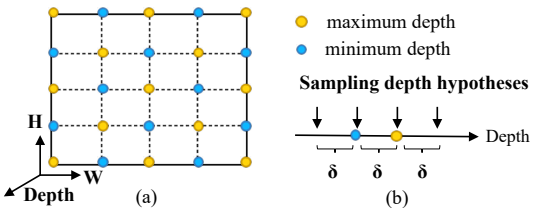


Figure 5. **Checkerboard selection strategy and depth hypotheses sampling.** (a) We alternately select the minimum or maximum depth of the two predictions for each pixel. (b) We define the euclidean distance between the two estimated depths corresponding to a pixel as its uncertainty, which serves as the interval for the next round of depth sampling.

Most existing MVS methods do not impose constraints on depth cells, so their depth maps are distributed between geometries composed of one-sided and saddle-shaped cells, which determines the accuracy of 3D point reconstruction that falls between the performance of the two ideal geometries composed of singular cells (as shown in the first row of Table 1). Besides, without constraints on depth cells, despite better depth prediction performance, poorer 3D metrics may be obtained(Table 2). How can we constrain the network to generate a depth map with more saddle-shaped cells? In the above toy experiment, the method of flipping the estimated depth using the ground truth is a chicken-and-egg problem, which is not feasible in practical inference. In the next section, we will introduce a dual-depth prediction to address this dilemma.

4. Dual-Depth Prediction

4.1. Review of learning-based MVS

In traditional learning-based MVS pipelines, a weight-shared CNN is used to first extract feature maps $\{\boldsymbol{F}_i \in \mathbb{R}^{F^{'} \times H^{'} \times W^{'}}\}_{i=1}^{N}$ aligned with images $\{\boldsymbol{I}_i\}_{i=1}^{N}$, where $H^{'}$, $W^{'}$, and $F^{'}$ represent the height, width, and number of channels of the feature map, respectively. The depth hypotheses for a pixel $\{d_i\}_{i=1}^{M}$ is usually sampled within the range of $[\alpha_1, \alpha_2]$. With the depth hypotheses, as well as camera intrinsic matrix $\boldsymbol{K}$ and extrinsic matrix $\boldsymbol{T}$, a differentiable homography transformation is used to construct a feature volume $\{\boldsymbol{V}_i \in \mathbb{R}^{F^{'} \times M \times H^{'} \times W^{'}}\}_{i=1}^{N}$ in 3D space. At depth d, the homography matrix between the k-th view and the reference camera frustum is given by:

$$\boldsymbol{H}_k^d = d\boldsymbol{K}_k\boldsymbol{T}_k\boldsymbol{T}_1^{-1}\boldsymbol{K}_1^{-1}\,. \tag{1}$$

For the pixels $\boldsymbol{p} \in \mathbb{R}^{2 \times H^{'} \times W^{'}}$ in the reference image, the transformed pixels in the image of the k-th view at depth d are

$$\boldsymbol{p}_k^d = \boldsymbol{H}_k^d\boldsymbol{p}_1\,. \tag{2}$$

The feature volumes are constructed by warping feature maps from source images to the reference camera frustum

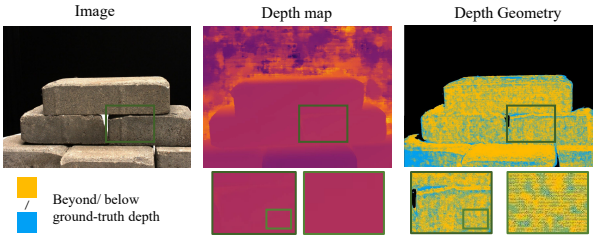


Figure 6. **Visualization of the depth geometry with saddle-shaped cells.** We colored the pixels whose estimated depth value is beyond/below the ground truth with orange/blue.

per to pixels and their transformed pixels at different depth hypotheses per to Eq. (2). By regularizing the cost volume generated by measuring the similarity of the feature volumes, a probability volume $\boldsymbol{P} \in \mathbb{R}^{M \times H^{'} \times W^{'}}$ can be obtained. The depth of a pixel located at coordinates (x, y) in the reference view can be obtained by using the following.

$$\overline{\boldsymbol{D}}(x, y) = \sum_{i=1}^{M} d_i\boldsymbol{P}(i, x, y)\,. \tag{3}$$

Loss function. In common MVS methods, L_1 loss is used to supervise the estimated depth map $\overline{\boldsymbol{D}}$ with

$$L_{est}(\overline{\boldsymbol{D}}, \boldsymbol{D}_g) = L_1(\overline{\boldsymbol{D}}, \boldsymbol{D}_g)\,, \tag{4}$$

where $\boldsymbol{D}_g$ is the ground-truth depth map. L_{est} aims to minimize the difference between the estimated depth map and the ground-truth depth map, thus reducing the estimated bias. However, it lacks the ability to enforce the geometry of the estimated depth, let alone predict a saddle-shaped depth map. Furthermore, the objective of the saddle-shaped cell depth map is inconsistent with the objective of L_{est}, which encourages the estimated depth map to approach the smooth depth map of the truth of the ground.

4.2. Dual-Depth

Aiming at an oscillating depth geometry with more saddle-shaped cells, we choose to predict two depth values for each pixel. If the dual depth is distributed on either side of the ground-truth depth, a heuristic selection strategy can achieve the target geometry.

Specifically, we generate two probability distributions for each pixel and use them to generate two corresponding depth maps $\boldsymbol{D} \in \mathbb{R}^{2 \times H^{'} \times W^{'}}$ (see Eq. (3)). To ensure the accuracy of the independently predicted dual depth, we take a L_1 loss to supervise their predicted values, as in previous works. Intuitively, without adding constraints on the joint distribution of the dual depth, the resulting prediction distribution is disordered. Therefore, we propose another novel loss to constrain the two depths to be symmetrically distributed around the ground truth.

$$\begin{aligned} L_{int}(\boldsymbol{D}, \boldsymbol{D}_g) = L_1(|\texttt{max}(\boldsymbol{D}) - \texttt{min}(\boldsymbol{D})|, \\ max(|\texttt{max}(\boldsymbol{D}) - \boldsymbol{D}_g|, |\texttt{min}(\boldsymbol{D}) - \boldsymbol{D}_g|))\,, \end{aligned} \tag{5}$$

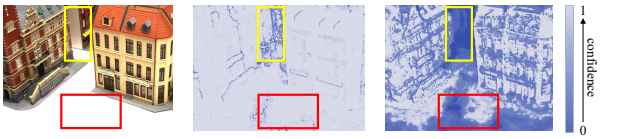


Image (a) Via variance (b) Via Dual-Depth

Figure 7. **Comparison of confidence maps**. The confidence map via Dual-Depth provides more accurate confidence for occluded regions or invalid backgrounds. The lighter region indicates more higher confidence.

where $|\cdot|$ indicates the absolute distance, $\texttt{max}(\cdot)$ and $\texttt{min}(\cdot)$ takes the maximum and minimum value along the first dimension, *e.g.* $\texttt{max}(\boldsymbol{D}) = max(\boldsymbol{D}[1,:,:], \boldsymbol{D}[2,:,:])$. L_{int} encourages estimated bias is no larger than $|\texttt{max}(\boldsymbol{D}) - \texttt{min}(\boldsymbol{D})|$, such that the interval increases as estimated bias increasing, which guarantees that dual-depth is distributed on either side of the ground-truth depth. If $\texttt{max}(\boldsymbol{D}) = \texttt{min}(\boldsymbol{D}) = \boldsymbol{D}_g$, L_{int} reaches the minimal value, suggesting an unbiased depth estimate and is consistent with the objective of L_{est}.

When the true depth value lies between the predicted dual-depth, we propose a checkerboard selection strategy to choose the appropriate depth prediction value for each pixel. Specifically, we alternate between selecting the maximum and minimum predicted depth values, creating a distribution that resembles a checkerboard. As shown in Fig. 5(a), the depth of pixel (x, y) is determined by

$$\boldsymbol{D}_c(x,y) = \begin{cases} \texttt{min}(\boldsymbol{D})(x,y)\,, & x\%2 == y\%2 \\ \texttt{max}(\boldsymbol{D})(x,y)\,, & otherwise \end{cases}, \quad (6)$$

which generates an oscillating depth map $\boldsymbol{D}_c$. As shown in Fig. 6, the depth map obtained by the dual-depth method achieves the geometry composed of saddle-shaped cells. The depth map within the box is smooth, indicating that the predicted values are close to the ground truth. At the same time, its corresponding depth geometry presents a saddle-shaped form, which is consistent with our expectations.

However, the above approach might carry a potential risk of increasing depth prediction errors when the true depth value at (x, y) is not within the range of $\texttt{min}(\boldsymbol{D})(x, y)$ and $\texttt{max}(\boldsymbol{D})(x, y)$. To address this issue, we propose using Cascade Dual-Depths, which will be illustrated in the next section.

4.3. Cascade Dual-Depths

Despite the fact that the encouraging estimated bias is not larger than $|\texttt{max}(\boldsymbol{D}) - \texttt{min}(\boldsymbol{D})|$ in the dual depth, the uncovered issue occurs when the estimated bias is too large, which is beyond the range of $|\texttt{max}(\boldsymbol{D}) - \texttt{min}(\boldsymbol{D})|$. The reason is that the fixed range of depth hypotheses $\alpha_2 - \alpha_1$ leads to a large depth estimation bias. Intuitively, when a pixel's estimated depth is unreliable, the range of depth hypotheses should be increased to ensure that the ground truth is included in the searching space. In contrast, the range can be appropriately narrowed for more reliable estimates. For example, UCS-Net [7] leverages the variance of the probability distribution to reflect uncertainty and dynamically adjusts the range of depth hypotheses for the corresponding sampling depths, resulting in smaller estimation biases. Inspired by it, we attempted to utilize uncertainty estimates to adaptively adjust the search range of depth hypotheses.

We first adopt the variance of the probability distribution to obtain a confidence map, similar to UCS-Net (as shown in Fig. 7(a)). However, the confidence map obtained through variance tends to predict similar confidence levels for most regions, making it unreliable for predicting confidence levels in weakly textured areas or at edges. Additionally, there exists inference conflict between the two corresponding confidence values since each pixel predicts the dual-depth. Therefore, we need to find an appropriate way to represent the uncertainty of the dual-depth, not solely relying on the probability distribution.

In daily life, when measuring an object with a ruler, people usually take two measurements and compare the results. If there is a large difference between the two measurements, the precision of the measurement is considered low. Similarly in dual-depth estimation, if the difference between the maximum and minimum depth values predicted by a pixel is large, we consider the estimation bias to be large. Consequently, the depth searching space for this pixel will be enlarged in the next iteration. To adaptively adjust the range of the depth hypotheses, we use the absolute distance between $\texttt{max}(\boldsymbol{D})(x, y)$ and $\texttt{min}(\boldsymbol{D})(x, y)$ as the boundary of the depth hypotheses, as shown in Fig. 5(b). Given the range of depth hypotheses, we can construct feature volumes, cost volumes, and probability distributions as described in Sec. 4.1. Following the principles outlined in Sec. 4.2, we calculate the refined dual depth $\boldsymbol{D}'$, which is then used to obtain the final depth $\boldsymbol{D}_r$ according to Eq. (6).

Confidence Map. The confidence map is used in the fusion process to mask out pixels with substantially deviated depths using a threshold, thereby preventing their projection into the 3D space. In this paper, the confidence is given by

$$\boldsymbol{F}(x,y) = 2\texttt{sigmoid}(\frac{1}{\boldsymbol{U}(x,y)}) - 1 \in (0,1)\,, \quad (7)$$

where $\boldsymbol{U}(x, y) = |\texttt{max}(\boldsymbol{D})(x, y) - \texttt{min}(\boldsymbol{D})(x, y)|$. The confidence map is shown in Fig. 7(b).

4.4. DMVSNet

To embed our double-dual depths method into the multi-view stereo (MVS) task, we propose a coarse-to-fine MVS framework, named DMVSNet. As shown in Fig. 8, the dual depths structure is incorporated into the depth regression stage through differentiable warping.

Specifically, we adopt a Feature Pyramid Network (FPN) [21] to extract multiscale features like

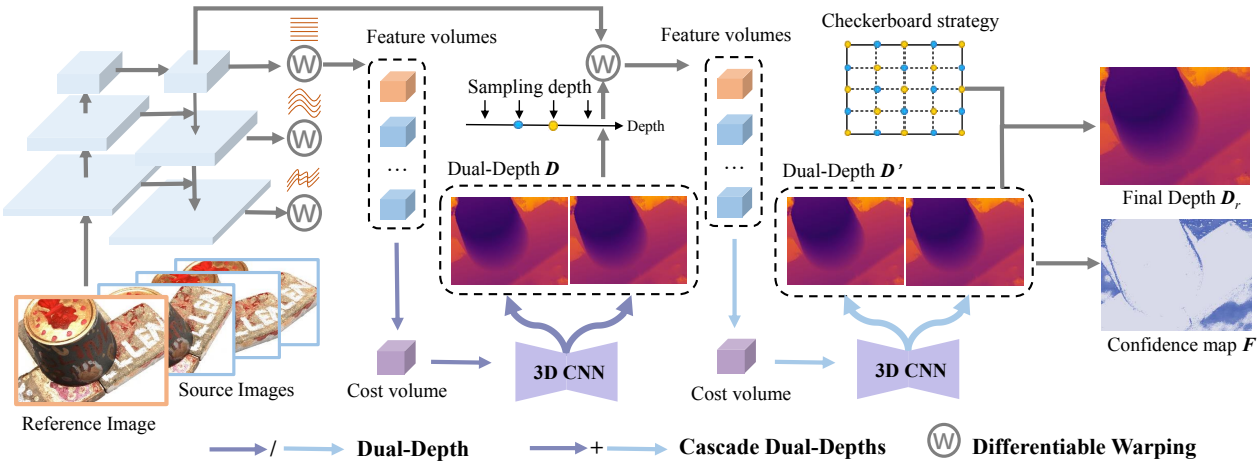


Figure 8. **Architecture details of DMVSNet.** Our backbone adopts a classical coarse-to-fine framework. We employ a shared feature pyramid network for extracting feature maps, and then warp them to obtain a cost volume. Next, we use a 3D convolutional neural network (CNN) to generate two depth maps, denoted as Dual-Depth $\boldsymbol{D}$. By resetting the depth sampling range using $\boldsymbol{D}$, we repeat the depth prediction process to obtain another dual-depth $\boldsymbol{D}^{'}$, which is utilized to construct the final depth $\boldsymbol{D}_r$ with the checkerboard strategy. Per to Eq. (7), the corresponding confidence map $\boldsymbol{F}$ can be calculated.

Method	Years	ACC.(mm)↓	Comp.(mm)↓	Overall(mm)↓
Gipuma [15]	ICCV2015	**0.283**	0.873	0.578
COLMAP [27]	CVPR2016	0.400	0.664	0.532
SurfaceNet [17]	ICCV2017	0.450	1.040	0.745
MVSNet [39]	ECCV2018	0.396	0.527	0.462
P-MVSNet [22]	ICCV2019	0.406	0.434	0.420
R-MVSNet [40]	CVPR2019	0.383	0.452	0.417
Point-MVSNet [6]	ICCV2019	0.342	0.411	0.376
CasMVSNet [16]	CVPR2020	0.325	0.385	0.355
CVP-MVSNet [38]	CVPR2020	<u>0.296</u>	0.406	0.351
UCS-Net [7]	CVPR2020	0.338	0.349	0.344
AA-RMVSNet [34]	ICCV2021	0.376	0.339	0.357
UniMVSNet [24]	CVPR2022	0.352	0.278	0.315
transMVSNet [9]	CVPR2022	0.321	0.289	**0.305**
MVSter [32]	ECCV2022	0.350	<u>0.276</u>	<u>0.313</u>
DMVSNet	-	0.349	<u>0.276</u>	<u>0.313</u>
DMVSNet⋆	-	0.338	**0.272**	**0.305**

Table 3. Results on DTU. We report our results with a vanilla checking strategy and dynamic checking strategy⋆. The best performance is in boldface and the second best is underlined.).

CasMVSNet [16], and double the output channels to obtain feature maps for a cascading dual-depth. Then, we construct feature volumes by warping feature maps via sampled depth hypotheses. The feature volumes are then aggregated into cost volumes using a similarity metric, such as the inner-product-based metric. A 3-D CNN is utilized to transform the cost volume into the probability distribution. To obtain dual-depth $\boldsymbol{D}$, we double the 3-D CNN to obtain two probability distributions and generate depth maps using Eq. (3). We repeat this process with adaptively sampled depth hypotheses to generate refined dual-depth $\boldsymbol{D}^{'}$, which is utilized to construct the final depth $\boldsymbol{D}_r$ with the checkerboard selection strategy.

Training Loss. We adopt L_{est} and L_{int} to respectively reduce estimation bias and interpolation bias. Since the interpolated biases result from the depth error of sub-pixels, we additonally supervise the depths of sub-pixels at coordinates $(x+0.5, y+0.5)$. The final loss function is

$$\begin{aligned} L = & L_{est}(\boldsymbol{D}, \boldsymbol{D}_g) + L_{int}(\boldsymbol{D}, \boldsymbol{D}_g) + L_{sub}(\boldsymbol{D}, \boldsymbol{D}_g) \\ & + L_{est}(\boldsymbol{D}^{'}, \boldsymbol{D}_g) + L_{int}(\boldsymbol{D}^{'}, \boldsymbol{D}_g) + L_{sub}(\boldsymbol{D}^{'}, \boldsymbol{D}_g) \end{aligned}, \tag{8}$$

where $L_{sub}(\boldsymbol{D}), \boldsymbol{D}_g) = L_1(\texttt{sub.}(\texttt{con.}(\boldsymbol{D}), \boldsymbol{D}_g))$, $\texttt{con.}(x)$ suggests constructing a depth map with the checkerboard selection strategy,and $\texttt{sub.}$ indicates taking the sub-pixels at coordinates $(x+0.5, y+0.5)$.

5. Experiments

Datasets. We conducted training and evaluation of our models using the DTU dataset [1]. We fine-tune the BlendedMVS [41] and subsequently evaluate on Tanks and Temples [19] for generalization. The DTU dataset comprises 124 indoor scenes that were captured under controlled camera and lighting conditions. To ensure consistency, We adapt the same training and evaluation split as MVSNet [39]. On the other hand, the BlendedMVS dataset is a synthetic dataset with 113 scenes that simulate both indoor and outdoor conditions. We follow the UniMVS and adapt the same training and validation split. Lastly, the Tanks and Temples benchmark comprises scenes captured in a complex and realistic environment and serves as an online benchmark. It is divided into intermediate and advanced sets based on the level of difficulty.

Metrics. For the depth metric, we utilize the absolute distance of the disparity between the predicted depth and the ground truth depth, commonly referred to as depth error or estimated bias. As for the 3-D representation, we report the standard metrics, namely accuracy, completeness, and overall score, using the official evaluation toolkit.

Method	Years	Intermediate(%)↑									Advanced(%)↑						
		Mean	Fam.	Fra.	Hor.	Lig.	M60.	Pan.	Pla.	Tra.	Mean	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
PointMVS [6]	ICCV2019	48.27	61.79	41.15	34.20	50.79	51.97	50.85	52.38	43.06	-	-	-	-	-	-	-
PatchmatchNet [31]	CVPR2021	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	32.31	23.69	37.03	30.04	41.80	28.31	32.29
CVP-MVSNet [38]	CVPR2020	54.03	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54	-	-	-	-	-	-	-
CasMVSNet [16]	CVPR2020	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	31.12	19.81	38.46	29.10	43.87	27.36	28.11
UCS-Net [7]	CVPR2020	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	-	-	-	-	-	-	-
D2HC-RMVSNet [37]	ECCV2020	59.20	74.69	56.04	49.42	60.08	59.81	59.61	60.04	53.92	-	-	-	-	-	-	-
AA-RMVSNet [34]	ICCV2021	61.51	77.77	59.53	51.53	**64.02**	64.05	59.47	60.85	55.50	33.53	20.96	40.15	32.05	46.01	29.28	32.71
EPP-MVSNet [23]	ICCV2021	61.68	77.86	60.54	52.96	62.33	61.69	60.34	**62.44**	55.30	35.72	21.28	39.74	35.34	49.21	30.00	38.75
UniMVSNet [24]	CVPR2022	64.36	81.20	66.43	53.11	63.46	**66.09**	**64.84**	62.23	57.53	38.96	28.33	44.36	39.74	52.89	33.80	34.63
transMVSNet [9]	CVPR2022	63.52	80.92	65.83	56.94	62.54	63.06	60.00	60.20	**58.67**	37.00	24.84	44.59	34.77	46.49	34.69	36.62
MVSter [32]	ECCV2022	60.92	80.2	63.51	52.30	61.38	61.47	58.16	58.98	51.38	37.53	26.68	42.14	35.65	49.37	32.16	39.19
Ours	-	**64.66**	**81.27**	**67.54**	**59.10**	63.12	64.64	64.80	59.83	56.97	**41.17**	**30.08**	**46.10**	**40.65**	**53.53**	**35.08**	**41.60**

Table 4. Quantitative results on Tanks and Temples benchmark. We report the F-score metric and "Mean" refers to the average F-score of all scenes. The best performance is in **boldface** and the second best is underlined.

Implementation Details. We performed network optimization over 16 epochs using a learning rate of 0.001, with semi-decay occurring in epochs 10, 12, and 14. During the evaluation process, we used the final depth $\boldsymbol{D}_r$ to calculate the metrics and perform the fusion. To ensure consistency with DTU standards, the input images were resized to 1152×864, and the number of input images was set to 5. Before evaluating on Tanks and Temples, our network underwent a fine-tuning process on BlendedMVS for 10 epochs, following established protocols. For training and evaluation, the number of input images was set to 9 and 11, respectively. Consistent with previous research [37], we also adopted the dynamic check strategy for fusion.

5.1. Results on DTU

To assess the effectiveness of our proposed approach, we adopt a vanilla checking strategy. Drawing inspiration from MVSNet [39], we use confidence maps and geometric constraints for depth filtering. Specifically, we set the probability threshold and the minimum number of consistent views to 0.3 and 5, respectively. Our baseline comprises traditional and learning-based MVS methods. Compared with previously published works, our vanilla check strategy achieves the highest completeness and the second-highest overall results, as depicted in Table 3. It is important to note that the fusion strategies used in previous works are not standardized, such as transMVSNet [9] used the dynamic check strategy [37]. For a fair comparison, we also present our method result with the dynamic check strategy*, which outperforms all other methods in terms of the overall performance.

5.2. Results on Tanks and Temples

Following the common setting, we evaluate the efficacy of our method in terms of generalization using the Tanks and Temples online benchmark, after fine-tuning on BlendedMVS. As shown in Table 4, we report the F-score that is defined as a harmonic mean of accuracy and completeness. Our method achieves SOTA performance in both intermediate and advanced sets. In the advanced set, a more challenging subset of Tanks and Temples, we achieve the state-of-the-art results in every scene, with an 5.6% improvement over the second-best result. It is worth noting that the accuracy of depth estimation for a scene is inversely proportional to the difficulty of that scene. Dual-depth prediction may not mitigate the bias in estimated depth, but it reduces interpolated bias via the depth geometry, indicating that it better fits for scenes with inaccurate depth estimation.

6. Ablation Study

Which part works. We first report the results of the ablation study in Table 5. The baseline choose the CasMVSNet framework. By introducing Dual-Depth, we observe a significant 7.5% enhancement in the 3-D representation. The Cascade Dual-Depths bring another 1.6% improvement over Dual-Depth in 3-D representation. It should be noted that the advancement in depth error (*i.e.*, the estimated bias), is not the primary factor that contributes to the improved 3-D representation achieved by Dual-Depth. Instead, the saddle-shaped geometry plays a crucial role.

To further illustrate the importance of the checkerboard strategy, we conduct an experiment in Table 6. We fixed the parameters of the model and only changed the depth selection strategy. The results showed that while the checkerboard strategy had little effect on the accuracy of depth prediction, it significantly improved 3-D reconstruction metrics. When the checkerboard strategy was not used, i.e., only one side with two predicted depth values was selected (such as selecting the minimum depth), the depth values of adjacent pixels were closer due to the lack of constraints. This made adjacent pixels more likely to be on the same side of the true surface, which disrupted some saddle-shaped cells that should have existed in the depth map using the checkerboard strategy. Therefore, the checkerboard strategy is indispensable for saddle-shaped depth geometry.

What Dual-Depth does. To understand the importance

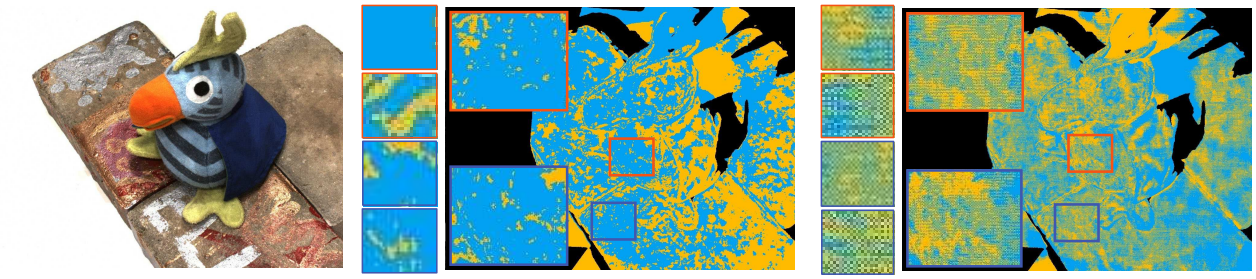
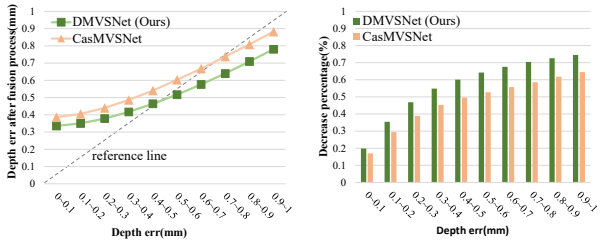


Image CasMVSNet DMVSNet (Ours)

Figure 9. **The comparison of the depth geometries.** We visualize the depth geometry by coloring the pixel whose estimated depth is beyond the ground truth with orange and the others with blue.

Settings	Depth Err.	ACC.	Comp.	Overall
w.o. Dual-Depth	10.47	0.366	0.324	0.345
w.o. Cascade Dual-Depths	10.03	0.352	0.288	0.320
DMVSNet (Ours)	**9.12**	**0.349**	**0.276**	**0.313**

Table 5. Ablation studies on DTU. We apply the same fusion setting as in Sec. 5.1



(a) Depth err. before/after fusion (b) Decrease percentage of depth err.

Figure 10. **How Dual-Depth works.** Compared with side prediction, the estimated biases of dual-depth decrease more(a). The ratios of depths whose estimated bias decreases are shown in (b) and ratios of Dual-Depth are higher in all intervals.

of Dual-Depth, we conducted an interesting experiment in scene 29 of DTU, as shown in Fig. 10. We analyzed the pixel depth error changes after fusion within the range of $0 \sim 1$ mm. A reference line is characterized by zero-change depth error. As shown in Fig. 10(a), our DMVSNet achieves smaller depth errors after fusion compared with CasMVSNet, indicating that the Cascade Dual-Depths can help reduce initial depth errors during fusion. As shown in Fig. 10(b), we illustrate the percentage of pixels with reduced errors after fusion for each depth error interval. Our method achieves a higher proportion of depth error reduction in all error intervals, which explains our superior performance in Fig. 10(a). We display the geometry of Dual-Depth Prediction and the baseline in Fig. 9, to understand what Dual-Depth Prediction does. The Dual-Depth Prediction presents depth predictions at the pixel level in an oscillating pattern, while the baseline method tends to predict depth regionally smooth. Specifically, the Dual-Depth Prediction generates a checkerboard-like depth geometry that approximates the expected geometry composed of saddle-shaped cells.

Why is Dual-Depth. needed. To better understand the contribution of the method details, we conducted additional

Settings	Depth Err.	ACC.	Comp.	Overall
w.o. checkerboard strategy	10.05	0.386	0.321	0.354
w. checkerboard strategy (Ours)	**10.03**	**0.352**	**0.288**	**0.320**

Table 6. Evaluations w.o./w. checkerboard strategy. We take the fixed model to obtain depth maps w.o./w. checkerboard strategy. With similar qualities of depth estimation, the checkerboard strategy achieves a significant improvement in the performance of 3D reconstruction.

Settings	Depth Err.	ACC.	Comp.	Overall
w.o. L_{int} and L_{sub}	11.09	0.356	0.318	0.337
w.o. double branches	9.26	0.356	0.297	0.327
Cascade Dual-Depths (Ours)	**9.12**	**0.349**	**0.276**	**0.313**

Table 7. Additional experiments.

ablation studies in Table 7. Firstly, we removed the constraints on the geometric shape of the depth estimation (by not using L_{int} and L_{sub}), which resulted in decreased depth and 3-Dreconstruction quality. Then, we did not apply the double branches in the 3-D CNN. Although it had a similar depth estimation accuracy, the 3-D reconstruction results were worse. They demonstrate the necessity of constraints on depth geometry and additional structural design.

7. Conclusion

In this article, we present a novel perspective to enhance the performance of learning-based MVS by incorporating constraints of the depth geometry. We demonstrate that the proposed saddle-shaped cells outperform other cells in terms of qualitative and quantitative measures. We introduce the Dual-Depth, checkerboard selection strategy, and Cascade Dual-Depths to implement the saddle-shaped depth geometry. By integrating them into a coarse-to-fine framework, we develop a novel DMVSNet approach. Extensive experiments prove that DMVSNet achieves superior performance in 3-D reconstruction and the importance of depth geometry.

Acknowledgments. This work was funded in part by the National Natural Science Foundation of China (Grant No. U1913602) and supported by the Huawei Technologies CO., LTD.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *Int. J. Comput. Vis.*, 120:153–168, 2016.

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.

[3] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Proc. Eur. Conf. Comput. Vis.*, pages 766–779. Springer, 2008.

[4] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer: Multi-view stereo by learning robust image features and temperature-based depth. *Transactions of Machine Learning Research*, 2023.

[5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 5410–5418, 2018.

[6] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1538–1547, 2019.

[7] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 2524–2534, 2020.

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, volume 1, pages 886–893. Ieee, 2005.

[9] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 8585–8594, 2022.

[10] Thomas Ebner, Ingo Feldmann, Sylvain Renault, Oliver Schreer, and Peter Eisert. Multi-view reconstruction of dynamic real-world objects and their integration in augmented and virtual reality applications. *Journal of the Society for Information Display*, 25(3):151–157, 2017.

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Proc. Adv. Neural Inf. Process. Syst.*, 27:2366–2374, 2014.

[12] Sudeep Fadadu, Shreyash Pandey, Darshan Hegde, Yi Shi, Fang-Chieh Chou, Nemanja Djuric, and Carlos Vallespi-Gonzalez. Multi-view fusion of sensor data for improved perception and prediction in autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2349–2357, 2022.

[13] Pascal Fua and Yvan G Leclerc. Object-centered surface reconstruction: Combining multi-image stereo and shading. *Int. J. Comput. Vis.*, 16(ARTICLE):35–56, 1995.

[14] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 873–881, 2015.

[15] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 873–881, 2015.

[16] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 2495–2504, 2020.

[17] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2307–2315, 2017.

[18] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proc. Eur. Conf. Comput. Vis.*, pages 573–590, 2018.

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

[20] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 2811–2820, 2018.

[21] 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.

[22] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 10452–10461, 2019.

[23] Xinjun Ma, Yue Gong, Qirui Wang, Jingwei Huang, Lei Chen, and Fan Yu. Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 5732–5740, 2021.

[24] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 8645–8654, 2022.

[25] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 5506–5514, 2016.

[26] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 4938–4947, 2020.

[27] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 4104–4113, 2016.

[28] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Proc. Eur. Conf. Comput. Vis.*, pages 501–518. Springer, 2016.

[29] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. In *Int. J. Comput. Vis.*, pages 1067–1073. IEEE, 1997.

[30] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 8922–8931, 2021.

[31] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 14194–14203, 2021.

[32] Xiaofeng Wang, Zheng Zhu, Guan Huang, Fangbo Qin, Yun Ye, Yijia He, Xu Chi, and Xingang Wang. Mvster: epipolar transformer for efficient multi-view stereo. In *Proc. Eur. Conf. Comput. Vis.*, pages 573–591. Springer, 2022.

[33] Yiran Wang, Zhiyu Pan, Xingyi Li, Zhiguo Cao, Ke Xian, and Jianming Zhang. Less is more: Consistent video depth estimation with masked frames modeling. In *Proc. ACM Int. Conf. Multimedia*, pages 6347–6358, 2022.

[34] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 6187–6196, 2021.

[35] Chenglei Wu, Bennett Wilburn, Yasuyuki Matsushita, and Christian Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 969–976. IEEE, 2011.

[36] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 611–620, 2020.

[37] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *Proc. Eur. Conf. Comput. Vis.*, pages 674–689. Springer, 2020.

[38] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 4877–4886, 2020.

[39] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proc. Eur. Conf. Comput. Vis.*, pages 767–783, 2018.

[40] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 5525–5534, 2019.

[41] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020.

[42] Xinyi Ye, Weiyue Zhao, Hao Lu, and Zhiguo Cao. Learning second-order attentive context for efficient correspondence pruning. In *Proc. AAAI Conf. Artificial Intell.*, volume 37, pages 3250–3258, 2023.

[43] Weiyue Zhao, Hao Lu, Zhiguo Cao, and Xin Li. A2b: Anchor to barycentric coordinate for robust correspondence. *Int. J. Comput. Vis.*, pages 1–25, 2023.

[44] Weiyue Zhao, Hao Lu, Xinyi Ye, Zhiguo Cao, and Xin Li. Learning probabilistic coordinate fields for robust correspondences. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–17, 2023.