Digging into Uncertainty in Self-supervised Multi-view Stereo

Hongbin Xu[1, 2], Zhipeng Zhou[4], Yali Wang[1], Wenxiong Kang[2, 5], Baigui Sun[4], Hao Li[4], Yu Qiao[1,3*]
[1]ShenZhen Key Lab of Computer Vision and Pattern Recognition,
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]South China University of Technology, [3]Shanghai AI Laboratory, [4]Alibaba Group, [5]Pazhou Laboratory
hongbinxu1013@gmail.com yu.qiao@siat.ac.cn

Abstract

Self-supervised Multi-view stereo (MVS) with a pretext task of image reconstruction has achieved significant progress recently. However, previous methods are built upon intuitions, lacking comprehensive explanations about the effectiveness of the pretext task in self-supervised MVS. To this end, we propose to estimate epistemic uncertainty in self-supervised MVS, accounting for what the model ignores. Specially, the limitations can be categorized into two types: ambiguious supervision in foreground and invalid supervision in background. To address these issues, we propose a novel Uncertainty reduction Multi-view Stereo (U-MVS) framework for self-supervised learning. To alleviate ambiguous supervision in foreground, we involve extra correspondence prior with a flow-depth consistency loss. The dense 2D correspondence of optical flows is used to regularize the 3D stereo correspondence in MVS. To handle the invalid supervision in background, we use Monte-Carlo Dropout to acquire the uncertainty map and further filter the unreliable supervision signals on invalid regions. Extensive experiments on DTU and Tank&Temples benchmark show that our U-MVS framework[1] achieves the best performance among unsupervised MVS methods, with competitive performance with its supervised opponents.

1. Introduction

Multi-view stereo (MVS) [30] is a fundamental computer vision problem which aims to recover 3D information from multiple images on different views. Standing on the shoulder of giants in traditional methods [11, 29], recent learning-based methods [37, 38] have extended the MVS pipeline to deep neural networks, achieving state-of-the-art performance in several benchmarks [1, 20]. However, the fully supervised learning paradigm suffers from the non-negligible problem of requiring tedious and expensive procedures for collecting ground truth depth annotations. Hence, it leads the community to consider competitive alternative of learning-based approaches which require fewer labels.

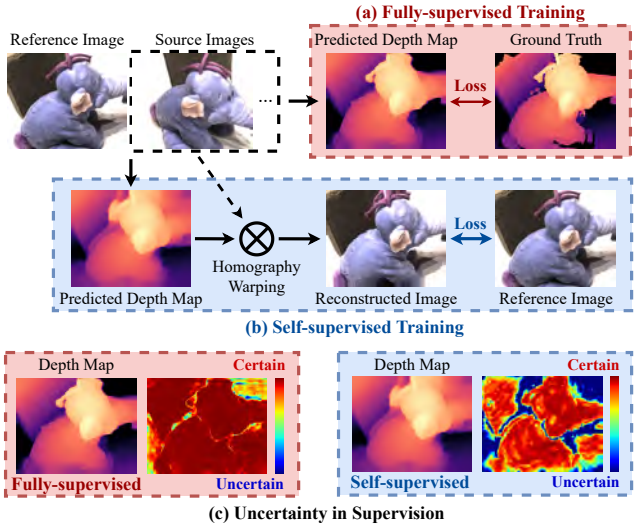


Figure 1. Illustration of the effectiveness of fully-supervised and self-supervised training in learning-based MVS via the visualization of *uncertainty in supervision*.

A prominent and appealing trend is to construct a self-supervised MVS task [12, 19, 7, 15], which further transforms the original depth estimation problem as RGB image reconstruction problem. However, previous methods are merely built upon intuitive motivations, lacking comprehensive explanations about which image regions such self-supervision signal can effectively work for multi-view depth estimation. For fully-supervised MVS (Fig. 1(a)), the regions where supervision exists are explicit, if given the ground truth depth map. Whereas, for self-supervised MVS shown in Fig. 1(b), the pretext task of image reconstruction actually provides indistinct supervision based on color similarity, which is agnostic to the exact presence of

*Corresponding author.
[1]Code: https://github.com/ToughStoneX/U-MVS

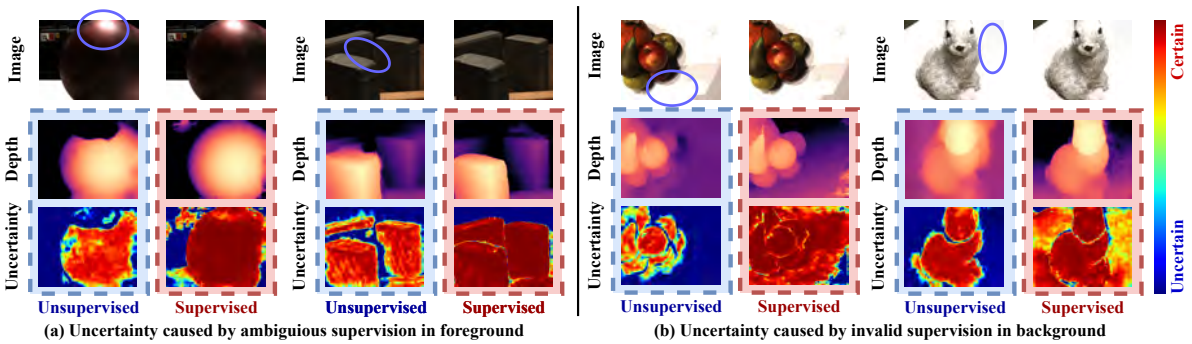


Figure 2. Visualization of epistemic uncertainty in fully-supervised and self-supervised MVS.

supervision in depth estimation. Hence, to provide a direct proof of the effectiveness in supervision, we utilize Monte-Carlo Dropout [18] to visualize the epistemic uncertainty for a comprehensive insight (Fig. 1(c)). In Bayesian modeling [8], the epistemic uncertainty inherently reflects what the supervision ignores.

What can we know from uncertainty? In Fig. 2, we provide a direct comparison of uncertainty between fully-supervised and self-supervised MVS to explicitly understand *what factors may lead to the failure of self-supervision*. It is found that the uncertain regions in self-supervision are more than the ones in fully-supervised training. Revisiting the premise of self-supervision as an image reconstruction task, the problem can be distinguished into two groups: (1) *Ambiguous supervision in foreground* (Fig. 2(a)). Under the influence of unexpected factors such as color variation and occlusion in the foreground object, the pretext task of image reconstruction is inconsistent with the photometric consistency and unable to reflect the correct depth information. (2) *Invalid supervision in background* (Fig. 2(b)). The textureless background provides no effective clues for depth estimation task, which is ignored in fully-supervised training. Whereas, the pretext task of image recostruction takes the whole image including the textureless backgrounds into consideration, involving invalid supervisions and oversmoothing the self-supervised results.

How to handle these uncertainties? To address these problems, we propose a novel Uncertainty reduction Multi-view Stereo framework U-MVS for self-supervised learning. It mainly consists of two distinct designs as follows: (1) To handle *ambiguous supervision in foreground*, we aim to append extra prior of correspondence to strengthen the reliability of self-supervision, and introduce a new multi-view flow-depth consistency loss. The intuition is that the dense 2D correspondence of optical flow can be utilized to regularize the 3D stereo correspondence in self-supervised MVS. A differentiable Depth2Flow module is proposed to convert the depth map to virtual optical flow among views and the RGB2Flow module unsupervisedly predict the optical flow from corresponding views. Then the virtual flow and the real flow are enforced to be consistent. (2) To handle *invalid supervision in background*, we suggest to filter the unreliable supervision signals on invalid regions, and propose an uncertainty-aware self-training consistency loss. In a totally unsupervised setting, we firstly annotate the dataset with a self-supervisedly pretrained model, while acquiring the uncertainty map with Monte-Carlo Dropout. Then the pseudo label filtered by the uncertainty map is used to supervise the model. Random data-augmentations on the input multi-view images are appended to enforce the robustness towards disturbance on the areas with valid supervision.

In summary, our contributions are: (1) We propose a novel self-supervised framework to handle the problems investigated from the visualization analysis about the uncertainty gap between supervised and self-supervised supervision signals. (2) We propose a novel self-supervision signal based on the cross-view consistency of optical flows and depth maps among arbitrary views. (3) We propose an uncertainty-aware self-training consistency loss for self-supervised MVS. (4) Experimental results on DTU and Tanks&Temples show that our proposed method can achieve competitive performance with its supervised counterparts with same backbones.

2. Related Work

Supervised Multi-view Stereo: With the flourishing of deep learning, convolutional neural networks (CNN) have now superseded classical techniques in Multi-view stereo. MVSNet [37] is a profound attempt that builds a standard MVS pipeline with end-to-end neural networks. They utilize 3D CNN to regularize the cost volume from features of CNN and get the depth map based on the soft-argmin operation from the output volume. Many efforts are further made to relieve the huge memory cost of cost volume. R-MVSNet [38] replace the 3D convolution with recurrent convolutional GRU unit. Many concurrent works are built on coarse-to-fine manner by separating the single cost volume regression into multiple stages, such as Fast-MVSNet [39], UCS-Net [6], CVP-MVSNet [36] and CascadeMVS-Net [14], achieving resounding success.

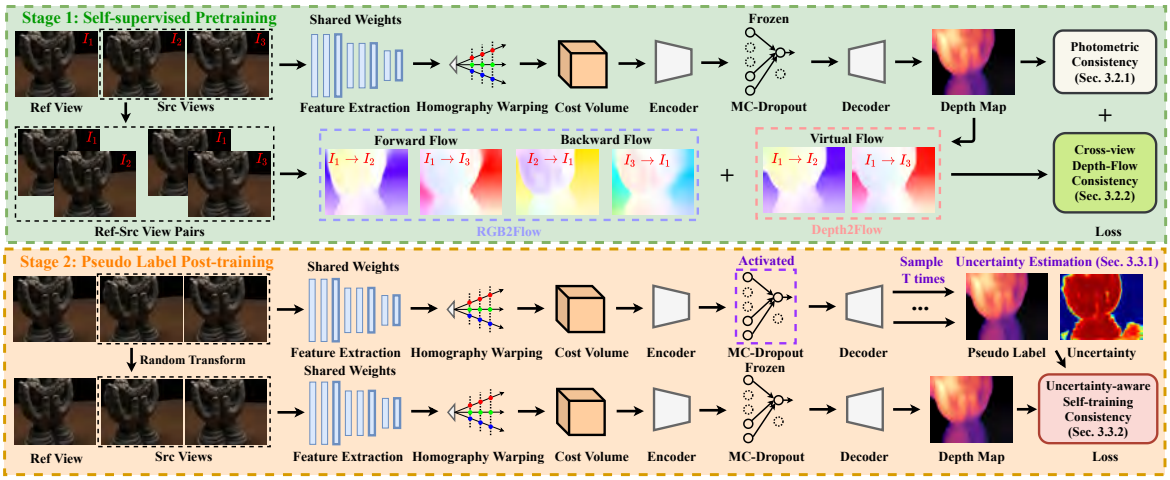


Figure 3. Illustration of our proposed self-supervised MVS framework U-MVS. "MC-Dropout" means Monte-Carlo Dropout.

Unsupervised / Self-supervised Multi-view Stereo: The burgeoning field of self-supervision [12] provide a competitive alternative for amazing performance and requiring no ground truth data. In Unsup_MVS [19], the predicted depth map and the input images are utilized to reconstruct the image on another view by homography warping, thus the photometric consistency is enforced to minimize the difference between the original and reconstructed images. MVS^2 [7] predicts the per-view depth maps simultaneously and automatically infer the occlusion relationship among views. M^{3}VSNet [15] enforce the consistency between surface normal and depth map to regularize the MVS pipeline. JDACS [34] revisit the color constancy hypothesis of self-supervision and propose a unified framework to enhance the robustness of self-supervision signal towards the natural color disturbance in multi-view images.

Uncertainty: The uncertainty [8] in deep learning models for vision tasks can be classified as aleatoric uncertainty and epistemic uncertainty. Aleatoric uncertainty captures the noise inherent in the training data, while epistemic uncertainty provides interpretation for the uncertainty in the model which can be remedied with enough data. [18] study the benefits of modeling epistemic and aleatoric uncertainty in Bayesian deep learning models for vision tasks. In this work, we aim to reject the unreliable pixels estimated by the epistemic uncertainty. Similar idea also appears in [27]. Confidence estimation is applied in MVS to filter the unreliable predictions, such as [21, 22]. UCS-Net [6] progressively reconstruct high-resolution depth map with a coarse-to-fine manner. The depth hypothesis of each stage adapts to the uncertainties of previous per-pixel depth predictions.

3. Method

In this section, we introduce the proposed self-supervised MVS framework, U-MVS. As Fig. 3 shows, the architecture of U-MVS is comprised of two stages: self-supervised pre-training stage and pseudo label post-training stage. The backbone model (Sec. 3.1) is firstly trained in the self-supervised pre-training stage (Sec. 3.2), and then trained in the pseudo label post-training stage (Sec. 3.3).

3.1. Backbone

Arbitrary MVS network can be utilized as backbone in our self-supervised MVS framework. In default, the representative MVSNet [37] is used. The network extracts feature from N input multi-view images and reprojects the feature maps in source views to the reference view by differentiable homography warping. The variance of the feature maps are used to construct a cost volume and a 3D U-Net is utilized to regularize the volume. Different from the standard 3D U-Net, we apply Monte-Carlo Dropout [18] on the bottleneck layer between the encoder and decoder, as shown in Fig. 3. In default, the Monte-Carlo Dropout is frozen when predicting depth map. It is only activated when estimating the uncertainty maps and pseudo labels.

3.2. Self-supervised Pre-training

The self-supervised pre-training stage contains two components of self-supervision loss: photometric consistency loss and cross-view depth-flow consistency loss. In the photometric consistency loss, the images on the source views are utilized to reconstruct the image on the reference view via homography warping relationship determined by the predicted depth map. As a solution to the *ambiguous supervision in foreground*, we add an extra branch of depth-flow consistency to endow extra correspondence prior to the self-supervision loss.

3.2.1 Photometric Consistency

The core insight of photometric consistency [2] aims at minimizing the difference between the real image and the syn-

thesized image from other views. It is denoted that the first view is the reference view and the $j(2 \leq j \leq V)$-th view is one of the $V-1$ source views. For a pair of multi-view images (I_1, I_j), it is attached with the intrinsic and extrinsic camera parameters $([K_1, T_1], [K_j, T_j])$. The output of a MVSNet backbone is the depth map D_1 on the reference view. We can compute the corresponding point position of pixel $\hat{p}_i$ in the source view j according to its position p_i^j in the reference view.

$$D_j(\hat{p}_i^j)\hat{p}_i^j = K_j T_j (K_1 T_1)^{-1} D_1(p_i) p_i \tag{1}$$

where $i(1 \leq i \leq HW)$ represents the index of pixels in the images. Since $D_j(\hat{p}_i^j)$ is a scale term in homogeneous coordinates, the $\hat{p}_i^j$ can be further described by the following equation:

$$\hat{p}_i^j = \text{Norm}[D_j(\hat{p}_i^j)\hat{p}_i^j] \tag{2}$$

where $\text{Norm}([x, y, z]^T) = [x/z, y/z, 1]^T$.

Then the synthesized image $\hat{I}_1^j$ from the source view j to the reference view can be calculated via differentiable bilinear sampling [16]. In Eq. 1, we can obtain a binary mask M_j indicating the valid corresponding pixels of I_j to the synthesized image $\hat{I}_1^j$. In a self-supervised MVS system, all source views are warped into the reference view to calculate the photometric consistency loss:

$$L_{pc} = \sum_{j=2}^{V} \frac{\|(I_1 - \hat{I}_1^j) \odot M_j\|_2 + \|(\nabla I_1 - \nabla \hat{I}_1^j) \odot M_j\|_2}{\|M_j\|_1} \tag{3}$$

where ∇I represents the gradient on x and y direction of image I and $\odot$ means point-wise product.

3.2.2 Cross-view Flow-Depth Consistency

As discussed in Sec. 1, one problem of basic self-supervised MVS is the *ambiguous supervision in foreground*. To handle this issue, we propose a novel flow-depth consistency loss to regularize the self-supervision loss. The flow-depth consistency loss is comprised of two modules: RGB2Flow and Depth2Flow, as shown in Fig. 3. In the Depth2Flow module, the predicted depth map is transformed as a virtual optical flow between the reference view and arbitrary source view. The whole Depth2Flow module is differentiable, which can be plugged in the training framework. In the RGB2Flow module, we use an unsupervised method [23] to predict the optical flow from corresponding reference-source view pairs. The forward and backward flows obtained from RGB2Flow module are required to be consistent with the virtual flow calculated from Depth2Flow module.

Depth2Flow Module: In a standard MVS system, the cameras are moving around the target object with fixed position when collecting the multi-view images. The relative motion of camera towards object can also be viewed as the motion of object towards a virtual camera with fixed position as shown in the Fig. 4. In this way, the correspondence can be represented by a dense 2D optical flow between arbitrary views and should be consistent with the 3D correspondence determined by homography warping in real MVS system. Given the definition of the aforementioned virtual optical flow: $\widehat{F}_{1j}(p_i) = \hat{p}_i^j - p_i$, where $\widehat{F}_{1j}(p_i)$ represent the optical flow between the corresponding point p_i in the reference view and $\hat{p}_i^j$ in the sourve view j. Considering the stereo correspondence defined in homography warping function (Eq. 1 and Eq. 2):

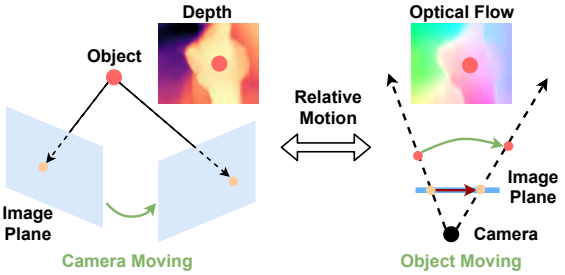


Figure 4. Intuition of Depth2Flow module. The relative motion of moving camera can be viewed as a special case of moving object represented by optical flow.

$$\widehat{F}_{1j}(p_i) = \text{Norm}[K_j T_j (K_1 T_1)^{-1} D_1(p_i) p_i] - p_i \tag{4}$$

With Eq. 4, the implicit correspondence modeled in the depth map can be explicitly transformed to the 2D correspondence of optical flow between the reference view and arbitrary source view j. This operation is fully differentiable which can be inserted into the training framework, namely Depth2Flow module in Fig. 3.

RGB2Flow module: We utilize a self-supervised method [23] to train a PWC-Net [31] on the dataset from scratch. All two-view pairs are enumerated among the provided multi-view pairs from the target MVS dataset. After unsupervisedly pretrained on the MVS dataset, the PWC-Net is used to predict the optical flow in the RGB2Flow module. As shown in Fig. 3, all two-view pairs combined with reference view and arbitrary source view are fed to RGB2Flow module. The output includes the forward flow and backward flow among each of the two views. The forward flow F_{1j} models the projection from reference view to source view j. In contrast, the backward flow F_{j1} represents the optical flow from source view j to reference view.

Loss function: The predicted depth map D_1 can be converted to virtual cross-view optical flow $\widehat{F}_{1j}$ by Depth2Flow module. The output of RGB2Flow module is forward flow F_{1j} and backward flow F_{j1}, which should be consistent with the virtual flow $\widehat{F}_{1j}$. For non-occluded pixels, the forward flow F_{1j} should be the inverese of the backward flow F_{j1}. To avoid learning incorrect deformation in occluded pixels, we mask out the occluded parts via the occlusion

mask O_{1j} infered by forward-backward consistency check [26]:

$$O_{1j} = \{|F_{1j} + F_{j1}| > \epsilon\} \tag{5}$$

where the threshold ϵ is set to 0.5. Then, the flow-depth consistency loss can be calculated:

$$L_{fc} = \sum_{i=1}^{HW} \min_{2 \le j \le V} \frac{\|(F_{1j}(p_i) - \widehat{F}_{1j}(p_i)) \cdot O_{1j}(p_i)\|_2}{\sum_{i=1}^{HW} O_{1j}(p_i)} \tag{6}$$

At each pixel, instead of averaging the difference between $\widehat{F}_{1j}$ and F_{1j} on all source views, we use the minimum error. The minimum error was firstly introduced in [13] to reject occluded pixels in self-supervised monocular depth estimation. Since the unsupervised RGB2Flow module may generate noisy predictions of optical flows, we utilize the minimum error to reject unreliable optical flow among views.

3.2.3 Overall Loss

In self-supervised pre-training stage, the overall loss is comprised of the photometric consistency loss L_{pc} and the flow-depth consistency loss L_{fc}:

$$L_{ssp} = L_{pc} + \lambda L_{fc} \tag{7}$$

where λ is a weight to balance the scale of L_{fc}, which is set to 0.1 in default. The flow-depth consistency loss aims to involve extra correspondence regularization to enhance the robustness of self-supervision loss towards real-world disturbances.

3.3. Pseudo-Label Post-training

To handle the aforementioned problem of *invalid supervision in background* in Sec. 1, the invalid regions like textureless backgrounds are ignored in the pseudo-label post-training stage. The uncertainty maps are estimated from the pretrained model in self-supervised pre-training stage via Monte-Carlo Dropout [18]. Then, the normalized uncertainty mask is adopted to filter the uncertain regions when calculating the uncertainty-aware self-training loss.

3.3.1 Uncertainty Estimation

In practice, the predictive uncertainty conveys skepticism about a model's output. As discussed in Sec. 3.1, Monte-Carlo Dropout [18] is applied to the bottleneck layers in the 3D U-Net of the backbone. Following the modification strategy suggested by [18], the original photometric consistency loss is modified as follows:

$$L'_{pc} = \sum_{j=2}^{V} \frac{\|(I_1 - \hat{I}_1^j) \odot M'_j\|_2 + \|(\Delta I_1 - \Delta \hat{I}_1^j) \odot M'_j\|_2}{\|M'_j\|_2} + \frac{1}{2} \log \Sigma^2 \tag{8}$$

where Σ^2 is the predicted variance of data noise, which is also called aleatoric uncertainty. Since Σ^2 is pixelwise uncertainty, the weighted mask is calculated by: $M'_j = \frac{1}{2} \exp(-\log \Sigma^2) \odot M_j$. Then the self-supervision loss (Eq. 7) is further modified as follows:

$$L'_{ssp} = L'_{pc} + \lambda L_{fc} \tag{9}$$

In our framework, a 6-layer CNN directly predicts the pixelwise aleatoric uncertainty Σ^2 from the input image. Then, the model is pre-trained with modified loss L'_{ssp} in Eq. 9 in the self-supervised pre-training stage.

Random Monte-Carlo Dropout [18] plays a role in sampling different model weights: $\mathbf{W}_t \sim q_\theta(\mathbf{W}, t)$, where $q_\theta(\mathbf{W}, t)$ is the random Dropout distribution in each sample. Denote that in the t-th time of sampling, with a model weight of $\mathbf{W}_t$, the predicted depth map is $D_{1,t}$. For our depth regression loss, the epistemic uncertainty is captured by the predictive variance of sampled depth maps:

$$U_1 = \frac{1}{T} \sum_{t=1}^{T} D_{1,t}^2 - (\frac{1}{T} \sum_{t=1}^{T} D_{1,t})^2 + \frac{1}{T} \sum_{t=1}^{T} \sigma_t^2 \tag{10}$$

where $(D_{1,t}, \sigma_t)_{t=1}^T$ is the sampled outputs with random Monte-Carlo Dropout. The mean prediction $\overline{D}_1 = \frac{1}{T} \sum_{t=1}^{T} D_{1,t}$ of the T sampled outputs is treated as the pseudo label.

3.3.2 Uncertainty-aware Self-training Consistency

To alleviate the invalid supervision in background, we utilize the generated pseudo label and uncertainty map in the previous section to construct an uncertainty-aware self-training consistency loss. A binary uncertainty mask $\widehat{U}_1$ can be calculated after normalizing the predicted uncertainty U_1:

$$\widehat{U}_1 = \{exp(-U_1) > \xi\} \tag{11}$$

where $\xi = 0.3$ is the threshold for calculating the binary mask $\widehat{U}_1$, which only retains the certain regions in self-supervison. Then, the uncertainty-aware self-training consistency loss can be calculated:

$$L_{uc} = \frac{\|(D_{1,\tau} - \overline{D}_1) \odot \widehat{U}_1\|_2}{\|\widehat{U}_1\|_1} \tag{12}$$

where $D_{1,\tau}$ represent the output of a randomly transformed multi-view images. All images (I_1, I_j) are transformed by data-augmentation operations (τ_1, τ_j) randomly. In the framework, we utilize standard data-augmentation operations [34] which do not move pixels, such as color jitter, gamma correction, random crop and etc. The output of the augmented input is required to be consistent with the pseudo label $\overline{D}_1$ on the valid regions filtered by $\widehat{U}_1$.

3.4. Overall Training Procedure

As shown in Fig. 3, our proposed self-supervised framework, U-MVS, is comprised of two stages: self-supervised

	Method	Acc.	Comp.	Overall
Geo.	Furu [10]	0.613	0.941	0.777
	Tola [32]	0.342	1.190	0.766
	Camp [3]	0.835	0.554	0.694
	Gipuma [11]	0.283	0.873	0.578
Sup.	Surfacenet [17]	0.450	1.040	0.745
	MVSNet [37]	0.396	0.527	0.462
	CIDER [35]	0.417	0.437	0.427
	P-MVSNet [24]	0.406	0.434	0.420
	R-MVSNet [38]	0.383	0.452	0.417
	Point-MVSNet [5]	0.342	0.411	0.376
	Fast-MVSNet [39]	0.336	0.403	0.370
	CascadeMVSNet [14]	0.325	0.385	0.355
	UCS-Net [6]	0.330	0.372	0.351
	CVP-MVSNet [36]	0.296	0.406	0.351
	PatchMatchNet [33]	0.427	0.277	0.352
UnSup.	Unsup_MVS [19]	0.881	1.073	0.977
	MVS^2 [7]	0.760	0.515	0.637
	M^3VSNet [15]	0.636	0.531	0.583
	Meta_MVS [25]	0.594	0.779	0.687
	JDACS[34]	0.571	0.515	0.543
	Ours+MVSNet	0.470	0.430	0.450
	Ours+CascadeMVSNet	0.354	0.3535	0.3537

Table 1. Quantitative results on DTU evaluation benchmark. "Geo."/"Sup." /"Unsup." are respectively the abbreviation of Geometric/Supervised/Unsupervised methods.

pretraining and pseudo-label post-training. In the first stage of self-supervised pre-training stage, the overall loss L_{ssp} includes photometric consistency loss L_{pc} and flow-depth consistency loss L_{fc}. As suggested by [18], the uncertainty is involved in L_{ssp} to construct the modified loss L'_{ssp}, which is used for training. In the second stage of pseudo-label post-training stage, the pseudo-label and uncertainty map are estimated from the pre-trained model in previous stage via Monte-Carlo Dropout [18]. In the uncertainty-aware self-training loss L_{uc}, the pseudo-label filtered by the uncertainty map is used to supervised the model. Standard random data-augmentation operations are involved in the post-training stage.

4. Experiment

Dataset: DTU [1] is a large-scale indoor MVS dataset collected by robotic arms. For each of the 124 scenes in total, high-resolution images are captured on 49 different views with 7 controlled light conditions. Tanks&Temples [20] is a outdoor MVS dataset, which contains challenging realistic scenes. Following the official split of MVSNet [37], we train the model on DTU training set and test on the DTU evaluation set. To validate the generalization performance of the proposed method, we test it on the *intermediate* and *advanced* partition of Tanks&Temples without any finetuning.

Error Metrics: In the DTU benchmark, *Accuracy* is measured as the distance from the result to the structured light reference, encapsulating the quality of reconstruction; *Completeness* is measured as the distance from the ground truth reference to the reconstructed result, encapsulating how much of the surface is captured; *Overall* is a the average of *Accuracy* and *Completeness*, acting as a composite error metric. In the Tanks&Temples benchmark, F-score in each scene is calculated following the official evaluation process.

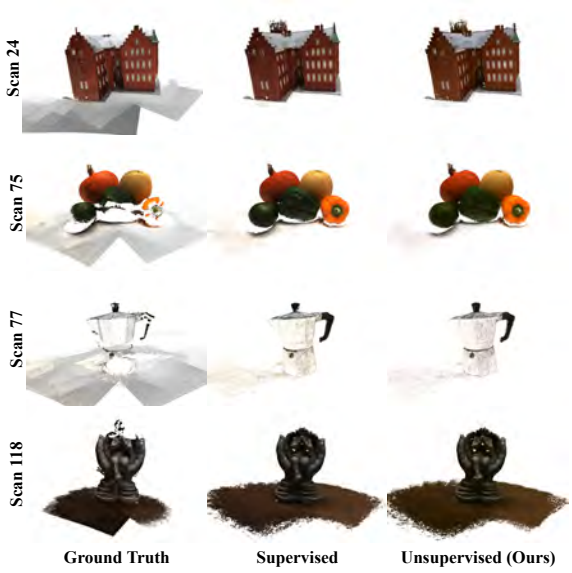


Figure 5. Qualitative comparison of the 3D reconstruction results on DTU evaluation benchmark. From left to right: Ground truth, results of SOTA supervised method, and our unsupervised method. CascadeMVSNet [14] is utilized as the backbone.

Implementation Details: The backbone of our U-MVS framework is inherited from the consice open implementations of MVSNet [37] and CascadeMVSNet [14]. In the preparation phase, we utilize a self-supervised method [23] to train an optical flow estimation network, PWC-Net [31], from the scratch on DTU dataset. The two-view pairs for optical flow estimation are selected by combining the reference view with each of the source views provided by MVS-Net [37]. Then, we utilize the self-supervised pretrained PWC-Net to estimate the optical flow from the aforementioned two-view pairs in the RGB2Flow module. More implementation details are provided in the supplementary materials.

4.1. Benchmark Results on DTU

Comparison with SOTA: To evaluate the performance of our proposed method, the quantitative results on the evaluation set of DTU benchmark [1] are presented in Table 1. In the table, state-of-the-art (SOTA) supervised and unsupervised methods are compared. From the figure, we can find that our proposed method performs better than previous unsupervised method. Under the error metric of *overall* in DTU benchmark, the performance of current SOTA su-

L_{pc}	L_{fc}	L_{uc}	Acc.	Comp.	Overall
✓			0.5527	0.5345	0.5436
✓	✓		0.5063	0.4576	0.4820
✓	✓	✓	**0.4695**	**0.4308**	**0.4501**

Table 2. Ablation study of different components of our proposed self-supervision framework using MVSNet as backbone.

L_{pc}	L_{fc}	L_{uc}	Acc.	Comp.	Overall
✓			0.4442	0.3641	0.4041
✓	✓		0.3745	0.3833	0.3789
✓	✓	✓	**0.3540**	**0.3535**	**0.3537**

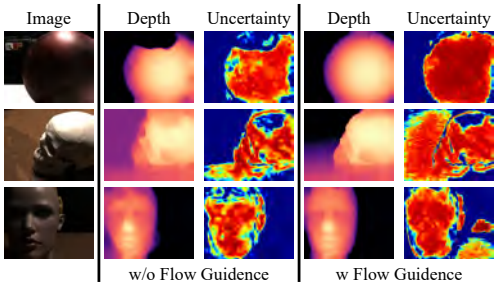
Table 3. Ablation study of different components of our proposed self-supervision framework using CasMVSNet as backbone.

pervised methods is about 0.351 - 0.355. Whereas, without utilizing any ground truth labels, our unsupervised model with a backbone of CascadeMVSNet can achieve 0.3537 on *overall* metric, which is comparable with supervised components. Fig. 5 shows the qualitative comparisons of the 3D reconstruction results on several scenes of DTU evaluation set. With the same CascadeMVSNet as backbone, our self-supervision framework can achieve a comparable performance with the supervised training.

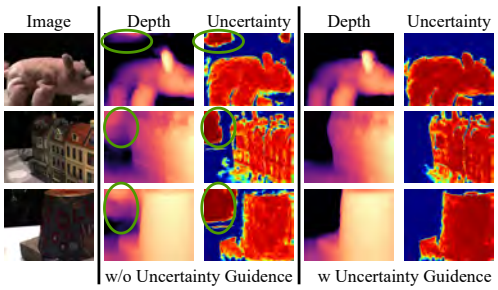
Supervised vs Self-supervised: To provide a fair comparison with the same backbone, we compare our proposed self-supervised MVS framework with the supervised training methods on MVSNet and CascadeMVSNet. The performance of supervised baselines are taken from previous papers (MVSNet [37], CascadeMVSNet [14]). From the italics in Table 1, it demonstrates that our self-supervised framework can perform slightly better than its supervised counterpart in an equal setting.

Ablation Studies: To evaluate the effect of different self-supervised components in the proposed framework, we respectively train the model with different combinations of the self-supervised losses. With a MVSNet as backbone, the quantitative results are presented in Table 2. With a CascadeMVSNet as backbone ,the ablation results are presented in Table 3. L_{pc}, L_{fc}, L_{uc} represent the basic photometric consistency loss (Eq. 3), flow-depth consistency loss (Eq. 8), uncertainty-aware self-training consistency loss (Eq. 12) respectively. From the tables, we can find that these self-supervised components can effectively improve the performance on all metrics.

Uncertainty Visualization: To find out whether the proposed self-supervised components can handle the aforementioned issues of uncertainties in foreground and background in Sec. 1, we provide the visualization results of the uncertainty estimated by Monte-Carlo Dropout in Fig. 6. For the first question, the uncertainty maps of the models respectively trained with or without our proposed flow-depth consistency loss L_{fc} are presented in Fig. 6(a). With the guidance of the dense 2D correspondence in flow-depth consistency loss, it is found that the certain regions in self-supervision become larger and more certain. It demonstrates that effective supervision towards disturbance such as reflection and low-texture is involved via the extra correspondence prior of flow-depth consistency. For the second question, the uncertainty maps of the models respectively trained with or without uncertainty guidance in the self-training loss L_{uc} are shown in Fig. 6(b). From the figure, we can find that if the model is trained without the guidance of uncertainty, the interfused uncertain supervision may be mistaken for correct pseudo label, further misleading the self-supervision. With the guidance of uncertainty, the misleading effect is alleviated, as shown in Fig. 6(b). It shows that the proposed uncertainty-aware self-training loss can enhance the supervision signals and get rid of the negative effect of uncertain supervision signals in self-supervised MVS.



(a) Uncertainty visualization about the effect of flow-depth consistency loss



(b) Uncertainty visualization about the effect of uncertain-aware self-training loss

Figure 6. Visualization results of the uncertainty under the effect of our proposed flow-depth consistency loss L_{fc} (Eq. 4) and uncertainty-aware self-training loss L_{uc} (Eq. 12).

4.2. Generalization

In order to evaluate the generalization ability of the proposed method, we compare the performance of our proposed method with state-of-the-art supervised and unsupervised methods on Tanks and Temples benchmark. For a fair comparison, we utilize the model merely trained on DTU dataset without any finetuning to test on Tanks&Temples dataset. For evaluation, the input image is set to 1920×1056, and the number of views is 7. We use CascadeMVSNet as backbone without using any ground truth in the training phase. The quantitative comparisons of performance on the *intermediate* partition of Tanks and Temples benchmark is presented in Table 4. The experimental results in the ta-

Method	Sup.	Mean	Family	Francis	Horse	Lighthouse	M60	Panther	Playground	Train
OpenMVG [28] + MVE [9]	-	38.00	49.91	28.19	20.75	43.35	44.51	44.76	36.58	35.95
OpenMVG [28] + OpenMVS [4]	-	41.71	58.86	32.59	26.25	43.12	44.73	46.85	45.97	35.27
COLMAP [29]	-	42.14	50.41	22.25	25.63	**56.43**	44.83	46.97	48.53	42.04
MVSNet [37]	✓	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69
CIDER [35]	✓	46.76	56.79	32.39	29.89	54.67	53.46	53.51	50.48	42.85
R-MVSNet [38]	✓	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38
CVP-MVSNet [36]	✓	54.03	76.50	47.74	36.34	55.12	**57.28**	**54.28**	57.43	47.54
CascadeMVSNet [14]	✓	56.42	76.36	58.45	46.20	55.53	56.11	54.02	**58.17**	46.56
MVS^2 [7]	×	37.21	47.74	21.55	19.50	44.54	44.86	46.32	43.38	29.72
M^3VSNet [15]	×	37.67	47.74	24.38	18.74	44.42	43.45	44.95	47.39	30.31
JDACS [34]	×	45.48	66.62	38.25	36.11	46.12	46.66	45.25	47.69	37.16
Ours + CascadeMVSNet	×	**57.15**	**76.49**	**60.04**	**49.20**	55.52	55.33	51.22	56.77	**52.63**

Table 4. Quantitative results on the *intermediate* partition of *Tanks and Temples* benchmark without any finetuning. We present the *f-score* result of all submissions from the official leaderboard of *Tanks and Temples* benchmark.

Method	Sup.	Mean	Auditorium	Ballroom	Courtroom	Museum	Palace	Temple
COLMAP [29]	-	27.24	16.02	25.23	34.70	41.51	18.05	27.94
R-MVSNet [38]	✓	24.91	12.55	29.09	25.06	38.68	19.14	24.96
CIDER [35]	✓	23.12	12.77	24.94	25.01	33.64	19.18	23.15
CascadeMVSNet [14]	✓	**31.12**	19.81	**38.46**	**29.10**	**43.87**	27.36	28.11
Ours + CascadeMVSNet	×	30.97	**22.79**	35.39	28.90	36.70	**28.77**	**33.25**

Table 5. Quantitative results on the *advanced* partition of *Tanks and Temples* benchmark without any finetuning. We present the *f-score* result of all submissions from the official leaderboard of *Tanks and Temples* benchmark.

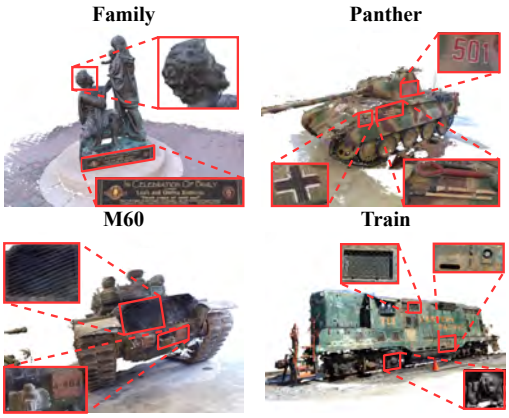


Figure 7. Visualization of the reconstructed 3D model on the *intermediate* partition of *Tanks and Temples* benchmark.

ble demonstrate that our proposed method has the highest score compared with unsupervised methods. Furthermore, the mean F-score on the *intermediate* benchmark is 57.15 which also outperforms previous supervised opponents including CascadeMVSNet. On the more complex *advanced* partition of Tanks and Temples benchmark, the comparison results are provided in Table 5. Without using any ground truth annotations, our proposed method can still present comparable performance with the SOTA supervised methods. The visualization results of the reconstructed 3D model on the *intermediate* partition of *Tanks and Temples* benchmark is provided in Fig. 7. Our proposed method achieves the best performance among unsupervised MVS methods on both partitions of *Tanks and Temples* benchmark until March 17, 2021.

5. Conclusions

In this paper, we have proposed a novel Uncertainty reduction Multi-view Stereo framework (U-MVS) for self-supervised learning, aiming to handle the two discovered problems via uncertainty analysis: 1) *Ambiguous supervision in foreground*; 2) *Invalid supervision in background*. For the first problem, we propose a flow-depth consistency loss to endow dense 2D correspondence of optical flows to regularize the 3D stereo correspondence in self-supervised MVS. For the second problem, we use Monte-Carlo Dropout to estimate the uncertainty map and filter the uncertain parts from supervision. The experimental results demonstrate the effectiveness of our proposed U-MVS framework.

6. Acknowledgement

This work is partially supported by National Natural Science Foundation of China (61876176, 61976095), the Science and Technology Service Network Initiative of Chinese Academy of Sciences(KFJ-STS-QYZX-092), Guangdong NSF Project (No. 2020B1515120085), the Shanghai Committee of Science and Technology, China (Grant No. 20DZ1100800 and 21DZ1100100). This work is supported by Alibaba Group through Alibaba Innovative Research (AIR) Program.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016.

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

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

[4] Dan Cernea. OpenMVS: Multi-view stereo reconstruction library. 2020.

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

[6] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020.

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

[8] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[9] Simon Fuhrmann, Fabian Langguth, and Michael Goesele. Mve-a multi-view reconstruction environment. In *GCH*, pages 11–18. Citeseer, 2014.

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

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

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[13] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[14] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020.

[15] Baichuan Huang, Can Huang, Yijia He, Jingbin Liu, and Xiao Liu. Mˆ 3vsnet: Unsupervised multi-metric multi-view stereo network. *arXiv preprint arXiv:2005.00363*, 2020.

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

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

[18] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[19] Tejas Khot, Shubham Agrawal, Shubham Tulsiani, Christoph Mertz, Simon Lucey, and Martial Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. *arXiv preprint arXiv:1905.02706*, 2019.

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

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

[22] Zhaoxin Li, Wangmeng Zuo, Zhaoqi Wang, and Lei Zhang. Confidence-based large-scale dense multi-view stereo. *IEEE Transactions on Image Processing*, 29:7176–7191, 2020.

[23] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6489–6498, 2020.

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

[25] Arijit Mallick, Jörg Stückler, and Hendrik Lensch. Learning to adapt multi-view stereo by self-supervision. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020. to appear.

[26] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[27] Christian Mostegel, Markus Rumpler, Friedrich Fraundorfer, and Horst Bischof. Using self-contradiction to learn confidence measures in stereo vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4067–4076, 2016.

[28] P Moulon, P Monasse, R Marlet, et al. Openmvg. an open multiple view geometry library.(2013).

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

[30] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006.

[31] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

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

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

[34] Hongbin Xu, Zhipeng Zhou, Yu Qiao, Wenxiong Kang, and Qiuxia Wu. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

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

[36] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020.

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

[38] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5534, 2019.

[39] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1949–1958, 2020.