Learning Deformable Hypothesis Sampling for Accurate PatchMatch Multi-View Stereo

Hongjie Li*, **Yao Guo***, **Xianwei Zheng**†, **Hanjiang Xiong**

The State Key Lab. LIESMARS, Wuhan University, China
{lihongjie,guoyao_gy,zhengxw,xionghanjiang}@whu.edu.cn

Abstract

This paper introduces a learnable Deformable Hypothesis Sampler (DeformSampler) to address the challenging issue of noisy depth estimation for accurate PatchMatch Multi-View Stereo (MVS). We observe that the heuristic depth hypothesis sampling modes employed by PatchMatch MVS solvers are insensitive to (i) the piece-wise smooth distribution of depths across the object surface, and (ii) the implicit multi-modal distribution of depth prediction probabilities along the ray direction on the surface points. Accordingly, we develop DeformSampler to learn distribution-sensitive sample spaces to (i) propagate depths consistent with the scene's geometry across the object surface, and (ii) fit a Laplace Mixture model that approaches the point-wise probabilities distribution of the actual depths along the ray direction. We integrate DeformSampler into a learnable PatchMatch MVS system to enhance depth estimation in challenging areas, such as piece-wise discontinuous surface boundaries and weakly-textured regions. Experimental results on DTU and Tanks & Temples datasets demonstrate its superior performance and generalization capabilities compared to state-of-the-art competitors. Code is available at https://github.com/Geo-Tell/DS-PMNet.

Introduction

Multi-View Stereo (MVS) aims to reconstruct dense 3D scene geometry from image sequences with known cameras, which has been widely used in robot perception, 3D reconstruction, and virtual reality. MVS is typically treated as a dense correspondence search problem (Galliani, Lasinger, and Schindler 2015), but many traditional methods have difficulty in achieving reliable matching within the low-texture, specular, and reflective regions. Learning-based MVS has recently attracted interest in solving this problem by introducing global semantic information for robust matching (Yao et al. 2018; Zhang et al. 2023b). Although achievements have been made, they still face the challenge of bridging the gap between accuracy and efficiency.

Learning-based methods commonly involve building a 3D cost volume, followed by a regularization using the 3D CNN for depth regression (Yao et al. 2018). Consequently, the 3D forms of both cost volume and CNN are undoubtedly restricted by limited resources. To overcome these limitations, many efforts have been made to reduce the cost volume size (Gu et al. 2020; Cheng et al. 2020) and modify the regularization techniques (Yao et al. 2019; Yan et al. 2020). Recently, a promising solution has emerged, which forgoes the common learning paradigm and re-evolves the traditional PatchMatch MVS into an end-to-end framework, like PatchMatchNet (Wang et al. 2021) and PatchMatch-RL (Lee et al. 2021). These methods follow the idea of patch-based searching and achieve improved results in efficiency and quality. However, we observe that they only transform the traditional pipeline into a trainable one, without adequately considering the implicit depth distribution within scenes for guiding depth hypothesis sampling during depth propagation and perturbation. This flaw directly degenerates the depth estimation performance, as shown in Figure1(d). Although PatchMatchNet introduces variability to sampling with CNNs, it remains insensitive to the underlying depth distribution. This will hamper the sampling of optimal hypotheses, thereby imposing additional burden on the subsequent learning modules. Therefore, our study raises two crucial questions for hypothesis sampling: (i) *What implicit depth distributions should be learned?* (ii) *How the learned distribution be leveraged to guide hypothesis sampling?*

At the propagation stage, hypotheses of neighboring pixels are sampled to generate a collection for enhancing each pixel's hypothesis space. An implicit piece-wise smooth depth distribution is contained in the depth map due to the scene regularity in the real world. In other words, the depth distribution tends to be smooth within coherent surfaces but can have abrupt shifts between distinct objects or scene elements. However, a preset sampling template is vulnerable when dealing with this implicit distribution (Lee et al. 2021; Duggal et al. 2019). This unreasonable hypothesis sampling results in a lot of noises with significant hypothesis differences in the collection, thereby causing unstable hypothesis evaluation, as revealed by Figure 1(b). Thus, a well-designed sampling scheme is required to select hypotheses from pixels that align more closely with the object's surface.

At the perturbation stage, fine-grained hypotheses over the scene depth range are expected to be sampled for refining the previously estimated depths. The optimization at this stage has received little attention in recent works. Gipuma

*These authors contributed equally.
†Corresponding author.

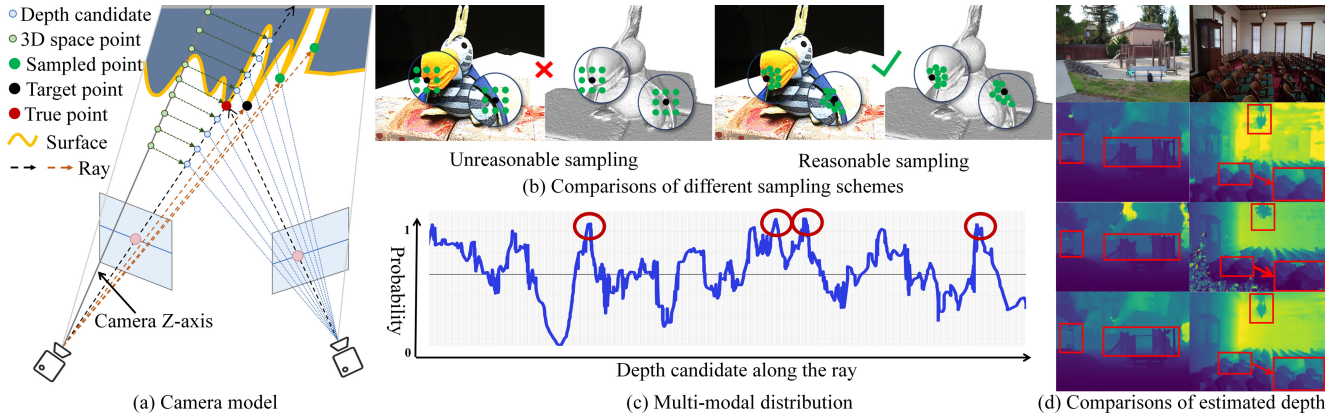


Figure 1: (a) A camera model is used to understand the hypothesis sampling from the 3D viewpoint. (b) Comparisons between different sampling schemes. The unreasonable sampling template often results in sampling hypotheses from plausible neighboring pixels during propagation. Pixels that appear to be closely adjacent in the image might correspond to significantly distant 3D points in space, and these 3D points could even belong to different objects, like the black and green points in the left subfigure. (c) An example of the multi-model distribution of depth prediction probabilities illustrates a multi-peak case regarding the minimum cost (or maximum matching probability). (d) Qualitative comparison of depth estimation on the courtroom and playground scenes of the Tanks & Temples dataset, respectively. From top to bottom: PathchMatchNet (Wang et al. 2021), PatchMatch-RL (Lee et al. 2021) and our DS-PMNet.

(Galliani, Lasinger, and Schindler 2015) employs a bisection strategy to refine sampling, while COLMAP (Schönberger et al. 2016) combines randomly perturbed samples with previous results in various ways. These methods lack the consideration of the uncertainty inherent in previous estimates, leading to redundant and coarse sampling. In this work, we intend to utilize this uncertainty to adaptively adjust the range of perturbations, rather than uniformly sampling for each pixel. In other words, for pixels with high confidence, the sampling should focus on hypotheses closely distributed around the previous estimates. Conversely, the sampling should encompass more dispersed hypotheses for pixels with significant uncertainty to provide a higher likelihood for correcting the estimates. In fact, we find that the cost distribution induced by previous sampling hypotheses offers a good representation of the uncertainty. However, due to the influence of varying imaging conditions, such as lighting, viewpoint, and other factors, this distribution often possesses multi-modal characteristics, as illustrated in Figure 1(c). This means that there is not a single distinct peak representing the lowest cost, which leads to even the true hypothesis failing to derive the lowest cost.

Based on the discussion above, we develop a learnable Deformable Hypothesis Sampler (DeformSampler) to learn the implicit depth distributions to guide reliable sampling in the learning-based PatchMatch framework. DeformSampler supports each pixel to sample optimal hypotheses at the stages of propagation and perturbation. Two modules are designed to drive this sampler: a plane indicator and a probabilistic matcher. The plane indicator encodes the intra-view feature consistency to learn the implicit depth distribution across the object surface. Using a Laplace Mixture model, the probabilistic matcher models multi-model distribution of depth prediction probabilities along the ray direction. By integrating this sampler into a learning-based PatchMatch framework, we can achieve excellent depth estimation performance, especially in the challenging piece-wise discontinuous surface boundaries and weakly-textured regions. Our method also shows strong generalization ability in both outdoor and indoor scenes, as shown in Figure 1(d).

In summary, our contributions are as follows:

- We develop a learnable PatchMatch-based MVS network (**DS-PMNet**) embedded with DeformSampler to learn implicit depth distribution for guiding the deformable hypothesis sampling.
- A plane indicator is designed to capture piece-wise smooth depth distribution for structure-aware depth propagation.
- A probability matcher is designed to model the multi-modal distribution of depth prediction probabilities for uncertainty-aware perturbation.

Related Works

Traditional MVS

Traditional MVS methods mainly rely on 3D representations, such as voxel, level-set, polygon mesh, and depth map (Seitz et al. 2006) for dense scene geometry reconstruction. Among them, the depth map-based methods usually gain more robust performance for large-scale dense 3D scene recovery by treating MVS as a dense correspondence search problem. In this line of research, PatchMatch MVS is a milestone, which replaces the costly dense point-based search with efficient patch-based search via a random and iterative strategy. Later, the propagation scheme of PatchMatch MVS was optimized for higher efficiency in some works, like Gipuma (Galliani, Lasinger, and Schindler 2015) and COLAMP (Schönberger et al. 2016) .

Recently, some methods have promoted the performance of PatchMatch MVS in terms of propagation and evaluation for accurate depth estimation. For propagation, ACMH (Xu and Tao 2019) adopted an Adaptive Checkerboard Sampling scheme to prioritize good hypotheses. HPM-MVS (Ren et al. 2023) enlarged the local propagation region by introducing a Non-local Extensible Sampling Pattern. While these methods provide sensitivity to scene region information, they still have difficulty explicitly capturing such details. To improve hypothesis evaluation in weakly-textured regions, approaches such as utilizing multiscale evaluation strategies, incorporating planar priors (Xu et al. 2022; Xu and Tao 2020a; Romanoni and Matteucci 2019), and employing deformable evaluation regions (Wang et al. 2023) have been adopted. In this work, we develop the DeformSampler to effectively impose awareness to scene structures and learn the implicit depth distribution from the current viewpoint for improved hypothesis propagation.

Learning-based MVS

While traditional solutions perform well in ideal Lambertian scenes, learning-based methods offer better semantic insight and stronger robustness in challenging scenarios. Most learning-based methods were built on MVSNet's foundation. They use warped multi-view features to create cost volumes and adopt 3D CNNs for regularization. Finally, the depths are predicted via regression. Recent works aim to enhance the quality of 3D cost volumes, reduce their size, and refine regularization techniques. To improve quality, techniques like attention mechanisms (Lee, Zou, and Hoiem 2022; Zhu et al. 2021; Cao, Ren, and Fu 2022), epipolar-assembling kernels (Ma et al. 2021), and pixel-wise visibility computation modules (Zhang et al. 2023a; Xu and Tao 2020b) were utilized. For computational efficiency, a common approach is to utilize a coarse-to-fine strategy (Gu et al. 2020; Yang et al. 2020), which involves a multi-stage hypothesis sampling. Some variants, like UCS-Net (Cheng et al. 2020) and IS-MVSNet (Wang et al. 2022), adaptively adjust sampling by incorporating uncertainty from depth estimation in earlier stages. For regularization, several studies adopted hybrid 3D U-Net (Luo et al. 2019; Sormann et al. 2020), RNN-CNN (Wei et al. 2021). In our work, the probabilistic matcher within the DeformSampler employs the same coarse-to-fine strategy but utilizes a more powerful modeling approach to capture the implicit multi-modal distribution of depth prediction probabilities, guiding fine-grained sampling.

Learning-based PatchMatch MVS

Recent advancements have integrated the idea of PatchMatch MVS into end-to-end training frameworks, such as PatchMatchNet (Wang et al. 2021) and PatchMatch-RL (Lee et al. 2021), which have partially bridged the gap between quality and efficiency for learning-based MVS. PatchMatchNet incorporated adaptive propagation and evaluation strategies to achieve efficient depth estimation. PatchMatch-RL argued that traditional methods perform better than learning-based MVS in wide-baseline scenes due to their joint optimization over pixel-wise depth, normals, and visibility estimates. In their follow-up work (Lee, Zou, and Hoiem 2022), they further considered the optimization over many high-resolution views and the usage of geometric consistency constraints. Our learnable DS-PMNet is also built upon the PatchMatch MVS but addresses the unreasonable hypothesis sampling issue. The core module, DeformSampler, provides more reliable guidance for propagation and perturbation, leading to a significant performance improvement.

Method

Figure 2 shows the entire pipeline of our method. In the following subsections, we first provide an overview of the end-to-end learnable PatchMatch MVS embedded with the DeformSampler (DS-PMNet). Then, we discuss how the two modules (the Plane Indicator $\mathcal{P}_\theta$ and Probability Matcher $\mathcal{M}_\theta$) learn the implicit depth distribution to drive the sampler for implementing deformable depth sampling.

Overview

In the PatchMatch MVS paradigm, each image is sequentially used as a reference image I^r, while the remaining images serve as source images $\{I_i^s\}_{i=1}^{N-1}$ to assist in estimating the depth map of I^r. The estimation process involves stages of initialization, propagation, evaluation, and perturbation, with the latter three stages iterating multiple times. In this work, we perform optimization at four different feature scales, with only one iteration per scale. The detailed DS-PMNet framework is presented in Algorithm 1 of the supplementary material.

We first extract a feature pyramid $\Psi = \{\varphi_\ell\}_{\ell=1}^{L}$ for each input image to capture the low-level details and high-level contextual information denoted as follows,

$$\{\varphi_\ell^r\}_{\ell=1}^{L}, \{\varphi_\ell^s\}_{\ell=1}^{L} = F_\theta\left(I^r, I^s\right), \tag{1}$$

where F_θ is an encoder, and $\ell \in \{1, ..., L\}$ is the indices for the multi-scale features. In this work, the feature pyramid is constructed with four scales, denoted as $L = 4$, corresponding to 1/8, 1/4, 1/2, and 1 of the original image size. To avoid confusion, we only describe four stages of one iteration in the following content, i.e., the subscript ℓ is discarded.

In stage **I**, we randomly initialize a depth map D^0 for I^r. The known depth range is first divided into m_0 intervals in the inverse depth space. Then, we randomly sample a depth candidate for each pixel at each interval. This means that each pixel is assigned with m_0 candidates $\{d_j\}_{j=1}^{m_0}$, which ensures that true hypotheses can be propagated quickly under a limited number of iterations.

In stage **II**, the plane indicator $\mathcal{P}_\theta$ guides the structure-aware hypothesis propagation by capturing implicit piecewise smooth depth distribution of the object's surface. $\mathcal{P}_\theta$ encodes the intra-view feature consistency to estimate a plane flow field $\mathcal{F}$ for I^r. For each pixel, $\mathcal{F}$ provides m_1 neighboring coplanar points to sample hypotheses, resulting in a reliable hypothesis collection $\{d_j\}_{j=1}^{m_0+m_1}$.

In stage **III**, the probabilistic matcher $\mathcal{M}_\theta$ enhances the evaluation of depth candidates in $\{d_j\}_{j=1}^{m_0+m_1}$ by modeling implicit multi-modal distribution of depth prediction probabilities, and outputs the prediction uncertainty to guide the

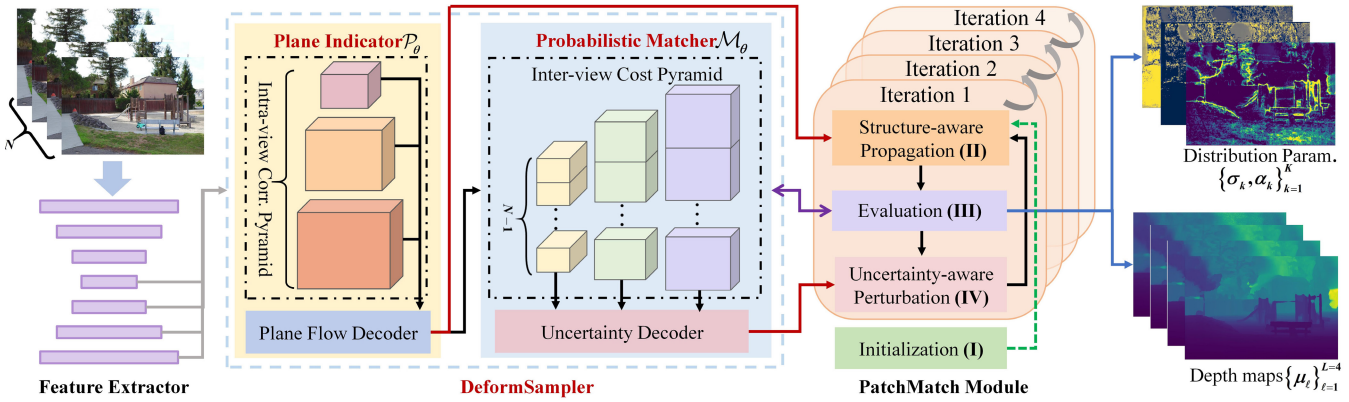


Figure 2: An overview of the proposed DS-PMNet, which is built upon the PatchMatch framework with the DeformSampler embedded to achieve a deformable hypothesis sampling. DeformSampler learns implicit depth distribution using Plane Indicator $\mathcal{P}_\theta$ and Probability Matcher $\mathcal{M}_\theta$. The purple double sided arrow line indicates that $\mathcal{M}_\theta$ belongs to evaluation.

subsequent perturbation. $\mathcal{M}_\theta$ first generates a multi-view cost volume $\mathcal{S} = \{\mathrm{S}_i\}_{i=1}^{N-1}$, where each element S_i encodes a matching cost introduced by the depth-induced homography matrix set $\{\mathrm{H}_j\}_{j=1}^{m_0+m_1}$ between φ^r and φ_i^s. Then, for each pixel in I^r, the parameter set of Laplace Mixture distribution $\{\psi_i\}_{i=1}^{N-1}$ is decoded from $\mathcal{S}$ to predict depth map D and the corresponding uncertainty map set $\mathcal{U} = \{U_i\}_{i=1}^{N-1}$.

In stage **IV**, the inferred Laplace Mixture distribution is used to guide the uncertainty-aware perturbation, and a fine-grained hypothesis collection $\{d_j\}_{j=1}^{m_2}$ is sampled. Then, this collection is further input into stage **II**, and $m_0 \leftarrow m_2$.

Plane Indicator for Deformable Propagation

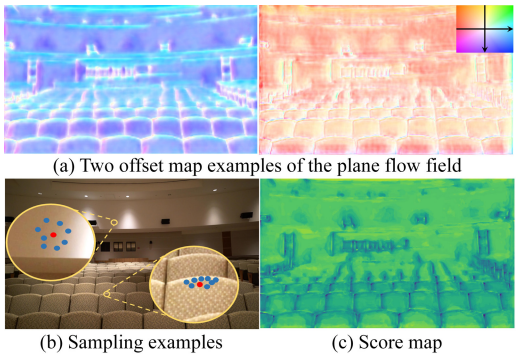
The plane indicator $\mathcal{P}_\theta$ encodes the self-similarity of features within the reference view to learn the relationship between scene structure and depth under the whole PatchMatch solver, thereby decoding a plane flow field $\mathcal{F} \in \mathbb{R}^{H\times W\times 2M}$ that represents the planar regions of the scene. This field contains M offset maps, where each element in the offset map represents the directional displacement between a location and its neighboring points in the same scene plane. Examples of offset maps are shown in Figure 3(a). Utilizing this $\mathcal{F}$, each pixel is guided to sample reliable depth hypotheses from $m_1 (m_1 \leq M)$ neighboring points, as shown in Figure 3(b). In general, our $\mathcal{P}_\theta$ consists of two components: an intra-view correlation pyramid $\mathrm{C} = \{\mathcal{C}_\ell\}_{\ell=1}^{L-1}$ and a plane flow decoder $\mathcal{D}_\theta$ in Figure 4.

Intra-view Correlation Pyramid Construction Each $\mathcal{C}_\ell$ is generated by calculating the dot product between every pixel in the ℓ_{th} feature map and all the pixels within its designated neighborhood. The search radius R_1 determines the neighborhood range. Specifically, given the feature map φ_ℓ^r, each element $c_\ell(p,\eta)$ in $\mathcal{C}_\ell \in \mathbb{R}^{H_\ell\times W_\ell\times R_1}$ is defined as

$$c_\ell\left(\mathrm{p},\eta\right) = \frac{1}{\sqrt{h_\ell}}\left\langle \varphi_\ell^r[\mathrm{p}], \varphi_\ell^r[\mathrm{p}+\eta]\right\rangle, \|\eta\|_\infty \leq R_1, \quad (2)$$

where h_ℓ represents the channel number of ℓ_{th} feature map, $\mathrm{p} \in \mathbb{R}^2$ is a coordinate on the feature image and η is the offset from this location. The offset is constrained to $\|\eta\|_\infty \leq R_1$. The symbol $[\cdot]$ is used to extract the features at a specific coordinate from the feature map. Each level's search radius R_1 remains fixed. Therefore, the radius covers the largest feature map area at the top level, gradually reducing with each subsequent level.

Plane Flow Decoder This decoder $\mathcal{D}_\theta$ is designed to infer a plane flow field $\mathcal{F}_\ell \in \mathbb{R}^{H_\ell\times W_\ell\times 2M}$ progressively from the pyramid, which achieves a refined field $\mathcal{F}$ at last. Inspired by (Melekhov et al. 2019), the decoder incorporates four *Conv-BN-LeayReLU* (CBL) blocks and one predictor, as shown in Figure 4. Dense connections are added among the four blocks to enhance information exchange. Here, a slight adjustment is made in the predictor when inputting elements from different pyramid levels. At level $\ell = 1$, the predictor gives a coarse plane flow field $\mathcal{F}_1 \in \mathbb{R}^{H_1\times W_1\times 2M}$. At subsequent levels, the predictor generates a residual $\tilde{\mathcal{F}}_\ell \in$



(a) Two offset map examples of the plane flow field

(b) Sampling examples (c) Score map

Figure 3: Visualization of the inferred plane flow field $\mathcal{F}$. (a) The offset map is visualized in the form of optical flow, where colors indicate neighboring point directions relative to the current pixel, while color intensity represents offset magnitude. (b) Examples of deformable sampling in the weak texture and object boundary regions. (c) The scene-aligned distribution scoring for the plane flow field.

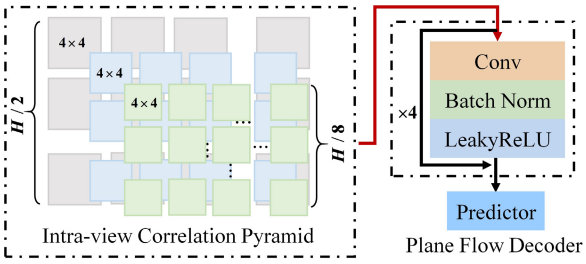


Figure 4: Illustration of the plane indicator $\mathcal{P}_\theta$. Element in each level of the pyramid is sequentially fed into the plane flow decoder to iteratively refine the plane flow field $\mathcal{F}$.

$\mathbb{R}^{H_\ell \times W_\ell \times 2}$ to refine the coarse field further, i.e.,

$$\mathcal{F}_\ell\left[\mathrm{p}\right] = \gamma \cdot \mathcal{F}_1^{\uparrow}\left[\mathrm{p}\right] + \tilde{\mathcal{F}}_\ell\left[\gamma \cdot \mathcal{F}_1^{\uparrow}\left[\mathrm{p}\right] + \mathrm{p} \otimes 1_M\right], \quad (3)$$

where $\mathrm{p} \in \mathbb{R}^2$ is the coordinate on the field, γ is the up-sampling factor, 1_M is a $1 \times M$ identity matrix. The symbols $\uparrow$, $[\cdot]$, and $\otimes$ represent the up-sampling, fetch operation, and Kronecker product operation, respectively.

Probabilistic Matcher for Deformable Perturbation

The probability matcher $\mathcal{M}_\theta$ is designed to model the multi-modal distribution of the depth prediction probabilities for guiding the fine-grained sampling during perturbation. We adopt a Laplace Mixture model containing two components ($K = 2$) to tackle this multi-peak issue, i.e.,

$$p\left(y|\psi\right) = \alpha_1 \frac{1}{\sqrt{2}\sigma_1} e^{\frac{\sqrt{2}}{\sigma_1}|y-\mu|} + \alpha_2 \frac{1}{\sqrt{2}\sigma_2} e^{\frac{\sqrt{2}}{\sigma_2}|y-\mu|}, \quad (4)$$

where $\psi = \{\mu, \alpha_1, \alpha_2, \sigma_1, \sigma_2\}$ is the distribution parameter set to be estimated, $\alpha_1 + \alpha_2 = 1$, y is the specific depth hypothesis. The mean μ of both components is set to be the same to ensure only one peak exists. Additionally, we achieve this by setting σ_1 as a constant for the former to represent the most accurate depth prediction and imposing the constraint $\sigma_2 > \sigma_1 > 0$ for the latter to model larger errors. Figure 5 visualizes an example of pixel-wise distribution parameters.

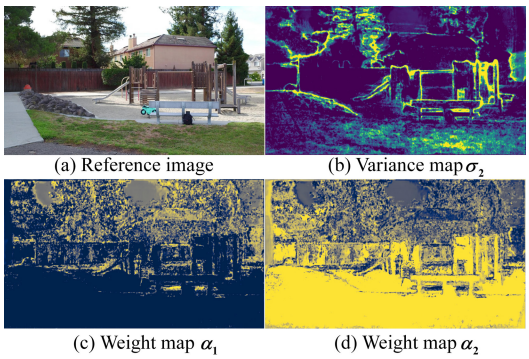


Figure 5: Visualization of pixel-wise distribution parameters. Brighter colors indicate larger values. Object edge regions exhibit higher uncertainty, i.e., higher σ_2.

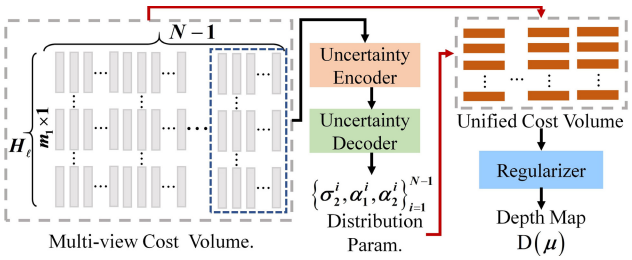


Figure 6: Illustration of the proposed probabilistic matcher $\mathcal{M}_\theta$. Only a single level of structure is displayed here.

Probabilistic Matcher Matcher $\mathcal{M}_\theta$ takes the inter-view cost pyramid as input, where each level of the pyramid contains a multi-view cost volume $\mathcal{S} = \{\mathrm{S}_i\}_{i=1}^{N-1}$ introduced by the source images. For each level of the pyramid, the matcher predicts the distribution parameter set $\{\psi_i\}_{i=1}^{N-1}$ for each pixel in the reference image, representing the matching uncertainty between the reference and different source images. The detailed structure can be seen in Figure 6. This matcher contains two branches. In the first branch, $\mathcal{S}$ is encoded into an uncertainty embedding, which is then decoded into $\{\sigma_2^i, \alpha_1^i, \alpha_2^i\}_{i=1}^{N-1}$ for each pixel. According to the inferred parameters, an uncertainty map set $\mathcal{U} = \{\mathrm{U}_i\}_{i=1}^{N-1}$ between the reference and source images can be obtained by computing depth prediction probabilities on each pixel, i.e., $u_i = P\left(|y-\mu| < R_2\right)$,

$$u_i = \alpha_1^i \left[1 - e^{-\sqrt{2}R_2}\right]^2 + \alpha_2^i \left[1 - e^{-\sqrt{2}\frac{R_2}{\sigma_2^i}}\right]^2, \quad (5)$$

where u_i is an element of a specific coordinate on U_i, R_2 is the hyper-parameter determining the acceptable deviation between ground truth and predicted depth map μ. Utilizing those visibility maps $\{\mathrm{U}_i\}_{i=1}^{N-1}$, an unified cost volume can be obtained by integrating $\sum_{i=1}^{N-1} \mathrm{U}_i\mathrm{S}_i$. In the second branch, a 3D CNN-based regularizer is adopted to process the unified cost volume to estimate a weighted depth map $D(\mu)$ with the sampled depth hypotheses. More details can be seen in the supplementary.

Probabilistic Perturbation σ_2 is adopted to guide hypothesis sampling because it represents high matching uncertainty. We first integrate the variances from different views into a unified value $E\left(\sigma_2\right) = \sum_{i=1}^{N-1} u_i\sigma_2^i$. Then, a sampling space is defined as $\left[\mu \pm \varepsilon E\left(\sigma_2\right)\right]$ for each pixel, where ε is the hyper-parameter. Then, we divide this range into m_2 bins, each containing an equal portion of probability mass. This ensures that pixels with low uncertainty sample candidates are closer to μ while pixels with high uncertainty sample more dispersed candidates to rectify μ. Subsequently, we sample the midpoint of each bin as a potential depth candidate. Thus, j_{th} depth candidate is defined as

$$d_j = \frac{1}{2}\left[\Phi\left(\frac{j-1}{m_2}\tilde{P} + \frac{P^*}{2}\right) + \Phi\left(\frac{j}{m_2}\tilde{P} + \frac{P^*}{2}\right)\right], \quad (6)$$

where $\Phi\left(\cdot\right)$ is used to transform the cumulative probability into coordinates of the Laplace distribution, m_2 is the

number of bins, $\tilde{P}$ is the probability mass covered by range $[\mu \pm \varepsilon E(\sigma_2)]$, $P^* = 1 - \tilde{P}$.

Loss Function

We first compute the loss $\mathcal{L}_{depth} = \sum_{\ell=1}^{L} \left\| \mathrm{D}^{gt} - \mathrm{D}_\ell \right\|$ between all predicted depth maps $\{\mathrm{D}_\ell\}_{\ell=1}^{L}$ and ground truth D^{gt}. Then, a negative log-likelihood loss $\mathcal{L}_{NLL}$ is adopted to supervise the fitted mixed Laplace distribution, i.e.,

$$\mathcal{L}_{NLL} = -\frac{1}{N-1} \sum_{\ell=1}^{L} \sum_{i=1}^{N-1} \log p\left(\mathrm{D}^{gt} | \psi_i\right). \quad (7)$$

Thus, the total loss $\mathcal{L}_{total}$ is defined as $\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{depth} + \lambda_2 \mathcal{L}_{NLL}$, where λ_1, λ_2 are the weight factors.

Experiments

Implementation Setup

Training and Testing Following the previous works like TransMVSNet (Ding et al. 2022), we initially train our DS-PMNet on the DTU dataset (Jensen et al. 2014) for DTU evaluation. Then, we fine-tune the model on the BlendedMVS dataset (Yao et al. 2020), and test it on the Tanks and Temples benchmark (Knapitsch et al. 2017). For training, we use 6 DTU images cropped to 512×640 as input in each batch. Our model is trained for 16 epochs with Adam optimizer, starting with a learning rate of 0.001, reduced by 0.2 at epochs 5, 9, and 13. To stabilize training against initial errors from random depth hypotheses, we set the initial learning rate of the probability matcher $\mathcal{M}_\theta$ to 10^{-5}. As for Fine-tuning on BlendedMVS, our model undergoes 10 epochs with an initial learning rate of 0.0002, using 6 input images at a resolution of 576 × 768. The batch size is set to two on NVIDIA RTX 3090 for DTU and one for BlendedMVS.

When assessing the DTU, we use 6 input images at 1152×1600 resolution (N=6). For the Tanks and Temples dataset, N is set to 8, with images at 1024×1920 resolution. We report the results in terms of the accuracy (Acc.), completeness (Comp.), and overall metrics for DTU dataset and evaluate the performance of precision (Pre.), recall (Rec.), and F_1-score (F_1) for Tanks and Temples benchmark.

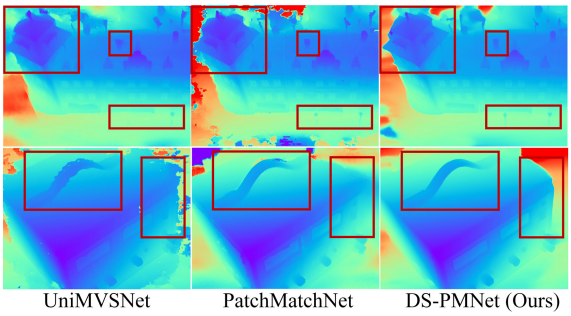


Figure 7: Qualitative results on DTU testing set. Our model stands out on the edges and thin structures of the depth map, as highlighted by the red box.

	Methods	Acc. ↓	Comp. ↓	Overall ↓
Tra.	Gipuma$_{ICCV2015}$	**0.283**	0.873	0.578
	COLMAP$_{CVPR2016}$	0.400	0.664	0.532
MVSNet	MVSNet$_{ECCV2018}$	0.396	0.527	0.462
	R-MVSNet$_{CVPR2019}$	0.383	0.452	0.417
	Point-MVSNet$_{ICCV2019}$	0.342	0.411	0.376
	CasMVSNet$_{CVPR2020}$	0.325	0.385	0.355
	CVP-MVSNet$_{CVPR2020}$	0.296	0.406	0.351
	UCS-Net$_{CVPR2020}$	0.338	0.349	0.344
	AA-RMVSNet$_{ICCV2021}$	0.376	0.339	0.357
	UniMVSNet$_{CVPR2022}$	0.352	0.278	0.315
	TransMVSNet$_{CVPR2022}$	0.321	0.289	0.305
	MVSTER$_{ECCV2022}$	0.350	0.276	0.313
	IS-MVSNet$_{ECCV2022}$	0.355	0.351	0.359
	RA-MVSNet$_{CVPR2023}$	0.326	0.268	0.297
	DispMVS$_{AAAI2023}$	0.354	0.324	0.339
	EPNet$_{AAAI2023}$	0.299	0.323	0.311
	GeoMVSNet$_{CVPR2023}$	0.331	0.259	0.295
	DMVSNet$_{ICCV2023}$	0.338	0.272	0.313
PM	PatchMatchNet$_{CVPR2021}$	0.427	0.277	0.352
	Ours	0.323	**0.257**	**0.290**

Table 1: Quantitative results of reconstructed point clouds on DTU testing set by using the distance metric $[mm]$(lower is better). Comparison methods are divided into three categories: Traditional approaches (Tra.), MVSNet variants, and PatchMatch series (PM). Acc. and Comp. stand for accuracy and completeness, respectively.

Benchmark Performance

Results on DTU We first evaluate our DS-PMNet on the DTU testing set, where the model is only trained on the DTU dataset. The quantitative results are reported in Table 1. Our DS-PMNet outperforms traditional and learning-based methods in the overall metric, achieving a highest score of 0.290. In terms of the accuracy, Gipuma (Galliani, Lasinger, and Schindler 2015) achieves the best results, while our approach demonstrates state-of-the-art performance for completeness. Additionally, we compare our method with PatchMatchNet (Wang et al. 2021) and UniMVSNet (Peng et al. 2022) for estimated depth maps. As shown in Figure 7, our method excels in recovering the depth of thin structures and object boundaries, where the edges align better with object boundaries than other methods.

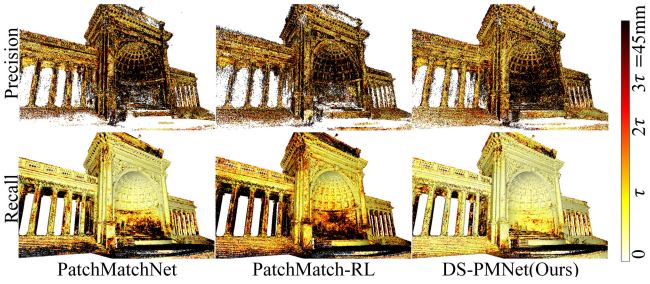


Figure 8: Qualitative comparison of point cloud reconstruction on the Tanks and Temples benchmark (Dark colors indicate large errors). The distance threshold τ is scene-dependent and is set to 15mm for the Temple scene.

	Methods	Intermediate (↑)			Advanced (↑)		
		Pre.	Rec.	F_1	Pre.	Rec.	F_1
Tra.	COLMAP$_{CVPR2016}$	43.16	44.48	42.14	33.65	23.96	27.24
	ACMMP$_{TPAMI2023}$	53.28	68.50	59.38	33.79	44.64	37.84
	APD-MVS$_{CVPR2023}$	55.58	75.06	63.64	33.77	49.41	<u>39.91</u>
MVSNet	MVSNet$_{ECCV2018}$	40.23	49.70	43.48	-	-	-
	CasMVSNet$_{CVPR2020}$	47.62	74.01	56.84	29.68	35.24	31.12
	IterMVS$_{CVPR2022}$	46.82	73.50	56.22	28.04	42.60	33.24
	Effi-MVS$_{CVPR2022}$	47.53	71.58	56.88	32.23	41.90	34.39
	EPNet$_{AAAI2023}$	53.26	71.60	60.46	30.75	44.12	35.80
MVSNet*	EPP-MVSNet$_{ICCV2021}$	53.09	75.58	61.68	**40.09**	34.63	35.72
	IS-MVSNet$_{ECCV2022}$	55.62	74.49	62.82	37.03	35.13	34.87
	RayMVSNet$_{CVPR2022}$	53.21	69.21	59.48	-	-	-
	IterMVS$_{CVPR2022}$	47.53	74.69	56.94	28.70	44.19	34.17
	TransMVSNet$_{CVPR2022}$	55.14	76.73	63.52	33.84	44.29	37.00
	DispMVS$_{AAAI2023}$	49.93	73.37	59.07	26.37	**53.67**	34.90
	EPNet$_{AAAI2023}$	**57.01**	72.57	63.68	34.26	<u>50.54</u>	**40.52**
PM	PatchMatchNet$_{CVPR2021}$	43.64	69.37	53.15	27.27	41.66	32.31
	PatchMatch-RL$^{*}_{ICCV2021}$	45.91	62.30	51.81	30.57	36.73	31.78
	PatchMatch-RL$^{*+}_{arXiv2022}$	50.48	63.27	55.32	<u>38.82</u>	32.35	33.80
	Ours	48.02	76.26	58.23	32.77	41.96	36.03
	Ours*	<u>56.02</u>	**76.76**	**64.16**	34.29	48.73	39.78

Table 2: Quantitative results on Tanks and Temples dataset (unit: %, higher is better). Methods are also divided into four categories: Traditional approaches (Tra.), MVSNet variants trained on DTU, MVSNet variants trained or fine-tuned on BlendedMVS (*), and PatchMatch series (PM). PatchMatch-RL^{+} (Lee, Zou, and Hoiem 2022) is an extension of PatchMatch-RL. Bold represents the best while underlined represents the second-best.

Results on Tanks and Temples We further validate the generalization ability of our method on the challenging Tanks and Temples dataset. As shown in Table 2, our DS-PMNet achieves competitive performance in precision and recall compared to the MVSNet variants. Specifically, our method ranks 1st and 3rd on the intermediate set and advanced set in terms of F_1 score respectively, outperforming most of methods. Compared with the existing learnable PatchMatch MVS methods, DS-PMNet outperforms better in all metrics. Figure 8 provides a qualitative comparison of the reconstructed point clouds for the differen methods, where our method show an enhanced precision and comprehensiveness. The above results demonstrate the robustness and generalization ability of our method.

Ablation Studies

Ablation Studies are first conducted to independently validate the two modules of DeformSampler (i.e., Plane Indicator $\mathcal{P}_\theta$ and Probability Matcher $\mathcal{M}_\theta$). Then, comprehensive analysis are made on various ε settings in the probability matcher to showcase our choice of the parameters.

Effectiveness of DeformSampler We first establish a baseline by incorporating Gipuma's fixed sampling modes in the propagation and perturbation stages into our learnable PatchMatch solver. Then, we progressively replace the sampling modes with our proposed plane indicator and probabilistic matcher for validating the effectiveness of our DeformSampler. Quantitative results evaluated on DTU are reported in Table 3. The results show that the solver incorporating both modules achieves the highest accuracy and completeness. This demonstrates the capability of our method in effectively learning the underlying depth distributions, guiding reliable hypothesis sampling during the propagation and perturbation stages. Additionally, our baseline outperforms PatchMatchNet in all metrics, demonstrating the superiority of our network design.

$\mathcal{P}_\theta$	$\mathcal{M}_\theta$	Acc.(mm)	Comp.(mm)	Overall.(mm)
		0.374	0.310	0.342
✓		0.368	0.296	0.322
	✓	0.356	0.284	0.320
✓	✓	**0.323**	**0.257**	**0.290**
PatchMatchNet		0.427	0.277	0.352

Table 3: Ablation studies for the plane indicator $\mathcal{P}_\theta$ and probabilistic matcher $\mathcal{M}_\theta$ on DTU's testing set (lower is better).

ε	Acc.(mm)	Comp.(mm)	Overall.(mm)
-,1,1,1	0.345	0.278	0.312
-,2,2,2	0.352	0.271	0.312
-,3,2,1	0.332	0.267	0.300
-,2,2,1	**0.323**	**0.257**	**0.290**

Table 4: Parameter sensitivity testing on DTU for uncertainty-aware perturbation in different stages (lower is better).

Parameter Sensitivity During the perturbation process, the parameter ε controls the range of the perturbation, i.e., $[\mu - \varepsilon\sigma, \mu + \varepsilon\sigma]$, and different perturbation ranges result in different sampling fineness. Therefore, we constrain the parameters to the subset $\{1, 2, 3\}$ to verify the optimality of our setting. Due to the step-by-step refinement in each iteration, ε in subsequent iterations must be less than or equal to the previous one. Additionally, the first iteration does not involve the perturbation process. The results of quantitative analysis on DTU are reported in Table 4. The best performance is achieved when ε is set to $\{2, 2, 1\}$, followed by $\{3, 2, 1\}$. These differences primarily stem from the initial ε setting. If ε is set too small (ε=1), the potentially valid hypotheses are excluded, while if too large (ε=3), the redundant noises hamper fine-grained sampling.

Conclusion

This paper presents a learnable DeformSampler that is embedded into PatchMatch MVS framework to facilitate the accurate depth estimation in complex scenarios. The proposed DeformSampler can help to sample distribution-sensitive hypothesis space during the propagation and perturbation. Extensive Experiments conducted on several challenging MVS datasets show that DeformSampler can effectively learn the piece-wise smooth depth distribution on the object surface for reliably propagating depth, while successfully capture the multi-modal distribution of depth prediction probabilities to allow for fine-grained hypothesis sampling. Comparisons with existing methods also demonstrate that our method can achieve state-of-the-art performance on MVS benchmarks.

Acknowledgments

This research was supported by NSFC-projects under Grant 42071370, the Fundamental Research Funds for the Central Universities of China under Grant 2042022dx0001, and Wuhan University-Huawei Geoinformatics Innovation Laboratory.

References

Cao, C.; Ren, X.; and Fu, Y. 2022. MVSFormer: Multi-View Stereo by Learning Robust Image Features and Temperature-based Depth. *Transactions on Machine Learning Research*.

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

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

Duggal, S.; Wang, S.; Ma, W.-C.; Hu, R.; and Urtasun, R. 2019. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4384–4393.

Galliani, S.; Lasinger, K.; and Schindler, K. 2015. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, 873–881.

Gu, X.; Fan, Z.; Zhu, S.; Dai, Z.; Tan, F.; and Tan, P. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2495–2504.

Jensen, R.; Dahl, A.; Vogiatzis, G.; Tola, E.; and Aanæs, H. 2014. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 406–413.

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Lee, J. Y.; DeGol, J.; Zou, C.; and Hoiem, D. 2021. Patchmatch-rl: Deep mvs with pixelwise depth, normal, and visibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6158–6167.

Lee, J. Y.; Zou, C.; and Hoiem, D. 2022. Deep PatchMatch MVS with Learned Patch Coplanarity, Geometric Consistency and Adaptive Pixel Sampling. *arXiv preprint arXiv:2210.07582*.

Luo, K.; Guan, T.; Ju, L.; Huang, H.; and Luo, Y. 2019. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10452–10461.

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

Melekhov, I.; Tiulpin, A.; Sattler, T.; Pollefeys, M.; Rahtu, E.; and Kannala, J. 2019. Dgc-net: Dense geometric correspondence network. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1034–1042. IEEE.

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

Ren, C.; Xu, Q.; Zhang, S.; and Yang, J. 2023. Hierarchical Prior Mining for Non-local Multi-View Stereo. *arXiv preprint arXiv:2303.09758*.

Romanoni, A.; and Matteucci, M. 2019. Tapa-mvs: Textureless-aware patchmatch multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10413–10422.

Schönberger, J. L.; Zheng, E.; Frahm, J.-M.; and Pollefeys, M. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 501–518. Springer.

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

Sormann, C.; Knöbelreiter, P.; Kuhn, A.; Rossi, M.; Pock, T.; and Fraundorfer, F. 2020. Bp-mvsnet: Belief-propagation-layers for multi-view-stereo. In *2020 International Conference on 3D Vision (3DV)*, 394–403. IEEE.

Wang, F.; Galliani, S.; Vogel, C.; Speciale, P.; and Pollefeys, M. 2021. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14194–14203.

Wang, L.; Gong, Y.; Ma, X.; Wang, Q.; Zhou, K.; and Chen, L. 2022. Is-mvsnet: importance sampling-based mvsnet. In *European Conference on Computer Vision*, 668–683. Springer.

Wang, Y.; Zeng, Z.; Guan, T.; Yang, W.; Chen, Z.; Liu, W.; Xu, L.; and Luo, Y. 2023. Adaptive Patch Deformation for Textureless-Resilient Multi-View Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1621–1630.

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

Xu, Q.; Kong, W.; Tao, W.; and Pollefeys, M. 2022. Multi-scale geometric consistency guided and planar prior assisted multi-view stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4945–4963.

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

Xu, Q.; and Tao, W. 2020a. Planar prior assisted patchmatch multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12516–12523.

Xu, Q.; and Tao, W. 2020b. Pvsnet: Pixelwise visibility-aware multi-view stereo network. *arXiv preprint arXiv:2007.07714*.

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

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

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

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

Yao, Y.; Luo, Z.; Li, S.; Zhang, J.; Ren, Y.; Zhou, L.; Fang, T.; and Quan, L. 2020. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1790–1799.

Zhang, J.; Li, S.; Luo, Z.; Fang, T.; and Yao, Y. 2023a. Vis-mvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision*, 131(1): 199–214.

Zhang, Z.; Peng, R.; Hu, Y.; and Wang, R. 2023b. GeoMVSNet: Learning Multi-View Stereo With Geometry Perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21508–21518.

Zhu, J.; Peng, B.; Li, W.; Shen, H.; Zhang, Z.; and Lei, J. 2021. Multi-view stereo with transformer. *arXiv preprint arXiv:2112.00336*.