# Bi-ClueMVSNet: Learning Bidirectional Occlusion Clues for Multi-View Stereo

Zhe Zhang[1], Yuxi Hu[2], Huachen Gao[1], Ronggang Wang[1]*

[1]School of Electronic and Computer Engineering, Peking University, China
[2]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

***Abstract*—Deep learning-based multi-view stereo (MVS) methods have achieved promising results in recent years. However, very few existing works take the occlusion issues into consideration, leading to poor reconstruction results on the boundaries and occluded areas. In this paper, the Bidirectional Occlusion Clues-based Multi-View Stereo Network (Bi-ClueMVSNet) is proposed as an end-to-end MVS framework that explicitly models the occlusion obstacle for depth map inference and 3D modeling. To this end, we use bidirectional projection for the first time to reduce the propagation and accumulation of incorrect matches and build the occlusion-enhanced network to further advance the representational ability from 2D visibility maps to 3D occlusion clues. As for depth map estimation, we combine the characteristics of both regression and classification approaches to propose the adaptive depth map inference strategy. Besides, the robustness of the training process is further guaranteed and elevated by the occlusion clues-based loss function. The proposed method significantly improves the accuracy of depth map inference in boundaries and heavily occluded areas and brings the overall quality of the reconstructed point cloud to a new altitude. Extensive experiments are performed on DTU, Tanks and Temples, and BlendedMVS datasets to demonstrate the persuasiveness of the proposed framework.**

## I. Introduction

Multi-View Stereo (MVS) is a method for reconstructing the three-dimensional (3D) model of a scene from multiple overlapping images, which is one of the fundamental problems in computer vision and robotics ecosystems extensively studied for decades. In recent years, intensive research has been devoted to deep learning-based MVS methods which have shown competitive results compared to previous traditional methods [1]. Traditional MVS often fails to obtain accurate reconstruction results in the case of texture starvation, texture repetition, or illumination changes. In contrast, learning-based methods [2]–[8] usually extract deep features using neural networks, which implicitly introduce global information such as specularity and reflection priors. Furthermore, learning-based methods usually apply 3D CNN for the cost volume regularization, which is more efficient and robust than hand-designed cost matches strategies in traditional methods.

Although learning-based methods have shown promising results, there is a widespread unsolved problem that corresponding matches are not visible among all views and incorrect matches will influence the reliability of the reconstruction. In traditional MVS methods, taking COLMAP [11] as an example, the occlusion issue can be modeled under the probabilistic framework. However, very few learning-based methods have taken the occlusion problem into consideration.

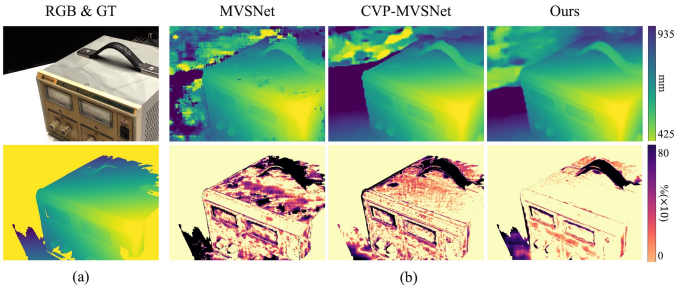



Fig. 1. **Comparison between the baseline methods** [9] [10] **and our Bi-ClueMVSNet.** (a) One view of RGB input and corresponding depth ground truth; (b) estimated depth maps (top) and error maps (bottom), all errors are magnified by 10 times for visualization and darker means larger error.

Motivated by the aforementioned challenges, this work proposes an end-to-end learning-based MVS network by explicitly modeling the occlusion issues. Previous MVS methods use homography for warping source views to reference view only, but we broaden the horizon of bidirectional projection. And apart from two-step aggregating [12] [13] from pair-wise cost volume, we directly construct the occlusion-aware cost volume from original warped features which lessens the propagation and accumulation of incorrect matches at the root. Besides, we further extend the 2D visibility maps to 3D occlusion clues using our proposed occlusion-enhanced network, and adaptively combine regression and classification approaches for depth map inference using occlusion clues. The training loss is also redesigned for robust and better convergence.

The proposed method is evaluated on DTU and Tanks and Temples datasets. Our method achieves very competitive results compared with previous methods and gets significantly better reconstruction results in boundaries and areas with severe occlusion. Ablation studies demonstrate the considerable improvement brought by the system we designed. In summary, the primary contributions of Bi-ClueMVSNet are as follows:

- We introduce bi-directional projection and occlusion-aware aggregation for constructing the cost volume,

*Ronggang Wang is the corresponding author (rgwang@pkusz.edu.cn).

This work is supported by the National Natural Science Foundation of China U21B2012 and 62072013, Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents RCJC20200714114435057, Shenzhen Research Projects of 201806080921419290.

2023 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-8867-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IJCNN54540.2023.10191325

which reduces the risk of propagation and accumulation of incorrect matches at the root.
- We propose the occlusion-enhanced network to transfer 2D visibility maps to 3D occlusion clues, improving the representational ability of regularization.
- We combine regression and classification strategies for adaptive depth map inference and build occlusion clues-based loss for robust convergence.

## II. Related Work

Our work focuses on occlusion-aware depth estimation to boost the performance of learning-based multi-view stereo methods. Therefore, some multi-view stereo techniques and occlusion issues are discussed below.

### A. Traditional Multi-View Stereo

Existing MVS methods can be classified into three categories. Point cloud-based methods [14] [15] process points in 3D space to iteratively densify the results. It is not always satisfactory due to their demanding requirements for the quality of extracted feature points. Voxel-based methods [16] [17] compute the bounding box containing the scene to find voxels near the surface. However, the memory consumption increases with the voxel resolution which determines the reconstruction accuracy. In contrast, depth map-based [18] [19] methods decouple the task of reconstruction into depth estimation on every image [1] and then obtain the point cloud through depth map fusion. Therefore, it has shown more flexibility in reconstructing the three-dimensional geometry of scenes [20].

### B. Learning-Based Multi-View Stereo

Recently, learning-based methods have shown great potential to solve MVS problems. MVSNet [9] aims to learn the depth map for each view by constructing a cost volume. They obtain the geometry by fusing the estimated depth maps from multiple views. To reduce memory consumption, R-MVSNet [21] adopts GRUs to regularize the cost volume in a sequential manner but this leads to increased run-time.

Current research targets to estimate high-resolution depth maps while improving inference efficiency. CasMVSNet [22] and UCS-Net [23] adopt cascade cost volumes and estimate the depth map coarse-to-fine. CVP-MVSNet [10] constructs an image pyramid and a cost volume pyramid. Similar coarse-to-fine frameworks [4] [6] [8] [24]–[33] are used to lower the GPU cost of 3D regularization or increase the depth quality. However, these methods all apply a variance-based cost metric, assuming that a given pixel is visible in all input images. In fact, many pixels only appear at some baseline angles and are occluded in others. As a result, an increasing number of input images lead to even worse depth estimation.

### C. Occlusion in Multi-View Stereo

Occlusion and visibility estimation are very core and difficult tasks, whether for basic computer vision tasks or 3D reconstruction and multi-view geometry. The previous methods mainly include the heuristic cost threshold method [15] [34] [35] and the depth-visibility joint estimation method [36]. However, most of these approaches are modeled under the framework of probability theory, which is difficult to integrate with the existing learning-based pipeline.

The existing frameworks generally model occlusion issues as an implicit representation and hope it to be processed through regularization. Vis-MVSNet [12] and PVSNet [13] explicitly model the visibility information at the pixel level. The visibility information is obtained through feature aggregation and matching between image pairs to aggregate the pair-wise cost volume. However, the cost volume obtained by two-by-two aggregation itself already contains erroneous matching information due to occlusion. This error can propagate and accumulate through the network, and it is difficult to solve the occlusion problem from the root. Meanwhile, it is difficult to constrain the depth dimension using the 2D visibility map, which restricts the ability of visibility processing. Therefore, we abandon the aggregated pair-wise cost volume and directly aggregate from the most original warped image features, and expand the 2D occlusion map into 3D occlusion clues through the occlusion enhancement network to further advance the effect of attention weights.

## III. Methodology

The overall architecture of the Bi-ClueMVSNet is illustrated in Fig. 2. Given a reference image $I_{ref} = I_0$ and arbitrary source images $I_{src} = \{I_i\}_{i=1}^{N}$, our proposed network predicts the corresponding depth map $\tilde{D}_0$ aligned with $I_0$, and explicit model the occlusion issue from the root.

The deep image feature extraction is first described in Sec. III-A. Then we introduce the proposed Occlusion-Aware Cost Volume construction in Sec. III-B and its regularization in Sec. III-C. Finally, we describe the Adaptive Depth Map Inference strategy in Sec. III-D, and the Occlusion Clues-based Loss function during training in Sec. III-E.

### A. Multi-scale Feature Pyramid

As raw images and traditional features suffer from illumination changes as well as reflections. Multi-scale Feature Pyramid Network (M-FPN) [37] is used for extracting deep features from $\{I_i\}_{i=0}^{N}$. The general works usually use high-resolution images as input and downsample them in the network [9] [21] [38]. Differently, we use a 9-layer 2D CNN to extract full-scale features from the low-resolution images to avoid information degradation [10] [39]. The pyramid network contains $L$-level, and the output deep features are denoted as $\{F_i^l\}_{i=0}^{N} \in \mathbb{R}^{\frac{H}{2^l} \times \frac{W}{2^l} \times M}$, where $H$ and $W$ are the height and width of the input images respectively, and $M = 24$ is the number of feature channels used in our experiments.

### B. Occlusion-Aware Cost Volume

Given the above deep features, most end-to-end learning-based MVS frameworks follow the pipeline of MVSNet [9] to construct the cost volume. As for our pyramid-based framework, let's consider the $k$th-level of the pyramid for simplicity. Cost volume $C \in \mathbb{R}^{M \times D \times H \times W}$ identifies the matching cost

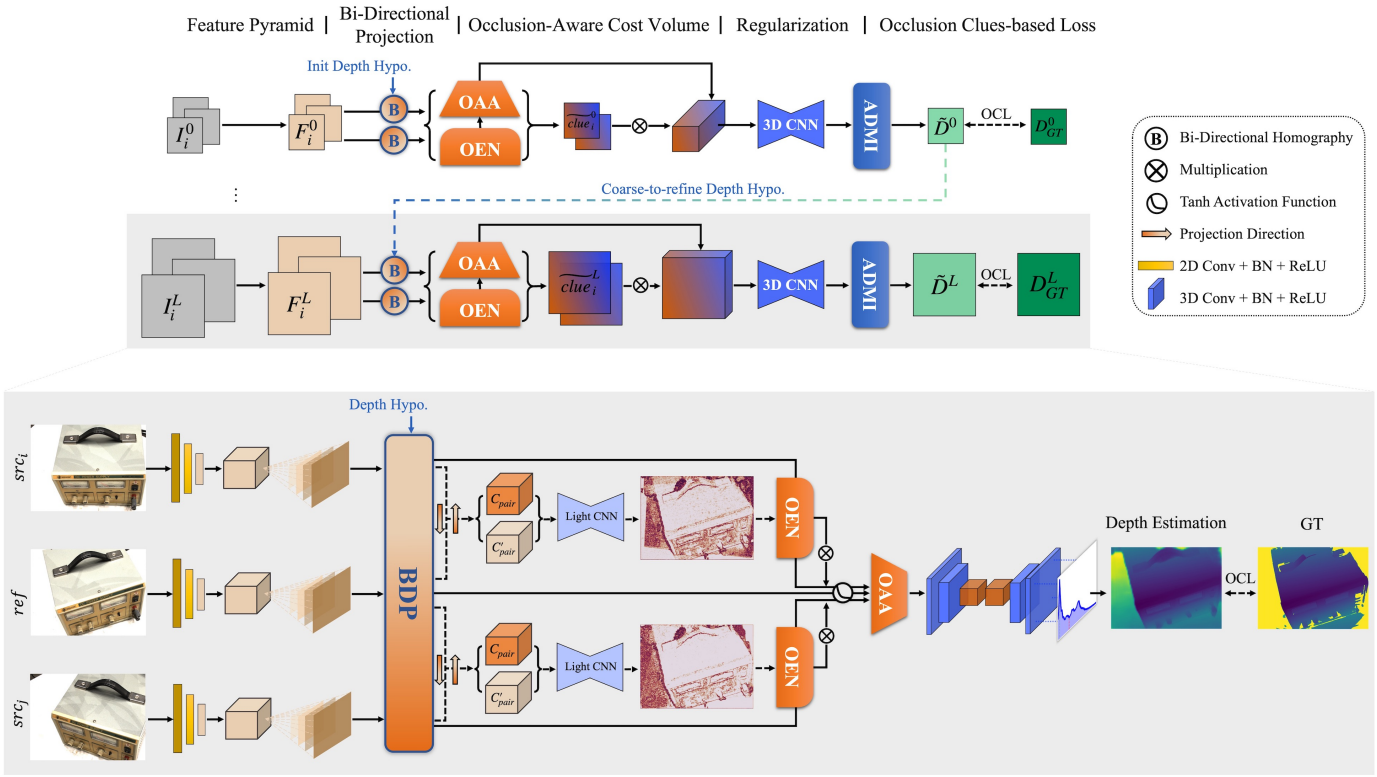



Fig. 2. **Network Structure**: the coarse-to-fine pipeline (top) and detailed architecture (below). Multi-scale Feature Pyramid Network (M-FPN) is first used to extract depth features from input images $I_i$. We then construct the cost volume from the proposed Bi-Directional Projection (BDP) and Occlusion-Aware Aggregation (OAA) modules and combining with the Occlusion-Enhanced Network (OEN) for regularization. At last, we use the Adaptive Depth Map Inference (ADMI) strategy to estimate depth $\tilde{D}$, and Occlusion Clues-based Loss (OCL) is used for supervised training.

of spatial points under the feature representation, where $D$ is the total sampling number of the global depth hypotheses which uniformly sample from $[d_{min}, d_{max}]$ here. However, previous methods do not take the occlusion problem into consideration when constructing and aggregating the cost volume. To this end, we propose the Bi-Directional Projection to expand the capabilities of warped features representation, and the Occlusion-Aware Aggregation strategy to reduce the impact of occlusion at the source.

*1) Homography and Bi-Directional Projection:* In MVSNet [9], inspired by the traditional plane sweeping stereo [40], features extracted from Sec. III-A are warped into $D$ front-parallel planes of the reference camera to build feature volumes $\{V_i\}_{i=0}^{N}$. Identically, according to the reversibility of camera parameters, we can also build the reverse homography warping from source views to the reference view. And the mathematical representation of the differentiable homography which warps each pixel of $\{I_i\}_{i=0}^{N}$ at depth $d$ as

$$\begin{cases} H_{i\to 0}(d) = K_i R_i (I - \frac{(t_0 - t_i) n_0^T}{d}) R_0^{-1} K_0^{-1} \\ H_{0\to i}(d) = K_0 R_0 (I - \frac{(t_i - t_0) n_i^T}{d}) R_i^{-1} K_i^{-1} \end{cases}, \tag{1}$$

where $\{K_i, R_i, t_i\}_{i=0}^{N}$ are the camera parameters denote intrinsics, rotations and translations respectively, $n_i$ denotes the normal axis of corresponding cameras, $I$ is the identity matrix.

Then we can naturally use bidirectional projection $H_{i\to 0}(d)$ and $H_{0\to i}(d)$ , as well as differential bilinear interpolation [41] to warp deep features between the reference image and source images to obtain warped feature volume $\{\mathcal{W}(V_i)\}_{i=0}^{N}$ in reference view and $\{\mathcal{W}'(V_i)\}_{i=0}^{N}$ in source views.

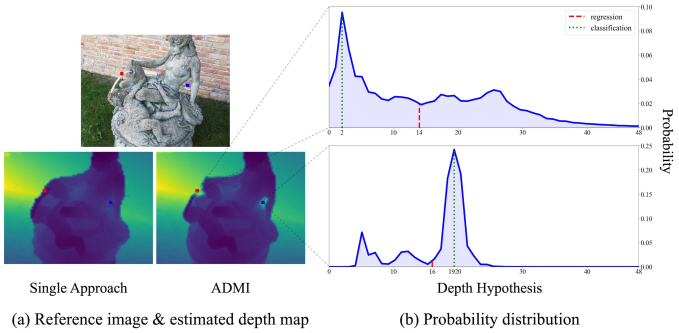
*2) Occlusion-Aware Aggregation:* As for unidirectional projection, similar to [9] [10] [22], the variance-based cost metric is used for pair-wise feature volume aggregation since explicit occlusion processing cannot be modeled through only two viewpoints (one reference view and one source view). Specifically, the pair-wise cost volume is calculated as

$$C_{pair} = \frac{\sum_{i=0}^{N_{pair}} (\mathcal{W}(V_i) - \overline{\mathcal{W}(V_i)})^2}{N_{pair}}, \tag{2}$$

where $N_{pair} = 2$ and $\overline{\mathcal{W}(V_i)}$ is the numerical average of warped cost volumes. $C'_{pair}$ constructs by $\mathcal{W}'(V_i)$ from the opposite direction is also the case.

After obtaining pair-wise cost volumes $C_{pair}$ and $C'_{pair}$ for each viewpoint, a light-weight 3D hourglass CNN [10] [22] will be used for matching cost enhancement between warped features. Then we can get pair-wise probability volume $P_{pair}$ and $P'_{pair}$ by applying the $softmax$ operation along the depth direction, to estimate the "confidence" which precisely *occlusion clues* here, denoted as $\{O_i\}_{i=0}^{N}$ and $\{O'_i\}_{i=0}^{N}$.

Different from PVSNet [13] and Vis-MVSNet [12], we directly use the occlusion clues (terms *visibility maps* and *uncertainly maps* in their works respectively) for aggregating




(a) Reference image & estimated depth map (b) Probability distribution

Fig. 3. **Illustrations on adaptive depth map inference strategy (ADMI).** (a) One reference image on BlendedMVS dataset [42] and estimated depth maps; (b) probability distributions for boundary (top) and occlusion area (bottom). ADMI achieves sharp boundaries and correct perspective relationship (blue rectangle) compared to the single approach, and according to strong occlusion clues, the classification method (green dotted line) predicts reasonable depth.

origin warped feature volume $\{\mathcal{W}(V_i)\}_{i=0}^{N}$ (vs. pair-wise cost volumes $\{C_{pair}\}_{i=0}^{N}$) along the reference axis. This avoids the propagation and accumulation of incorrect matches caused by occlusion from the root. The occlusion-aware aggregation can be finally modeled as

$$C = \frac{1}{N-1}\sum_{i=1}^{N-1} \widetilde{clue}_i \odot (\mathcal{W}(V_i) - \mathcal{W}(V_0))^2 \ , \tag{3}$$

$$\widetilde{clue}_i = 1 - tanh(OEN(O_i, O_i')) \ , \tag{4}$$

where $\widetilde{clue}$ is the final bidirectional occlusion clue, *OEN* is the Occlusion-Enhanced Network to be presented in Sec. III-C, and $\odot$ denotes element-wise multiplication.

### C. Occlusion-Enhanced Cost Volume Regularization

As we have mentioned above, warped features remain vulnerable to contamination due to occlusion. The bidirectional projection and occlusion-aware cost volume aggregation model the occlusion problem explicitly, and $\{O_i\}_{i=0}^{N}$ and $\{O_i'\}_{i=0}^{N}$ already encode the spatial occlusion clues of each pixel.

Additionally, we propose Occlusion-Enhanced Network to further extend occlusion clues to channel dimension and depth hypothesis dimension. First, we concatenate $\{O_i\}_{i=0}^{N}$ and $\{O_i'\}_{i=0}^{N}$ to have a unified input representation, and we use a 2D CNN acting like U-Net [43] which expands the searching receptive field to promote the origin occlusion clues. The last convolution layer produces $M$-channel features followed by the $tanh$ activation function and truncation that normalizes the occlusion-aware weights to $0 \sim 1$. And according to the uniform depth hypothesis, we use the channel-expanded clues for depth-wise interpolation. And finally, the occlusion-aware cost volume can be aggregated by the following Equ. 3, and the final occlusion clues $\{\widetilde{clue}\}_{i=1}^{N}$ can be derived from Equ. 4.

The occlusion-aware cost volume $C$ indicates the occlusion clues from spatial-wise, channel-wise, and depth-wise currently. And we also follow the classic works [9] [10] [22] to apply 3D CNN for further regularization.

### D. Depth Map Inference

Probability volume $P \in \mathbb{R}^{D\times H\times W}$ is generated by applying $softmax$ operation along the depth direction, and the depth estimation $\tilde{D}_0$ aligned with $I_0$ can be derived from it. Existing learning-based methods adopt *Regression* [9] [10] [13] [22] or *Classification* [21] [44] [45] approaches. We will introduce the individual characteristics first, and then propose our adaptive depth map inference strategy.

*1) Regression:* MVSNet [9] uses the expectation of probability volume and depth hypothesis to calculate the sub-pixel depth estimation for the first time:

$$\tilde{D} = \sum_{d=d_{min}}^{d=d_{max}} d \times P(d) \ . \tag{5}$$

The weighted sum in Equ. 5 reflects the impact of each depth hypothesis inherently and assigns larger weights for planes holding smaller matching costs. The regression achieves sub-pixel precision but boundaries usually suffer from lacking details due to the smoothing effect of the sum of weights.

*2) Classification:* Different from the weighted sum approach, the classification model directly applies pixel-wise winner-take-all [46] and chooses a depth index holding minimal matching cost as the final depth estimation:

$$\tilde{D} = \underset{d\in\{d_i\}_{i=1}^{N}}{argmax}\ P(d) \ . \tag{6}$$

The classification method can prevent the estimated plane from being polluted by the surrounding depth hypothesis (only one plane will be selected). However, it can easily cause depth discontinuity, especially in the foreground object.

*3) Adaptive Depth Map Inference Strategy:* By combining the advantages of both regression and classification approaches, we propose the Adaptive Depth Map Inference strategy which adaptively chooses the appropriate method for depth map estimation using the occlusion clues mentioned above. More specifically, we first take the average over $\widetilde{clue}_i$ along the depth dimension and set $\delta$ as the probability continuity threshold. When the reflecting probability is sufficiently continuous, we adaptively use *Regression* strategy for depth map estimation. And vice versa, the Bi-ClueMVSNet transitions to *Classification* mode for occlusion areas and boundaries holding a significantly discontinuous depth probability distribution. The visualization of probability distribution in boundary and occlusion area is shown in Fig. 3, and our ADMI strategy predicts smoother subject, sharper boundaries, and correct perspective relationship.

### E. Occlusion Clues-based Loss Function

The regression approaches [9] [10] [22] usually apply $l_1$ loss between the ground truth depth map $D_{GT}$ and estimated depth map $\tilde{D}$ directly, and classification methods [21] [44] [45] calculate cross-entropy between the latent depth representation $P$ and its ground truth distribution $P_{GT}$. Hence, for our adaptive

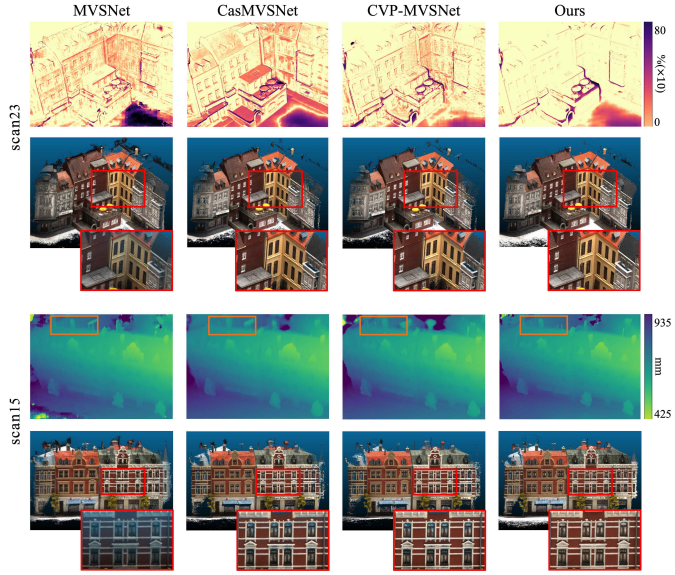



Fig. 4. **Qualitative results of scan 23 and scan 15 on the DTU dataset.** The first and third rows are error maps and estimated depth maps respectively, the top right scale bar represents the proportion of the *absolute depth error* to the DTU maximum depth (we magnified all errors by a factor of 10 for easy observation). The second and last rows are reconstructed point clouds.

depth map inference strategy, the supervised occlusion clues-based training loss and total loss are defined as:

$$OCL(p) = \begin{cases} \sum_{p\in\Omega} \|D_{GT}(p) - \tilde{D}(p)\|_1, & \widetilde{clue}(p) < \delta \\ \sum_{p\in\Omega} (-P_{GT}(p)\ log[P(p)]), & \widetilde{clue}(p) \geq \delta \end{cases}, \tag{7}$$

$$Loss = \sum_{l=1}^{L} \lambda^l OCL^l\ , \tag{8}$$

where $p$ denotes the pixel of a reference image, $\Omega$ denotes the set of valid pixels with ground truth precision, $l$ refers to the pyramid level discussed in Sec. III-A, and we adopt the combination weight $\lambda^l = 1$ for each level.

## IV. Experiments

In this section, we describe the datasets and evaluation metrics. After that, we validate the proposed Bi-ClueMVSNet and compare it with several state-of-the-art methods. Adequate ablation studies are presented at last.

### A. Dataset

**DTU Dataset** [56] is an indoor dataset consisting of 124 different objects, each training scene is recorded from 49 views with 7 brightness levels, and it contains the ground-truth point clouds for evaluation. We adopt the same data split as defined in SurfaceNet [57] and MVSNet [9] for a fair comparison.

**Tanks and Temples Dataset** [55] contains a more challenging realistic environment with large-scale variations and illumination changes. We use the intermediate split which consists of 8 scenes to evaluate the generalization of the method.

TABLE I

**Quantitative comparison on the DTU dataset.** Bold numbers represent the best while underlined numbers represent the second-best (lower is better).

| | Method | Acc. | Comp. | Overall↓ |
|---|---|---|---|---|
| Traditional | Furu [15] | 0.613 | 0.941 | 0.777 |
| | Tola [34] | 0.342 | 1.190 | 0.766 |
| | Camp [47] | 0.835 | 0.554 | 0.695 |
| | Gipuma [48] | **0.283** | 0.873 | 0.578 |
| | COLMAP [11] | 0.400 | 0.664 | 0.532 |
| Learning | MVSNet [9] | 0.396 | 0.527 | 0.462 |
| | R-MVSNet [21] | 0.383 | 0.452 | 0.417 |
| | Point-MVSNet [38] | 0.342 | 0.411 | 0.376 |
| | CasMVSNet [22] | 0.325 | 0.385 | 0.355 |
| | P-MVSNet [49] | 0.406 | 0.434 | 0.420 |
| | PVSNet [13] | 0.337 | 0.315 | 0.326 |
| | Vis-MVSNet [12] | 0.369 | 0.361 | 0.365 |
| | CVP-MVSNet [10] | <u>0.296</u> | 0.406 | 0.351 |
| | PatchmatchNet [50] | 0.427 | <u>0.277</u> | 0.352 |
| | EPP-MVSNet [27] | 0.413 | 0.296 | 0.355 |
| | CER-MVS [30] | 0.359 | 0.305 | 0.332 |
| | RayMVSNet [51] | 0.341 | 0.319 | 0.330 |
| | Effi-MVSNet [29] | 0.321 | 0.313 | 0.317 |
| | CDS-MVSNet [52] | 0.352 | 0.280 | 0.316 |
| | NP-CVP-MVSNet [28] | 0.356 | **0.275** | <u>0.315</u> |
| | UniMVSNet [53] | 0.352 | 0.278 | <u>0.315</u> |
| | Ours | 0.320 | 0.296 | **0.308** |

TABLE II

**Comparison of runtime and GPU consumption on DTU dataset.** All methods are tested under the same conditions.

| Method | Num. of Param. | Resolution | Runtime | GPU Mem. |
|---|---|---|---|---|
| CVP-MVSNet | 551585 | 1600 × 1184 | 3.06s | 7.40G |
| Vis-MVSNet | 1162696 | 1280 × 720 | 3.98s | 13.95G |
| | | 1440 × 960 | 4.98s | 15.84G |
| Ours | 989756 | 1280 × 720 | 1.74s | 9.63G |
| | | 1440 × 960 | 2.62s | 12.60G |

**BlendedMVS Dataset** [42] consists of over 17000 high-resolution rendered images with ground truth. We use it for qualitative analysis due to its large variety of scenes.

**Evaluation metrics.** For point cloud evaluation, the accuracy and completeness of the distance metric are adopted for DTU dataset [56] while the accuracy and completeness of the percentage metric for Tanks and Temples dataset [55].

### B. Implementation Details

For training, the number of input images is set to $N = 5$ with a resolution of $160 \times 128$. We use $L = 2$ layer pyramids and $\delta^1, \delta^2 = 0.3, 0.6$ for each level respectively. The initial depth planes hypothesis $D = 48$ and the depth sampling interval is from $425mm$ to $935mm$. We use PyTorch [58] for implementation and train the model with the Adam optimizer for 16 epochs from a start learning rate of $0.001$ on NVIDIA Tesla V100 (16G virtual memory).

We evaluate our model with $L = 5$ levels for high-resolution image input. We use $N = 5$ for DTU [56] and $N = 7$ for Tanks and Temples [55]. Both photometric and geometric filtering is considered similar to previous works [9] [10], and point clouds are reconstructed by fusion [48] depth maps.

TABLE III

**PERFORMANCE ON THE INTERMEDIATE SUBSET OF THE TANKS AND TEMPLES DATASET.** BOLD NUMBERS REPRESENT THE BEST WHILE UNDERLINED NUMBERS REPRESENT THE SECOND-BEST (HIGHER IS BETTER).

| **Method** | **Mean↑** | Family | Francis | Horse | LH | M60 | Panther | PG | Train |
|---|---|---|---|---|---|---|---|---|---|
| COLMAP [11] | 42.14 | 50.41 | 22.25 | 25.63 | 56.43 | 44.83 | 46.97 | 48.53 | 42.04 |
| ACMH [54] | 54.82 | 69.99 | 49.45 | 45.12 | 59.04 | 52.64 | 52.37 | 58.34 | 51.61 |
| MVSNet [9] | 43.48 | 55.99 | 28.55 | 25.07 | 50.79 | 53.96 | 50.86 | 47.90 | 34.69 |
| R-MVSNet [21] | 48.40 | 69.96 | 46.65 | 32.59 | 42.95 | 51.88 | 48.80 | 52.00 | 42.38 |
| Point-MVSNet [38] | 48.27 | 61.79 | 41.15 | 34.20 | 50.79 | 51.97 | 50.85 | 52.38 | 43.06 |
| P-MVSNet [49] | 55.62 | 70.04 | 44.64 | 40.22 | **65.20** | 55.08 | 55.17 | 60.37 | 54.29 |
| CasMVSNet [22] | 56.42 | 76.36 | 58.45 | 46.20 | 55.53 | 56.11 | 54.02 | 58.17 | 46.56 |
| PVSNet [13] | 56.88 | 74.00 | 55.17 | 39.85 | 61.37 | 60.22 | 56.87 | 58.02 | 49.51 |
| UCS-Net [23] | 54.83 | 76.09 | 53.16 | 43.03 | 54.00 | 55.60 | 51.49 | 57.38 | 47.89 |
| Vis-MVSNet [12] | 60.03 | 77.40 | 60.23 | 47.07 | 63.44 | **62.21** | 57.28 | 60.54 | 52.07 |
| CVP-MVSNet [10] | 54.03 | 76.50 | 47.74 | 36.34 | 55.12 | 57.28 | 54.28 | 57.43 | 47.54 |
| PatchmatchNet [50] | 53.15 | 66.99 | 52.64 | 43.24 | 54.87 | 52.87 | 49.54 | 54.21 | 50.81 |
| EPP-MVSNet [27] | **61.68** | 77.86 | 60.54 | **52.96** | 62.33 | 61.69 | **60.34** | **62.34** | **55.30** |
| Ours | 60.59 | **79.25** | **64.72** | 48.56 | 60.49 | 61.84 | 56.26 | 59.8 | 53.83 |

Fig. 5. **Qualitative result of the reconstructed point clouds on intermediate set on Tanks and Temples dataset [55].**

### *C. Benchmark Performance on DTU Dataset*

We compare our results with traditional geometric-based methods and other learning-based methods. The qualitative results are shown in Fig. 4. We calculated the absolute error between predicted depth maps and ground-truth depth maps. The top right corner of Fig. 4 represents the proportion of the absolute depth error to the maximum value of DTU depth (we magnified all errors by a factor of 10 for easy observation). For example, the darkness represents $80\%/10 = 8\%$ absolute error compared to ground truth. The proposed method generates more complete point clouds with finer details, especially in boundaries and occlusion areas which are usually considered as the most challenging parts of MVS reconstruction.

For quantitative evaluation, we report accuracy (Acc.) and completeness (Comp.) using the MATLAB code provided by the official DTU dataset as shown in Tab. I. Our method outperforms other methods by a significant margin, especially for completeness. Although PatchmatchNet has better completeness than our method, this is due to the trade-off when fusing point clouds, not because their estimated depth maps are more accurate. It should be noted that our method is to improve completeness while ensuring accuracy.

We also compare the runtime and GPU consumption of our method with CVP-MVSNet (our baseline) and VisMVSNet in Tab. II. While under the same conditions, our running time has been reduced by $47\%$ and GPU consumption has been reduced by $20\%$ compared with Vis-MVSNet. Because we build the cost volume for not only the reference view but also source views, there is some increase in memory consumption compared with the baseline, but we get significantly better reconstruction results than others.

### *D. Generalization on Tanks and Temples Dataset*

The images of DTU are taken under a well-controlled indoor environment with a fixed camera pose. To further demonstrate the generalization ability of our method, we evaluate the model trained on DTU on the Tanks and Temples [55] dataset. Tab. III provides the quantitative results derived from each method and our corresponding point cloud reconstructions are visualized in Fig. 5. Due to computational resource constraints, our training condition is significantly different from Vis-MVSNet [12] and EPP-MVSNet [27] which are fully trained on BlendedMVS Dataset ($768 \times 576$). We believe it would be better if we could train on larger-scale data with higher resolution.

### *E. Ablation Study*

*1) Benefits from Bi-Directional Projection:* We first study the validity of the Bi-Directional Projection (BDP) in Fig. 6. As we can see, neither the projection from source views to the reference view nor the opposite one can have promising results. Our proposed BDP demonstrates significant advantages, especially in terms of fast training convergence.

*2) Benefits from ADMI and OCL:* The ablation results of the proposed Adaptive Depth Map Inference (ADMI) strategy and Occlusion Clues-based Loss (OCL) function are in

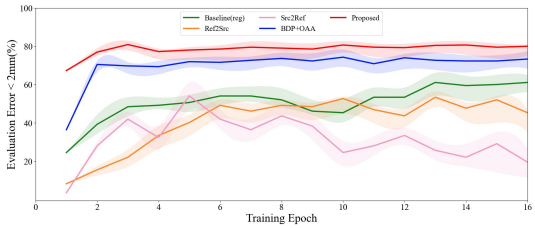



Fig. 6. **The percentage of evaluation error $< 2mm$ on the DTU dataset.** "Ref2src" and "Src2Ref" denote unidirectional projection, and the random error of training is indicated in shadow.

Tab. IV. As we have discussed in Fig. 3, the ADMI can reserve sharp boundaries and correct perspective relationships, and the occlusion clues-based loss function provides a better training strategy for superb convergence than single approaches (*Regression* or *Classification*).

*3) Benefits from Occlusion-Aware Cost Volume Construction and Regularization:* As shown in Tab. V, the Occlusion-Aware Aggregation (OAA) itself cannot promote completeness remarkably, let alone accuracy. However, after combining with the Bi-Directional Projection, the occlusion clues can significantly flow between the reference view and source views. And the Occlusion-Enhanced Network (OEN) further intensifies its characterization capabilities. We have also noticed that the baseline model (CVP-MVSNet [10]) provides better accuracy, which may be because our explicit occlusion clues still assign certain weights to correct matches and decrease the confidence for foreground objects during depth fusion. However, with comparable accuracy, the Bi-ClueMVSNet leads the way to a new era of completeness and overall reconstruction quality.

## V. Conclusion

In this paper, we present Bi-ClueMVSNet, a bidirectional occlusion clues-based MVS framework. We use bidirectional projection equipped with occlusion-aware aggregation for constructing cost volume and build the occlusion-enhanced network that further extends the 2D visibility maps to 3D occlusion clues to improve the representational ability. The estimated depth maps of the Bi-ClueMVSNet show great potential, especially on boundaries and occlusion areas, which is also credited to the application of the adaptive depth map inference strategy and occlusion clues-based loss function. Experiments on DTU and Tanks and Temples datasets demonstrate the effectiveness and generalization of Bi-ClueMVSNet. In the future, we intend to explore the application of the proposed method in the field of unsupervised or self-supervised MVS pipelines, which highlights the use of bidirectional occlusion clues to model boundaries and occluded areas.

TABLE IV
**Ablation results for ADMI and OCL on the DTU [56] evaluation set.** "Reg." and "Cla." refer to regression and classification strategies for depth map inference respectively. ◇ denotes completeness preference when fusing point clouds.

| Method | Inference | | | Loss | | | Num. | Acc. | Comp. | Overall↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| | Reg. | Cla. | ADMI | L1 | CE | OCL | | | | |
| Baseline(Reg)[3] | ✓ | | | ✓ | | | 3 | 0.296 | 0.406 | 0.351 |
| Baseline(Reg)[5] | ✓ | | | ✓ | | | 5 | **0.288** | 0.395 | 0.342 |
| Baseline(Cla)◇ | | ✓ | | | ✓ | | 5 | 0.362 | 0.311 | 0.337 |
| Baseline(ADMI) | | | ✓ | ✓ | | | 5 | 0.297 | 0.365 | 0.331 |
| Baseline(ADMI)◇ | | | ✓ | | ✓ | | 5 | 0.339 | 0.322 | 0.331 |
| Baseline(ADMI)+OCL◇ | | | ✓ | | | ✓ | 5 | 0.354 | **0.285** | **0.320** |

TABLE V
**Ablation results for BDP and OAA on the DTU [56] evaluation set.** "Var." refers to the variance-based aggregation. We choose regression strategy and $l_1$ loss for the baseline method, using $N = 5$ for training.

| Method | Cost Volume | | | Regularization | | Acc. | Comp. | Overall↓ |
|---|---|---|---|---|---|---|---|---|
| | Var. | OAA | BDP | 3D-CNN | OEN | | | |
| Baseline+Variance | ✓ | | | ✓ | | **0.288** | 0.395 | 0.342 |
| Baseline+OAA | | ✓ | | ✓ | | 0.293 | 0.387 | 0.340 |
| Baseline+OAA+BDP | | ✓ | ✓ | ✓ | | 0.296 | 0.360 | 0.328 |
| Baseline+OAA+BDP+OEN◇ | | ✓ | ✓ | ✓ | ✓ | 0.338 | 0.298 | 0.318 |
| Proposed◇ | | ✓ | ✓ | ✓ | ✓ | 0.320 | **0.296** | **0.308** |

## References


[1] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 1. IEEE, 2006, pp. 519–528.

[2] Y. Xue, J. Chen, W. Wan, Y. Huang, C. Yu, T. Li, and J. Bao, "Mvscrf: Learning multi-view stereo with conditional random fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4312–4321.

[3] K. Luo, T. Guan, L. Ju, Y. Wang, Z. Chen, and Y. Luo, "Attention-aware multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1590–1599.

[4] Z. Yu and S. Gao, "Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1949–1958.

[5] H. Yi, Z. Wei, M. Ding, R. Zhang, Y. Chen, G. Wang, and Y.-W. Tai, "Pyramid multi-view stereo net with self-adaptive view aggregation," in *European Conference on Computer Vision*. Springer, 2020, pp. 766–782.

[6] Q. Xu and W. Tao, "Learning inverse depth regression for multi-view stereo with correlation cost volume," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 508–12 515.

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

[8] J. Yan, Z. Wei, H. Yi, M. Ding, R. Zhang, Y. Chen, G. Wang, and Y.-W. Tai, "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking," in *European conference on computer vision*. Springer, 2020, pp. 674–689.

[9] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 767–783.

[10] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4877–4886.

[11] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] J. Zhang, Y. Yao, S. Li, Z. Luo, and T. Fang, "Visibility-aware multi-view stereo network." *british machine vision conference*, 2020.

[13] Q. Xu and W. Tao, "Pvsnet: Pixelwise visibility-aware multi-view stereo network," *arXiv preprint arXiv:2007.07714*, 2020.

[14] M. Lhuillier and L. Quan, "A quasi-dense approach to surface reconstruction from uncalibrated images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 3, pp. 418–433, 2005.

[15] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stere-

opsis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 8, pp. 1362–1376, 2009.

[16] J. S. De Bonet and P. Viola, "Poxels: Probabilistic voxelized volume reconstruction," in *Proceedings of International Conference on Computer Vision (ICCV)*, vol. 2, 1999.

[17] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *International journal of computer vision*, vol. 38, no. 3, pp. 199–218, 2000.

[18] S. B. Kang, R. Szeliski, and J. Chai, "Handling occlusions in dense multi-view stereo," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1. IEEE, 2001, pp. I–I.

[19] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *European conference on computer vision*. Springer, 2002, pp. 82–96.

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

[21] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent mvsnet for high-resolution multi-view stereo depth inference," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5525–5534.

[22] X. Gu, Z. Fan, Z. Dai, S. Zhu, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," *computer vision and pattern recognition*, 2019.

[23] S. Cheng, Z. Xu, S. Zhu, Z. Li, L. E. Li, R. Ramamoorthi, and H. Su, "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2524–2534.

[24] H. Xu, Z. Zhou, Y. Wang, W. Kang, B. Sun, H. Li, and Y. Qiao, "Digging into uncertainty in self-supervised multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6078–6087.

[25] F. Wang, S. Galliani, C. Vogel, and M. Pollefeys, "Itermvs: Iterative probability estimation for efficient multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8606–8615.

[26] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8645–8654.

[27] X. Ma, Y. Gong, Q. Wang, J. Huang, L. Chen, and F. Yu, "Epp-mvsnet: Epipolar-assembling based depth prediction for multi-view stereo," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5732–5740.

[28] J. Yang, J. M. Alvarez, and M. Liu, "Non-parametric depth distribution modelling based depth inference for multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8626–8634.

[29] S. Wang, B. Li, and Y. Dai, "Efficient multi-view stereo by iterative dynamic cost volume," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8655–8664.

[30] Z. Ma, Z. Teed, and J. Deng, "Multiview stereo with cascaded epipolar raft," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*. Springer, 2022, pp. 734–750.

[31] X. Wang, Z. Zhu, G. Huang, F. Qin, Y. Ye, Y. He, X. Chi, and X. Wang, "Mvster: Epipolar transformer for efficient multi-view stereo," in *European Conference on Computer Vision*. Springer, 2022, pp. 573–591.

[32] C. Cao, X. Ren, and Y. Fu, "Mvsformer: Learning robust image representations via transformers and temperature-based depth for multi-view stereo," *arXiv preprint arXiv:2208.02541*, 2022.

[33] Y. Ding, W. Yuan, Q. Zhu, H. Zhang, X. Liu, Y. Wang, and X. Liu, "Transmvsnet: Global context-aware multi-view stereo network with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8585–8594.

[34] E. Tola, C. Strecha, and P. Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Machine Vision and Applications*, vol. 23, no. 5, pp. 903–920, 2012.

[35] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5483–5492.

[36] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *European conference on computer vision*. Springer, 2016, pp. 501–518.

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

[38] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1538–1547.

[39] A. Fidler, U. Skaleric, and B. Likar, "The impact of image information on compressibility and degradation in medical image compression," *Medical physics*, vol. 33, no. 8, pp. 2832–2838, 2006.

[40] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," *Advances in neural information processing systems*, vol. 30, 2017.

[41] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *arXiv: Computer Vision and Pattern Recognition*, 2015.

[42] L. Quan, L. Zhou, J. Zhang, T. Fang, S. Li, Z. Luo, Y. Ren, and Y. Yao, "Blendedmvs: A large-scale dataset for generalized multi-view stereo networks," *computer vision and pattern recognition*, 2019.

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science*, 2015.

[44] J. Yan, Z. Wei, H. Yi, M. Ding, R. Zhang, Y. Chen, G. Wang, and Y.-W. Tai, "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking," *european conference on computer vision*, 2020.

[45] Z. Wei, Q. Zhu, C. Min, Y. Chen, and G. Wang, "Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network," *arXiv: Computer Vision and Pattern Recognition*, 2021.

[46] R. T. Collins, "A space-sweep approach to true multi-image matching," *computer vision and pattern recognition*, 1996.

[47] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *European Conference on Computer Vision*. Springer, 2008, pp. 766–779.

[48] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 873–881.

[49] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 10 451–10 460.

[50] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "Patchmatchnet: Learned multi-view patchmatch stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 194–14 203.

[51] J. Xi, Y. Shi, Y. Wang, Y. Guo, and K. Xu, "Raymvsnet: Learning ray-based 1d implicit fields for accurate multi-view stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8595–8605.

[52] K. T. Giang, S. Song, and S. Jo, "Curvature-guided dynamic scale networks for multi-view stereo," *arXiv preprint arXiv:2112.05999*, 2021.

[53] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8645–8654.

[54] Q. Xu and W. Tao, "Multi-view stereo with asymmetric checkerboard propagation and multi-hypothesis joint view selection," 2018. [Online]. Available: https://arxiv.org/abs/1805.07920

[55] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

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

[57] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "Surfacenet: An end-to-end 3d neural network for multiview stereopsis," *international conference on computer vision*, 2017.

[58] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Z. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.