# Self-supervised Multi-view Stereo via Inter and Intra Network Pseudo Depth


Ke Qiu*
qiuke@pku.edu.cn
School of Electronic and Computer Engineering,
Peking University
Shenzhen, China

Yawen Lai*
alanlyawen@pku.edu.cn
School of Electronic and Computer Engineering,
Peking University
Shenzhen, China

Shiyi Liu
shy_liu11@pku.edu.cn
School of Electronic and Computer Engineering,
Peking University
Shenzhen, China

Ronggang Wang†
rgwang@pkusz.edu.cn
School of Electronic and Computer Engineering,
Peking University
Shenzhen, China



## ABSTRACT

Recent self-supervised learning-based multi-view stereo (MVS) approaches have shown promising results. However, previous methods primarily utilize view synthesis as the replacement for costly ground-truth depth data to guide network learning, still maintaining a performance gap with recent supervised methods. In this paper, we propose a self-supervised dual network MVS framework with inter and intra network pseudo depth labels for more powerful supervision guidance. Specifically, the inter network pseudo depth labels are estimated by an unsupervised network, filtered by multi-view geometry consistency, updated iteratively by a pseudo depth supervised network, and finally refined by our efficient geometry priority sampling strategy. And we dynamically generate multi-scale intra network pseudo labels inside our cascade unsupervised network during training to provide additional reliable supervision. Experimental results on the DTU and Tanks & Temples datasets demonstrate that our proposed methods achieve state-of-the-art performance among unsupervised methods and even achieve comparable performance and generalization ability with supervised adversaries.



## CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**; **Matching**; *Scene understanding*.

## KEYWORDS

3d reconstruction, multi-view stereo, self-supervised, pseudo label


*Both authors contributed equally to this research.
†Correspondence author.



*MM '22, October 10–14, 2022, Lisboa, Portugal*



ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548212

**ACM Reference Format:**
Ke Qiu, Yawen Lai, Shiyi Liu, and Ronggang Wang. 2022. Self-supervised Multi-view Stereo via Inter and Intra Network Pseudo Depth. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548212


## 1 INTRODUCTION

Multi-view stereo (MVS) aims to reconstruct the observed 3D scene structure from overlapping multi-view images and camera parameters, which has been a topic of long-standing interest in computer vision for decades. While traditional methods have achieved impressive reconstruction performance, they still suffer from reflections, low-textured and specular regions leading to incomplete reconstructions. Recently, learning-based methods[10, 32, 36] adopt convolutional neural networks (CNNs) to obtain more complete and accurate matching for MVS, which have shown remarkable results and outperform most traditional methods on MVS benchmark datasets[1, 14]. However, most networks[10, 26, 36] greatly depend on vast quantities of ground-truth depth data to handily learn multi-view depth map inference. Merely collecting high-quality multi-view depth maps in large-scale scenes is a costly challenge.

Unsupervised learning-based MVS methods [7, 13] are proposed and attract increasing interest in recent years. These methods just learn multi-view depth maps inference directly from multi-view images, take the difference between the real target view and the views synthesized from source views as supervisory signals. Though previous unsupervised methods have shown promising results [11]. Supervised MVS methods [35, 41], driven by depth data, still maintain noticeable advantages. Empirically, the performance of MVS is tightly coupled with supervisory signals.

In this paper, we propose a self-supervised dual network MVS framework with inter and intra network pseudo depth labels for more comprehensive supervision guidance. Compared with previous unsupervised methods largely relying on single supervision, we aim to incorporate the advantages of view synthesis supervision and depth data supervision in a self-supervised framework. Our framework mainly consists of two distinct designs as follows:

A cascade unsupervised network supervised by multi-view images and intra network pseudo depth. The intuition is that the

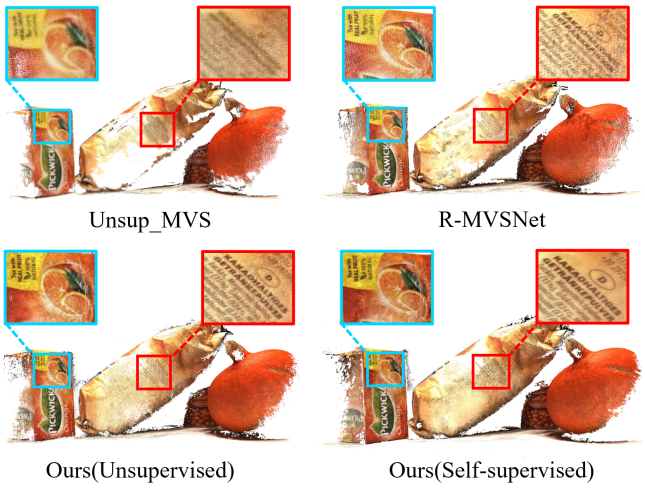


**Figure 1: Our reconstructed point clouds notably outperform existing unsupervised method Unsup_MVS[13], supervised method R-MVSNet[37] in completeness and accuracy. Best view on screen.**

resolution of training data is positively correlated with network performance because the supervisory signal from ambiguous low-resolution images is relatively weak. While our unsupervised network progressively estimates multi-scale depth maps in a coarse-to-fine manner, we treat the predictions of the training dataset as intra network pseudo depth to dynamically supervise the training of smaller parts of the cascade network. Therefore, intra network pseudo depth brings further supervision in form of a multi-scale depth consistency loss.

A similar network supervised by inter network pseudo depth. The key insight is that the high-quality predictions of the training dataset can handily bring effective supervision after filtering, also the room for optimization. Hence, we first create initial training depth data by trained unsupervised network and then leave the reliable part of our pseudo depth via a multi-view geometry consistency filter. After the pseudo depth is iteratively updated by the supervised network, we propose an efficient geometry priority sampling strategy to merge our inter and intra network pseudo depth. Multi-view geometry consistency filters with different tightness are adopted to judge the credibility of depth.

In summary, our main contributions are:

- We propose a self-supervised dual network MVS framework with inter and intra network pseudo depth labels, which brings more reliable and comprehensive supervision guidance.
- We propose a multi-scale depth consistency loss with intra network pseudo depth to bring further supervision for unsupervised MVS.
- We present an efficient training, filtering, and sampling strategy for iterative optimization of inter network pseudo depth.

Experiment results show that our proposed framework outperforms existing unsupervised MVS methods and achieves comparable performance and generalization ability with supervised adversaries.

## 2 RELATED WORK

Traditionally, a popular group of MVS methods is depth map based, which infers multi-view depth maps followed by a fusion process to obtain the point clouds. These methods [9, 15, 21, 23, 24, 30, 31] depend on the PatchMatch algorithm [2] to search the approximate pixel-wise correspondence. Zheng *et al.*[42] propose a probabilistic framework to jointly optimize the pixel-level view selection and depth estimation. And Schonberger *et al.*[23] present COLMAP, which further jointly infers pixel-wise depths and normals. ACMM[30] introduces an adaptive checkerboard sampling, multi-hypothesis joint view selection, and multi-scale patch matching scheme. These works greatly promote the development of MVS reconstruction.

**Supervised Learning-based MVS.** Different from traditional methods using hand-crafted features, learning-based MVS approaches [4–6, 10, 17, 18, 20, 26, 32, 33, 35–37, 39, 40] apply deep CNNs to learn more suitable features for better pair-wise patch matching. SurfaceNet[12] pre-warp the multi-view images to 3D space, and first adopt CNNs to regularize and aggregate the cost volume. Yao *et al.* propose MVSNet[36] to learn the depth maps by warping 2D image features to build the cost volume and applying 3D CNN for cost regularization. Recently, CascadeMVSNet[35] and CVP-MVSNet[10] significantly improve the depth prediction in a coarse-to-fine scheme and reduce the consumption of memory and time further. Therefore, the larger resolution depth map can be predicted in limited memory to obtain better reconstruction results. Although the pre-training models of these methods have certain generalization abilities, it is difficult to further obtain more ideal reconstruction results when they are directly applied to other scenes without ground truth for fine-tuning.

**Unsupervised Learning-based MVS.** Rather than using ground-truth depth, the unsupervised learning-based MVS methods leverage the underlying photometric and geometric consistency constraint as supervision. Khot *et al.*[13] present a robust multi-view photometric consistency loss between multiple images for learning multi-view depth maps. This method allows the network to implicitly handle lighting changes and occlusion across multiple views also leaves a lot of room for improvement, especially the network structure. Concurrently, Dai *et al.*[7] propose a multi-view symmetric unsupervised network MVS$^2$ and a cross-view consistency of multi-view depth building upon the symmetric network to detect occluded regions. Therefore, the learned multi-view depth maps naturally comply with the underlying 3D scene geometry. Due to the network predicting the depth map for all views simultaneously, the method consumes large GPU memory, which is hard to expand to more views. Huang *et al.*[11] present M$^3$VSNet with a multi-metric loss that combines pixel-wise and feature-wise loss to achieve better matching. Very recently, Yang *et al.*[34] carefully design a complex self-supervised pipeline with depth fusion, mesh generation, and depth rendering to improve pseudo depth. And Xu *et al.*[29] adopt Monte-Carlo Dropout to estimate the uncertainty

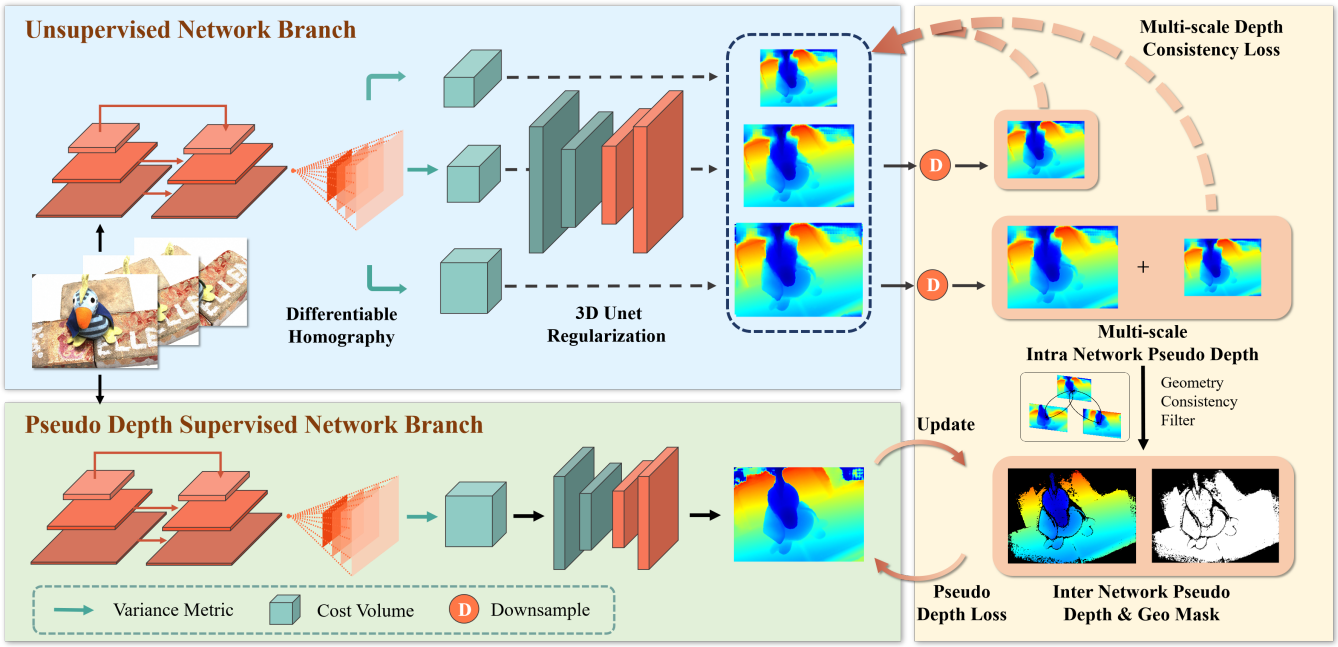


**Figure 2: Our self-supervised multi-view stereo framework. We dynamically generate the intra network pseudo labels for supervision during unsupervised network training. Then the initial inter network pseudo depth labels are estimated by trained unsupervised network and iteratively updated via self-training a pseudo depth supervised network.**

of depth prediction to generate the pseudo depth for post-training. These methods have shown a promising way for self-supervised MVS.

## 3 METHOD

In this section, we describe our proposed self-supervised MVS dual network framework, which introduces dual pseudo depth for more powerful supervision without the need for ground-truth data. We first introduce the approximate network structure design for both network branches in 3.1. Then, the intra network pseudo depth label design is illustrated in 3.2. Settings about inter network pseudo depth label and overall loss function are described in 3.3 and 3.4, respectively.

### 3.1 Network Architecture

We apply two network branches with similar structures to build the overall framework but set dissimilar supervision and training data individually, as shown in Figure 2. Specifically, the proposed cascade network mainly consists of three parts: feature pyramid extraction, cost volume construction, and cost volume regularization. Given $N$ input images of size $H \times W$, the first step is to extract multi-scale features of one reference image and $N-1$ source images via a common feature pyramid network[16]. The output feature maps of view $i$ are denoted as $\left\{F_i^l\right\}_{l=1}^{L} \in \mathbb{R}^{H/2^{l-1} \times W/2^{l-1} \times F}$, where $F$ and $l$ denote the feature dimension and the $l^{th}$ level in the feature pyramid. The next step is to apply classical plane sweeping based stereo pipeline and adopt differentiable homography[36] to wrap all source feature maps into $D$ hypothesis planes of the reference view. The homography $H_i$ between $i^{th}$ source image and the reference image at depth $d$ is defined as:

$$H_i(d) = K_i R_i \left( I - \frac{(t_1 - t_i) n_1^T}{d} \right) R_1^T K_1^{-1}, \tag{1}$$

where $I$ is the identity matrix, $\{K_i, R_i, t_i, n_i\}_{i=1}^{N}$ represent the corresponding camera intrinsics, rotations, translations, and camera principal axis of each view, respectively.

Subsequently, warped feature volumes are aggregated to one cost volume $C \in \mathbb{R}^{D \times H \times W \times F}$ via the variance metric. We adopt a cascade cost volume formulation to adaptively construct the cost volume at each level[10]. The predicted depth maps from the coarse levels are applied to adaptively update the depth hypothesis of the following level. Specifically, we set two reducing factors $w_R^l < 1$ and $w_I^l < 1$ to progressively narrow the depth hypothesis range $R^l$ and interval $I^l$ of the $l^{th}$ level to reduce time and memory consumption. Thus, the depth range and interval of the $(l+1)^{th}$ level are updated to $R^{l+1} = w_R^l \cdot R^l$ and $I^{l+1} = w_I^l \cdot I^l$.

After constructing a raw cost volume $C$ still contaminated by noise, a hybrid 3D U-Net network is applied to regularize it to generate a probability volume $P$. In this way, we infer the depth map $D$ from $P$ in a regression way. Specifically, we conduct Softmax operation $\sigma(\cdot)$ along the depth dimension to obtain the pixel-wise probability of each sampled depth $d$. The inference process is expressed as:

$$D = \sum_{d=d_{min}}^{d_{max}} d \times \sigma(P), \tag{2}$$

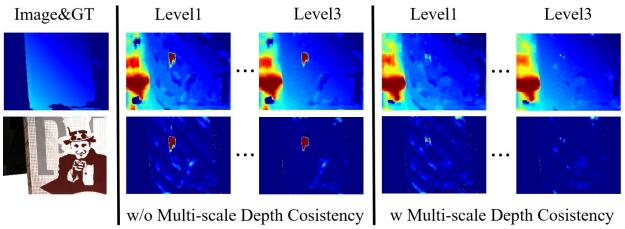


**Figure 3: Visualization of the effect of our proposed multi-scale depth consistency loss. The errors with large deviations have been corrected. Top row: Ground truth depth map and multi-scale depth predictions. Bottom row: Error maps of above predictions.**

where $d_{max}$ and $d_{min}$ denote the minimum and maximum depth value, respectively.

## 3.2 Intra Network Pseudo Depth Label

In the unsupervised network branch, we introduce loss evaluation from the multi-view view synthesis item and intra network pseudo depth item. Varying from existing unsupervised methods primarily relying on single view synthesis loss, we join extra depth data supervision for our unsupervised network. Considering that the supervisory signal mainly comes from the multi-view image pyramid, the resolution significantly affects the quality of depth estimation. As a result, the view synthesis supervision from ambiguous low-resolution images is relatively weak to generate a satisfactory coarse prediction. In the cascade network, some errors with large deviations in coarse depth maps are difficult to be completely rectified in the following levels.

To strengthen the supervision, we form intra network pseudo depth from multi-scale predictions and take high-resolution predictions as guidance to supervise the low-resolution network by downsampling operation. We sight that the predicted depth maps at finer scales significantly outperform predictions at coarse scales. In turn, the coarse depth maps with more sparse errors will further encourage the following network to generate better depth. Figure 3 shows the visualization of the effect of our proposed multi-scale depth consistency loss. It indicates that high-resolution predictions have significantly better quality and bring effective supervision to correct errors in the low-resolution depth.

Specifically, we design a multi-scale depth consistency loss to evaluate the consistency between multi-scale pseudo depth and network outputs. In practice, multi-scale depth predictions $\{D_l\}_{l=2}^{L}$ are first downsampled to each lower scale by interpolation. Next, the low-resolution depth predictions $D_l$ at scale $l$ are enforce to be consistent with one of the downsampled depth maps $\left\{D'_{s\to l}\right\}_{s=l+1}^{L}$. To further improve the robustness, we adopt the maximum operation to seek serious errors in multi-scale depth maps. The multi-scale depth consistency loss can be formulated as:

$$L_{mc} = max\left\{\left\|\left\{D'_{s\to l}\right\}_{s=l+1}^{L} - D_l\right\|_1\right\}_{l=1}^{L-1}, \tag{3}$$

which selects the most different pair $D_l$ and $D'_{s\to l}$.

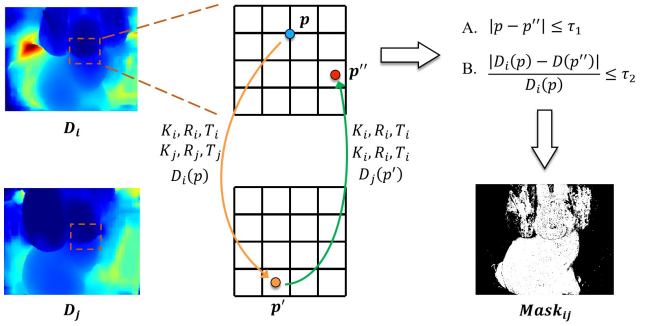


**Figure 4: Illustration of the pixel-wise depth re-projection error measurement process between reference view $i$ and source view $j$. We perform the error calculation between view $i$ and each corresponding source view.**

## 3.3 Inter Network Pseudo Depth Label

To dig for feature representations with stronger depth inference ability, we build an additional pseudo depth supervised network branch for iterative self-training to optimize inter network pseudo depth. Intuitively, the performance of MVS methods is tightly coupled with supervisory signals. Thus, given the trained unsupervised network, we first create high-quality inter network pseudo depth as training depth data. In the same way as intra network pseudo depth, we inference depth in the training set. In practice, our proposed pseudo depth optimization pipeline is comprised of three iterative stages: pseudo depth label filtering, iterative self-training, and geometry priority depth sampling.

**Pseudo Depth Label Geometry Consistency Filtering.** Given the raw inference results from our unsupervised network, we first construct an initial pseudo depth training dataset by geometry consistency filtering. We adopt the classical multi-view geometry consistency check pipeline and utilize depth re-projection error to measure the quality of pseudo depth labels. To judge the valid depth for each view, we perform the error calculation between one reference view and each corresponding source view and mark points that satisfy preset constraints.

In practice, we exploit pixel-wise depth re-projection error to obtain the valid mask of pseudo depth as shown in Figure 4. Given a reference view $D_i$ and a source view $D_j$, we first wrap each pixel $p$ in the view $i$ to view $j$ to obtain the corresponding pixel $p'$ and further get its depth $D_j(p')$ by bilinear interpolation. Then, by re-warping the pixel $p'$ back to the current view $i$, we can obtain pixel coordinate $p''$ and its re-projection depth $D(p'')$. Here the second projection process can be described by the following equation:

$$D(p'')K_i^{-1}p'' = T_{ij}D_j(p')K_j^{-1}p', \tag{4}$$

where $T_{ij}$ is the relative camera transform matrix from view $j$ to view $i$.

Ideally, the re-projection pixel coordinate $p''$ and depth $D(p'')$ should be consistent with pixel $p$ and depth $D_i(p)$. Therefore, the points that meet the constraints $|D(p'') - D_i(p)|\,/D_i(p) < \tau_1$ and $|p'' - p| < \tau_2$ in reference view $i$ are deemed to have passed the consistency check between view $i$ and $j$, where we set $\tau_1 = 0.01$ and $\tau_2 = 1$. Finally, we apply voting to integrate all the check results

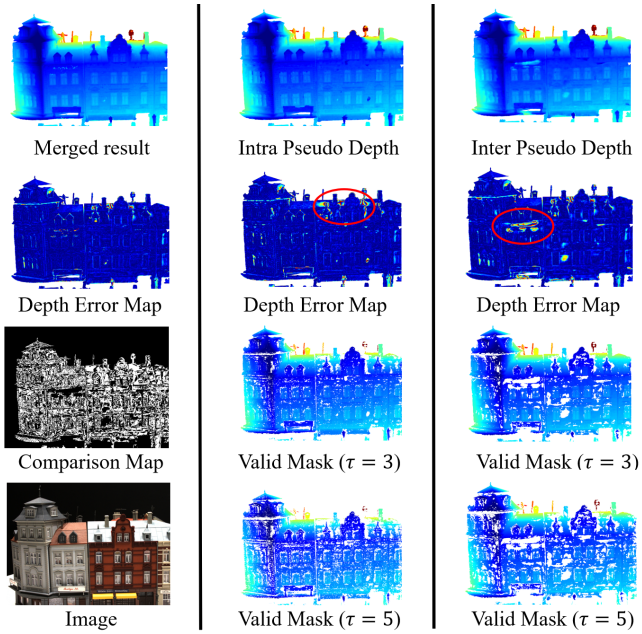


**Figure 5: Visualization results of our proposed efficient geometry priority pseudo depth sampling strategy. Merged reliable depth from two pseudo depth maps. Top row: Depth map before and after merging. Second row: Error maps of above depth. Third row: Quality comparison of two pseudo depth maps and filtered results after geometry check. Bottom row: Reference image and filtered results after stricter check.**

between each source views and reference view. More precisely, we mark points with less votes than the threshold $\tau_3 = 5$ as invalid.

**Iterative Self-training.** The pseudo depth supervised network branch follows the same cascade design as in our unsupervised network, but uses pseudo depth loss. After each iteration training, we regenerate the lastest pseudo depth dataset by inferring and filtering depth maps in the training set. Our pseudo depth loss can be formulated as:

$$L_{pseudo} = \sum_{l=1}^{L} \lambda^l \left\| D^l_{pseudo} - D^l \right\|_1 \odot M^l_{geo}, \tag{5}$$

where $D^l_{pseudo}$ and $D^l$ represent our inter network pseudo depth label and multi-scale predictions at the $l^{th}$ level, respectively. $M^l_{geo}$ is a binary mask that retains the reliable points in pseudo depth maps at the corresponding level.

**Geometry Priority Pseudo Depth Sampling.** To improve the quality of pseudo labels, we propose a refinement strategy that merges our inter and intra network pseudo depth. Figure 5 shows the visualization of geometry priority sampling. As shown in the comparison map, while the supervised network branch achieves better overall performance than our unsupervised network, intra network pseudo depth still owns higher quality in numerous local areas. To select reliable depth labels for refinement, we apply a series of geometry consistency filters with different thresholds to decide the priority of merging. In this way, we obtain several binary masks for each view, which indicate the credibility of the pseudo depth. Specifically, the priority of sampling is $M^{\tau 4}_{inter} > M^{\tau 4}_{intra} > M^{\tau 5}_{inter} > M^{\tau 5}_{intra}$, where we set the voting thresholds of geometry check $\tau_4 = 5$ and $\tau_5 = 3$, respectively. Such sampling process guarantees the unsatisfactory depth will be replaced and the final pseudo depth labels are consistent across multiple views.

## 3.4 Overall Loss Functions

Assume our hierarchical unsupervised network branch with $L$ levels generates a final depth map and $L-1$ intermedia predictions. To improve the quality of the final depth map, we design two aspects of unsupervised loss functions: 1) View synthesis loss computed at each output level which includes photometric loss and depth smoothness loss. 2) Multi-scale depth consistency loss for all outputs as discussed in 3.2. Given a reference image $I_r$ and several source images $(I_1, I_2, ..., I_n)$, the estimated depth maps are denoted by $\{D_l\}^L_{l=1}$, where $l$ indicates the $l^{th}$ output level. Similar to all unsupervised approaches, we create synthesized views for the supervision constraint based on photometric consistency. According to the depth and corresponding camera parameters, we generate $N$ synthesized views $\{I_{i\to r}\}^N_{i=1}$ at each output scale through warping source images into the reference view. Our total loss is defined as:

$$L_{unsup} = \sum_{l=1}^{L} \lambda^l \left( \alpha_{ph} L^l_{ph} + \alpha_{sm} L^l_{sm} \right) + \alpha_{mc} L_{mc}, \tag{6}$$

where $L_{ph}$, $L_{sm}$, $L_{mc}$ represent the photometric loss, depth smoothness loss, and multi-scale depth consistency loss. In practice, $\lambda^l = 2^l/4$, $\alpha_{ph} = 6$, $\alpha_{sm} = 0.18$, and $\alpha_{mc} = 0.2$.

**Photometric Loss.** Similar to [13], we adopt a Huber loss, a gradient consistency item, and the structure similarity SSIM [27] for similarity assessment between synthesized images $\{I_{i\to r}\}^N_{i=1}$ and the reference image $I_r$. Therefore, we obtain $N$ error maps from $N$ source views. To implicitly deal with occlusions, we only calculate the $K$ smallest loss values to enforce each point to be consistent with the best $K$ source views. The total photometric loss is:

$$L_{ph} = \sum_{p} \min_{1,2,...K} \sum_{i=1}^{N} (\left\| I_r - I'_{i\to r} \right\|_\epsilon + \left\| \nabla I_r - \nabla I'_{i\to r} \right\|_1 + \frac{\alpha}{2}(1 - SSIM\left(I_r, I'_{i\to r}\right))) \odot M, \tag{7}$$

where $p$ indexes over pixel coordinates, $M$ is a binary mask used to filter out invalid points that are projected outside the synthesized image bound, $\alpha$ and $K$ are set to 0.5 and 3. The threshold $\epsilon$ of Huber loss is set to 0.6.

**Depth Smoothness Loss.** An edge-aware term is adopted to enforce depth maps to be locally smooth in continuous regions of the image gradient, which is represented by:

$$L_{sm} = \frac{1}{m} \sum_{p} \left( |\nabla_x D| \, e^{-|\nabla_x I|} + |\nabla_y D| \, e^{-|\nabla_y I|} \right), \tag{8}$$

where $m$ denotes the total number of pixels in $I_r$.

# 4 EXPERIMENTS

In this section, we demonstrate the performance of proposed methods with a comprehensive set of experiments in standard MVS benchmarks including DTU[1] and Tank & Temples[14] datasets. We provide implementation details, qualitative and quantitative experimental results with analysis, and discussion about ablation studies.

## 4.1 Datasets

**DTU Dataset**[1] is a popular large-scale MVS benchmark consisting of 124 scenes scanned from 49 or 64 views under 7 different lighting conditions. The ground-truth point clouds and calibrated camera parameters are provided for each view. We adopt the same training, validation, and test sets as [7, 36].

**Tanks and Temples dataset**[14] contains indoor and outdoor realistic scenes with more challenging geometric layouts and lighting conditions. Each captured scene consists of hundreds of views. To validate the generalization ability of our method, we evaluate our method on the outdoor intermediate set using the model only trained on the DTU dataset [1] without any fine-tuning.

**BlendedMVS dataset[38]** is a large-scale dataset consisting of over 17k high-resolution images and 113 scenes covering various scenes including architectures and sculptures. Through the 3D reconstruction of large-scale scenes, Yao *et al.*[38] first build diversified mesh models and then obtain the images and depth maps from the 3D mesh rendering for MVS training.

## 4.2 Implementation details

During unsupervised training, we input 3 images with resolution $640 \times 512$ to our network with 3 hierarchy levels. Following the depth hypothesis strategies as [10], the number of depth hypotheses is set to 48, 32, 8, and the depth interval is narrowed to 50% at each level. Moreover, we adopt an image pyramid for supervision, which contains down-sampled images of 1/16, 1/4, and 1 of the input. During testing, we take 5 images with size $1600 \times 1184$ as input and reconstruct point clouds by depth fusion [9]. We quantitatively evaluate the cloud points by standard metrics [1]: accuracy and completeness, and overall score. After unsupervised training, we first generate pseudo depth of training sets and then conduct pseudo depth supervised training iteratively. Besides, in order to verify the generality of our proposed framework, we implemented the same scheme on the BlendedMVS dataset with the same parameter settings.

Our model is implemented in Pytorch [19]. We utilize a learning rate decay policy and Adam optimizer with initial learning rate $10^{-3}$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Our learning rate is reduced to 50% after 10,12,14 epochs. We train the network with two Nvidia Tesla V100s for 16 epochs with a batch size of 4.

## 4.3 Results on DTU Dataset

We first quantitatively compare our evaluation results against traditional geometric-based methods, recent learning-based supervised methods, and existing unsupervised methods including our unsupervised method and self-supervised method. As shown in Table 1, our self-supervised method outperforms other unsupervised MVS methods significantly in both completeness and overall score and achieves competitive performance compared to traditional and recent supervised methods.

**Table 1: Quantitative results of reconstruction quality on DTU dataset including mean accuracy, mean completeness, and the average of them (mm, lower is better). We mark noteworthy results by category (Unsupvised Methods & Others).**

| | Methods | Acc. ↓ | Comp. ↓ | Overall ↓ |
|---|---|---|---|---|
| Geometric | Furu [8] | 0.613 | 0.941 | 0.777 |
| | Tola [25] | 0.342 | 1.190 | 0.766 |
| | Camp [3] | 0.835 | 0.554 | 0.695 |
| | Gipuma [9] | **0.283** | 0.873 | 0.578 |
| | Colmap [22, 23] | 0.400 | 0.664 | 0.532 |
| Supervised | SurfaceNet [12] | 0.450 | 1.040 | 0.745 |
| | MVSNet [36] | 0.396 | 0.527 | 0.462 |
| | R-MVSNet [37] | 0.383 | 0.452 | 0.417 |
| | Point-MVSNet [4] | 0.361 | 0.421 | 0.391 |
| | CasMVSNet [10] | 0.325 | **0.385** | 0.355 |
| | CVP-MVSNet [35] | 0.296 | 0.406 | **0.351** |
| Unsup. | Unsup_MVS [13] | 0.881 | 1.073 | 0.977 |
| | MVS$^2$ [7] | 0.760 | 0.515 | 0.637 |
| | M$^3$VSNet [11] | 0.636 | 0.531 | 0.583 |
| | JDACS [28] | 0.571 | 0.515 | 0.543 |
| | Self-CVP (Self-Sup.) [34] | **0.308** | 0.418 | 0.363 |
| | **Ours (Unsup.)** | 0.398 | 0.381 | 0.389 |
| | **Ours (Self-Sup.)** | 0.408 | **0.316** | **0.362** |

The qualitative results of point clouds are shown in Figure 6, where notable details are highlighted in orange boxes. Compared with other unsupervised and supervised approaches, our model generates more complete and dense point clouds with finer details, especially for thin structures and textured surfaces such as windows, cartons, and the surfaces of the coffee can.

## 4.4 Generalization on Tanks & Temples

To verify the generalization ability of the proposed method, we use the models trained on the DTU dataset to reconstruct all scenes of the intermediate set of Tanks & Temples[14] without any fine-tuning. Besides, to verify the generality of our framework, we also finetune our model in BlendedMVS dataset [38] without using any ground-truth depth and then evaluate on Tanks & Temples[14] again (see Table 5). Then, we evaluate the F-score of point clouds via the official website [14]. As shown in Table 2, the results of the proposed unsupervised and self-supervised method obviously outperform existing unsupervised MVS methods and achieve comparable generalization ability with state-of-the-art supervised adversaries. As shown in Figure 7, from the enlarged details, we can see the excellent completeness and accuracy of reconstructed point clouds, which demonstrates the powerful generalization ability of our method. Specifically, the minute letters marked in orange boxes of *(f) Francis* can be seen clearly after zooming in.

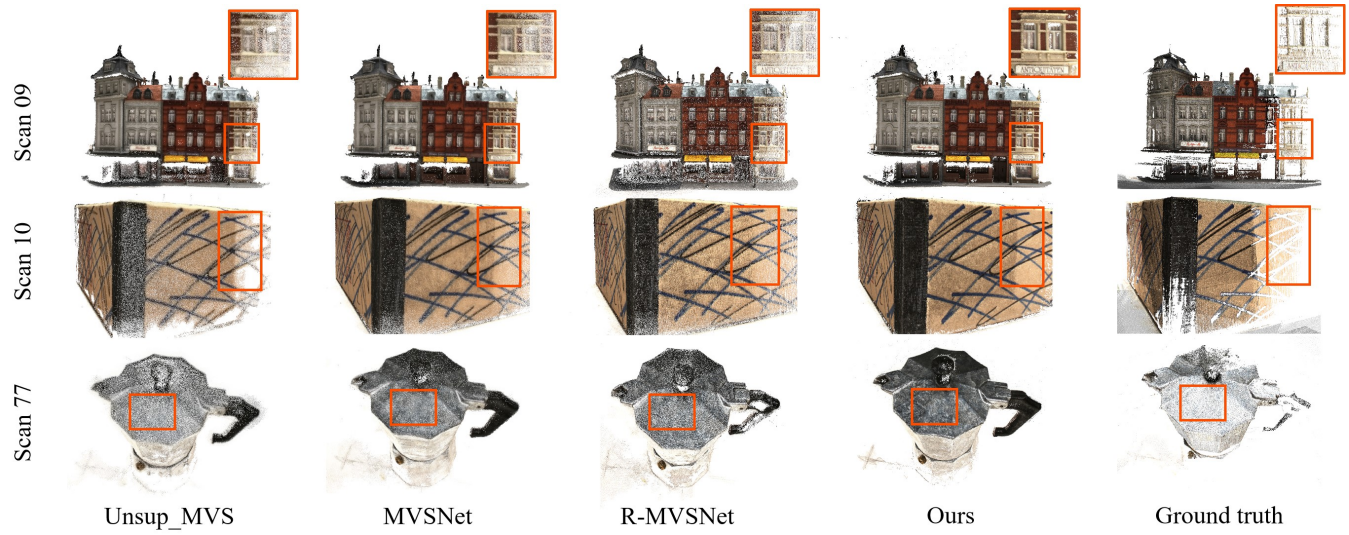


**Figure 6: Representative point cloud results on DTU dataset. Our method generates more complete and dense point clouds with finer details. Best viewed on screen.**

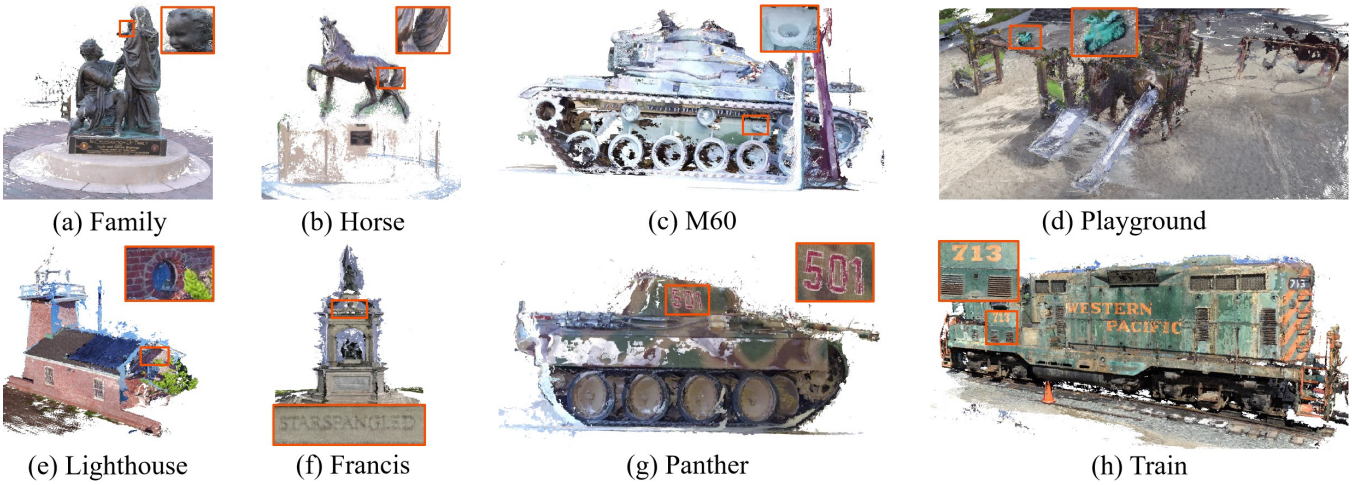


**Figure 7: Point clouds results of our method on the intermediate set of Tanks and Temples dataset without any fine-tuning.**

## 4.5 Ablation Study

We demonstrate the contribution from each module of the proposed framework through sufficient experiments by training on DTU Dataset[1], fine-tuning on BlendedMVS datasets[38] and testing on Tanks & Temples[14].

**Unsupervised MVS Loss.** To validate the effectiveness of the proposed intra network pseudo depth for unsupervised training, the ablation studies of loss items are conducted on the DTU test set [1] and use the same settings for evaluation. We analyze the contribution of each loss term on the quality of reconstruction. The quantitative results are summarized in Table 3. As shown, the proposed multi-scale depth consistency loss term (MC) brings a more critical improvement than the gradient consistency loss item (GC) from 0.407 to 0.389, which demonstrates the effectiveness of our multi-scale unsupervised loss design.

**Iterative Self-training Strategy.** To analyze the effect of the proposed iterative training, filtering, and sampling strategy. We conduct the ablation studies for the proposed geometry priority sampling strategy, and two crucial parameters including the multi-view geometry consistency threshold, the number of iterations for self-training. Results are listed in Table 4. As shown, both geometry filters with stricter settings and more iterations bring performance improvements. While using the depth sampling strategy, we obtain results with the best completeness and accuracy.

**Fine-tuning on BlendedMVS.** To verify the generality of the proposed framework, we additionally conducted sufficient experiments on the BlendedMVS dataset [38] without using any ground

**Table 2: Quantitative comparison on Tanks and Temples benchmark. The metric is the mean F score. Our results clearly outperform existing unsupervised MVS networks and achieve comparable generalization ability with state-of-the-art supervised adversaries.**

| Methods | Supervised | Mean ↑ | Family | Francis | Horse | Lighthouse | M60 | Panther | Playground | Train |
|---|---|---|---|---|---|---|---|---|---|---|
| MVSNet [36] | ✓ | 43.48 | 55.99 | 28.55 | 25.07 | 50.79 | 53.96 | 50.86 | 47.90 | 34.69 |
| R-MVSNet [37] | ✓ | 48.40 | 69.96 | 46.65 | 32.59 | 42.95 | 51.88 | 48.80 | 52.00 | 42.38 |
| Point-MVSNet [4] | ✓ | 48.27 | 61.79 | 41.15 | 34.20 | 50.79 | 51.97 | 50.85 | 52.38 | 43.06 |
| CasMVSNet [10] | ✓ | 56.84 | 76.37 | 58.45 | 46.26 | 55.81 | 56.11 | 54.06 | 58.18 | 49.51 |
| CVP-MVSNet [35] | ✓ | 54.03 | 76.50 | 47.74 | 36.34 | 55.12 | 57.28 | 54.28 | 57.43 | 47.54 |
| Vis-MVSNet [41] | ✓ | **60.03** | **77.40** | **60.23** | **47.07** | **63.44** | **62.21** | **57.28** | **60.54** | **52.07** |
| $MVS^2$ [7] | × | 37.21 | 47.74 | 21.55 | 19.50 | 44.54 | 44.86 | 46.32 | 43.48 | 29.72 |
| $M^3$VSNet [11] | × | 37.67 | 47.74 | 24.38 | 18.74 | 44.42 | 43.45 | 44.95 | 47.39 | 30.31 |
| JDACS [28] | × | 45.48 | 66.62 | 38.25 | 36.11 | 46.12 | 46.66 | 45.25 | 47.69 | 37.16 |
| Self-CVP[34](Self-Sup.) | × | 46.71 | 64.95 | 38.79 | 24.98 | 49.73 | **52.57** | **51.53** | 50.66 | 40.45 |
| **Ours (Unsupervised)** | × | 48.63 | 65.65 | 53.80 | 32.68 | 52.60 | 47.07 | 41.38 | **53.39** | 42.51 |
| **Ours (Self-supervised)** | × | **53.31** | **72.06** | **55.42** | **48.22** | **52.69** | 52.05 | 49.44 | 47.83 | **48.78** |

**Table 3: Ablation experiments with various loss functions. PC signifies the simple photometric consistency loss baseline without the gradient term (GC). Complete loss consists of PC, GC, and multi-scale depth consistency loss (MC).**

| Loss | Acc. ↓ | Comp. ↓ | Overall ↓ ($mm$) |
|---|---|---|---|
| PC | 0.411 | 0.417 | 0.414 |
| PC + GC | 0.412 | 0.403 | 0.407 |
| PC + GC + MC | **0.398** | **0.381** | **0.389** |

**Table 4: Comparisons of different iterations and thresholds of geometry check. $*$ indicates the geometry priority sampling is applied.**

| Iter. | Threshold | Acc. ↓ | Comp. ↓ | Overall ↓ ($mm$) |
|---|---|---|---|---|
| 1 | 3 | 0.430 | 0.334 | 0.382 |
| 1 | 5 | 0.408 | 0.342 | 0.375 |
| 2 | 3 | 0.412 | 0.403 | 0.377 |
| 2 | 5 | 0.416 | 0.317 | 0.367 |
| $2^*$ | 5 | **0.408** | **0.316** | **0.362** |

truth depth for fine-tuning both unsupervised and self-supervised training. Then, We directly conducted the evaluation on [14] without any fine-tuning. As illustrated in Table 5, fine-tuning on [38] can achieve better generalization especially the metric of recall.

## 5 CONCLUSION

In this paper, we propose a self-supervised dual network MVS framework with inter and intra network pseudo depth for more comprehensive supervision guidance. For the hierarchical unsupervised network, we first treat the multi-scale predictions as intra network pseudo depth to carefully bring extra supervision in form of a multi-scale depth consistency loss. Then, we create raw training depth data by the trained unsupervised network and leave the reliable part of them for the following network. Finally, an efficient training, filtering, and sampling strategy is designed for iterative optimizing the inter network pseudo depth. Our method significantly outperforms previous unsupervised methods and catches up with the performance of recent supervised methods. In the near future, we plan to extend our model to more stereo tasks and dynamic video sequences.

**Table 5: The ablation study of fine-tuning on BlendedMVS [38]. We evaluation the models on Tanks & Temples[14].**

| Methods | F Score ↑ | Precision ↑ | Recall ↑ |
|---|---|---|---|
| Unsup.(DTU) | 48.63 | **41.28** | 53.31 |
| Unsup.(Fine-tuning) | 51.32 | 39.14 | 77.72 |
| Self-Sup.(DTU) | 53.31 | 40.09 | 82.44 |
| Self-Sup.(Fine-tuning) | **53.80** | 40.48 | **82.90** |

## ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China U21B2012, 62072013 and 61902008, Shenzhen Cultivation of Excellent Scientific and Technological Innovation Talents RCJC20200714114435057, Shenzhen Research Projects of JCYJ20180503182128089 and 20180608092141929O.

# REFERENCES


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

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: a randomized correspondence algorithm for structural image editing. In *ACM Transactions on Graphics (TOG)*, Vol. 28. 24.

[3] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. 2008. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*. Springer, 766–779.

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

[5] Rui Chen, Songfang Han, Jing Xu, and Hao Su. 2020. Visibility-aware point-based multi-view stereo network. *IEEE transactions on pattern analysis and machine intelligence* 43, 10 (2020), 3695–3708.

[6] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. 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.

[7] Yuchao Dai, Zhidong Zhu, Zhibo Rao, and Bo Li. 2019. Mvs2: Deep unsupervised multi-view stereo with multi-view symmetry. In *2019 International Conference on 3D Vision (3DV)*. IEEE, 1–8.

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

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

[10] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. 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.

[11] Baichuan Huang, Can Huang, Yijia He, Jingbin Liu, and Xiao Liu. 2020. Mˆ 3VSNet: Unsupervised Multi-metric Multi-view Stereo Network. *arXiv preprint arXiv:2005.00363* (2020).

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

[13] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. 2019. Learning Unsupervised Multi-View Stereopsis via Robust Photometric Consistency. *arXiv preprint arXiv:1905.02706* (2019).

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

[15] Andreas Kuhn, Christian Sormann, Mattia Rossi, Oliver Erdler, and Friedrich Fraundorfer. 2020. DeepC-MVS: Deep confidence prediction for multi-view stereo reconstruction. In *2020 International Conference on 3D Vision (3DV)*. IEEE, 404–413.

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

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

[18] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. 2020. Attention-aware multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1590–1599.

[19] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).

[20] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. 2022. Rethinking Depth Estimation for Multi-View Stereo: A Unified Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8645–8654.

[21] Andrea Romanoni and Matteo Matteucci. 2019. Tapa-mvs: Textureless-aware patchmatch multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10413–10422.

[22] Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4104–4113.

[23] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*. Springer, 501–518.

[24] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*.

[25] Engin Tola, Christoph Strecha, and Pascal Fua. 2012. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications* 23, 5 (2012), 903–920.

[26] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. 2021. PatchmatchNet: Learned Multi-View Patchmatch Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14194–14203.

[27] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[28] Hongbin Xu, Zhipeng Zhou, Yu Qiao, Wenxiong Kang, and Qiuxia Wu. 2021. Self-supervised Multi-view Stereo via Effective Co-Segmentation and Data-Augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 2. 6.

[29] Hongbin Xu, Zhipeng Zhou, Yali Wang, Wenxiong Kang, Baigui Sun, Hao Li, and Yu Qiao. 2021. Digging Into Uncertainty in Self-Supervised Multi-View Stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 6078–6087.

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

[31] Qingshan Xu and Wenbing Tao. 2020. Planar prior assisted patchmatch multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12516–12523.

[32] Youze Xue, Jiansheng Chen, Weitao Wan, Yiqing Huang, Cheng Yu, Tianpeng Li, and Jiayu Bao. 2019. Mvscrf: Learning multi-view stereo with conditional random fields. In *Proceedings of the IEEE International Conference on Computer Vision*. 4312–4321.

[33] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. 2020. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European Conference on Computer Vision*. Springer, 674–689.

[34] Jiayu Yang, Jose M Alvarez, and Miaomiao Liu. 2021. Self-supervised Learning of Depth Inference for Multi-view Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7526–7534.

[35] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. 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.

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

[37] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. 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.

[38] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. 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.

[39] Hongwei Yi, Zizhuang Wei, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. 2020. Pyramid multi-view stereo net with self-adaptive view aggregation. In *European Conference on Computer Vision*. Springer, 766–782.

[40] Zehao Yu and Shenghua Gao. 2020. 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*. 1949–1958.

[41] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. 2020. Visibility-aware Multi-view Stereo Network. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press. https://www.bmvc2020-conference.com/assets/papers/0421.pdf

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