Uncertainty Guided Multi-View Stereo Network for Depth Estimation

Wanjuan Su, Qingshan Xu, and Wenbing Tao, *Member, IEEE*

***Abstract*— Deep learning has greatly promoted the development of multi-view stereo in recent years. However, how to measure the reliability of the estimated depth map for practical applications and make reasonable depth hypothesis sampling for the cost volume building in the coarse-to-fine architecture are still unresolved crucial problems. To this end, an Uncertainty Guided multi-view Network (UGNet) is proposed in this paper. In order to enable the network to perceive the uncertainty, an uncertainty-aware loss function is introduced, which not only can infer uncertainty implicitly in an unsupervised manner but also can reduce the bad impact of high uncertainty regions and the erroneous labels in the training set during training. Moreover, an uncertainty-based depth hypothesis sampling strategy is further proposed to adaptively determine the depth search range of each pixel for finer stages, which helps to generate more rational depth intervals compared with other methods and build more compact cost volumes without redundancy. Experimental results on DTU dataset, BlendedMVS dataset, Tanks and Temples dataset and ETH3D high-res benchmark show that our method achieves promising reconstruction results compared with other state-of-the-art methods.**

***Index Terms*— Multi-view stereo, uncertainty estimation, depth estimation, 3D dense reconstruction, deep learning.**

I. Introduction

MULTI-VIEW Stereo (MVS) is a task of recovering 3D scene geometry from a collection of calibrated images, and has been widely used in autonomous driving, augmented and mixed reality, robotics, etc [1], [2]. Similar to other tasks in computer vision, deep learning techniques are introduced in MVS to improve the reconstruction performance. In recent years, learning-based MVS methods have made tremendous progress, even surpassing some traditional methods on some MVS benchmarks [3], [4]. When learning-based MVS methods are applied to a safety-critical system, such as autonomous driving, it is important to know the reliability of these methods because rash decisions may be made with overconfidence. For this reason, the network should have the ability to accurately identify the reliability of its output for deployment to practical applications.

Manuscript received 10 February 2022; revised 18 April 2022 and 12 May 2022; accepted 9 June 2022. Date of publication 16 June 2022; date of current version 28 October 2022. This work was supported by the National Natural Science Foundation of China under Grant 62176096 and Grant 61991412. This article was recommended by Associate Editor X. Zhang. *(Corresponding author: Wenbing Tao.)*

The authors are with the National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: suwanjuan@hust.edu.cn; qingshanxu@hust.edu.cn; wenbingtao@hust.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3183836.

Digital Object Identifier 10.1109/TCSVT.2022.3183836

However, existing learning-based MVS methods lack the mechanism to estimate this reliability. To solve this problem, it is necessary to introduce the uncertainty estimation into MVS to measure the reliability. The uncertainty can be divided into epistemic uncertainty and aleatoric uncertainty [5], [6]. The former captures the model's limited understanding of the input data which can be eliminated by providing enough data, while the latter describes inherent and irreducible noises in the input data which cannot be reduced even if more data are given. For MVS, aleatoric uncertainty is mainly caused by Lambertian surfaces, low-texture regions and illumination changes which are also the reasons for matching ambiguity; epistemic uncertainty is mostly brought by simplification of matching process and insufficient representation of training data. In addition, the estimation of epistemic uncertainty often relies on Monte Carlo sampling which drastically increases the computational requirements of the model [7], while aleatoric uncertainty is able to be directly predicted from the data. For this reason, we only consider aleatoric uncertainty in this paper. Some methods that combine traditional methods with deep learning use confidence maps to fit the uncertainty, they typically use an extra deep neural network to predict confidence maps from depth maps, cost maps and RGB images [8]–[10]. This can be regarded as a post-processing of depth estimation, which cannot effectively use the information in the process of depth estimation. More importantly, this will introduce additional memory consumption and computation time, and it requires Ground Truth (GT) of uncertainty to supervise the training process.

In contrast, we build a probabilistic neural network to jointly estimate the depth map and uncertainty map at the same time by introducing an uncertainty-aware loss function [7]. Specifically, to embed the uncertainty estimation into the multi-view stereo network, we use a two heads 3D CNN to regress depth and uncertainty respectively. Additionally, in order to enhance the representation ability of uncertainty, entropy operation and a shallow Convolutional Neural Network (CNN) are applied to the branch of uncertainty estimation. The data-dependent aleatoric uncertainty can be regarded as noises corrupting the GT, it generally is modeled by the Gaussian prior or the Laplacian prior [7]. Since the residual parts of the Gaussian prior and the Laplacian prior are composed of L2 norm and L1 norm respectively, and learning-based MVS methods are usually trained with L1 loss which is consistent with the residual part of the Laplacian prior, we use the Laplacian prior to model the aleatoric uncertainty. As a result, an uncertainty-aware loss function derived from the Laplacian negative log likelihood $-\log p(y|x) \propto \frac{\sqrt{2}}{\sigma}|y-x| + \log\sigma$ will be obtained, where x and

y are the input and output of the model, σ denotes the model's observation noise parameter (i.e., the uncertainty in this case). It can be seen that allowing the network to estimate uncertainty means allowing the network to adaptive temper the residual loss by σ. In other words, the uncertainty is acted as loss attenuation of depth residual. This makes the network more robust to noisy data, and even allows the network to learn to attenuate the effect from erroneous labels, so as to contribute to more accurate depth estimation. Moreover, the uncertainty can be learned in an unsupervised manner in this way, which does not require additional computation and memory.

Based on our above uncertainty estimation, we obtain a prior knowledge of the reliability of the estimated depth map. Next, we further investigate how to leverage the learned uncertainty to facilitate high-resolution depth map refinement. Recently, in order to enable the network to estimate high-resolution depth maps, the coarse-to-fine strategy [11], [12] is widely used in learning-based MVS methods [4], [13], [14]. Specifically, they first build cost volumes with large depth ranges to estimate depth maps at the coarsest resolution, then progressively narrow the depth ranges of cost volumes in the following stages according to the estimation of previous stage, finally produce depth maps with both high resolution and high accuracy. This kind of architecture can produce high-precision predictions with high resolution while reducing memory consumption. Nevertheless, existing methods either use the same depth hypothesis interval for all pixels in finer stages [13], [14] which may cause information redundancy or narrow search ranges, or use the probability volume to determine the depth hypothesis interval of each pixel [4] which may lead to unreasonable search ranges. Fig. 1 illustrates the estimated depth map, estimated uncertainty and error map on DTU and BlendedMVS datasets. As shown, erroneous pixels in depth maps also have high uncertainty in uncertainty maps, which further shows that estimated uncertainty maps are able to precisely reflect the reliability of estimated depth maps. Accordingly, the estimated uncertainty can be regarded as an important clue to judge whether the estimated depth is accurate or not. Intuitively, the reliable depth should be given a narrower depth search range in the next stage, while the unreliable depth should be given a wider depth search range. Based on this, we propose to utilize the estimated uncertainty to adaptively determine the depth search range for each pixel at finer stages to construct more compact cost volumes and further boost the reconstruction performance.

In this paper, we propose an Uncertainty Guided multi-view stereo Network (UGNet) for depth estimation that enables the network to perceive its prediction reliability and explores more reasonable depth hypothesis sampling. Extensive experiments have been conducted on BlendedMVS dataset [15], DTU dataset [16], Tanks and Temples dataset [17] and ETH3D high-res benchmark [18], the experimental results show that our method achieves comparable reconstruction results compared to other state-of-the-art methods. To summarize, our contributions are as follows:

- We present a probabilistic multi-view stereo network UGNet to make the network be aware of the reliability of estimated depth, so that MVS can be more practical for safety-related tasks;
- We introduce the uncertainty-aware loss function to optimize the network, which not only can learn uncertainty implicitly without supervision but also can improve the reconstruction performance by attenuating adverse impact of high uncertainty areas and the erroneous labels of depth maps during training;
- To construct more compact cost volumes, we propose an uncertainty-based depth hypothesis sampling strategy which can effectively and adaptively determine the depth search range for finer stages in the coarse-to-fine architecture.

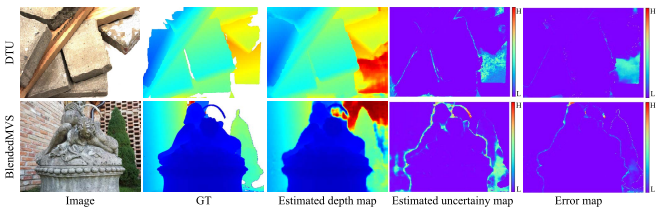


Fig. 1. Visualization of estimated depth map, estimated uncertainty and error map between estimated depth and GT, where the error map is obtained by using absolute depth error $|D_{gt}(x) - \hat{D}(x)|$. Note that since there are pixels without ground truth depth values in GT, the absolute depth errors of these pixels are inaccurate. We masked these areas with zeros in both estimated uncertainty map and error map, so as to show the relationship between their corresponding pixels more intuitively. The L and H denote the uncertainty/error is low and high, respectively.

II. Related Works

Methods of MVS can be roughly divided into point cloud-based methods [19], [20], voxel-based methods [21]–[24] and depth map-based methods [3], [25]–[29]. Since that our method is closely relevant to learning-based MVS methods for depth estimation, we will mainly review related literature of this filed and uncertainty estimation in stereo vision in this section.

A. Learning-Based Muiti-View Stereo

There have been significant progresses made by learning-based MVS methods compared with traditional ones [25]–[28] in recent years, owing to their much more powerful learning ability for discriminative representations than that of hand-crafted methods. The recent popular depth map-based pipeline of learning-based MVS method is MVSNet [3], which proposes an end-to-end architecture for depth map estimation by introducing the differentiable homography warping to build the cost volume. Based on this, an average group-wise correlation similarity measure is proposed to construct a lightweight cost volume in [30]. To reduce the memory consumption caused by the cost volume regularization, two kinds of methods are widely used. One uses the Recurrent Neural Network (RNN) [29], [31], [32] to regularize the cost volume instead of using CNN, which is a way of sacrificing the computational efficiency in exchange for the memory cost. On the contrary, the other one not only improves the performance and reduces the memory cost, but

also improves the computational efficiency, which introduces the coarse-to-fine strategy and reserves the CNN-based cost volume regularization [4], [13], [14], [33], [34].

When the coarse-to-fine architecture is adopted, the cost volume at the coarsest stage is constructed by uniform sampling depth hypotheses across the full depth range, and the depth map is refined iteratively by building a more compact cost volume according to the previous estimated depth with the proposed depth hypothesis sampling method. CVP-MVSNet [14] analyses the relationship between the resolution of images and depth search ranges, which determines the local depth search ranges by calculating the mean depth interval corresponding to 0.5 pixel distance on the epipolar line of the closest source images. The performance of this method is limited by the narrow search range. UCSNet [4] infers pixel-wise depth interval based on variance of the depth probability volume and depth residual. Since the depth probability volume may be over-confident or under-confident, the estimated depth interval may be too narrow or too wide, which is not conducive to predicting accurate depth maps. CascadeMVSNet [13] relies on the hyperparameters obtained from the experiments to control the depth interval. In addition, since the assigned depth interval of all pixels is the same in CascadeMVSNet [13], the depth search range for some pixels may be redundant. Different from these methods, we propose to leverage the uncertainty map to adaptively determine the depth interval for every pixel, which is more practical and helps to build a more compact cost volume without redundancy.

B. Uncertainty Estimation in Stereo Vision

As a key metric to measure the reliability of prediction, uncertainty has been widely studied in the field of stereo vision. For the traditional methods of stereo vision, uncertainty generally derived based on the characteristic of cost curve, a comprehensive review can refer to [35]. As for learning-based methods, uncertainty estimation methods can be divided into two categories. One uses the confidence to measure the uncertainty with an additional deep learning model, i.e., a score between 0 and 1 is assigned for each pixel. In [36], a deep CNN is designed to estimated the confidence from the disparity patches obtained by a stereo matching algorithm. The ConfNet [37] is proposed to exploit both local and global cues to predict confidence maps with the disparity maps and reference images. Similarly, Kim *et al.* [38] leverage disparity maps, matching costs and reference images to estimate confidence maps for stereo matching by the proposed locally adaptive fusion network. Specially, works in [8], [9] and [10] extend the methods in [36], [37] and [38] for the MVS task, respectively.

The methods mentioned above generally treat disparity/depth estimation and uncertainty estimation as two separate steps, which not only cannot make full use of the information in the process of disparity/depth estimation, but also increase the memory consumption and computation cost. More importantly, they require the GT of uncertainty for supervision which is usually not available in the public dataset. In contrast, the other changes the original stereo network into a probabilistic model [7], it estimates the task-related predictions and uncertainty jointly. Mehltretter and Heipke [39] introduce the probabilistic model into stereo matching for uncertainty estimation, which train a CNN based on the constructed cost volume for uncertainty prediction. But it regards the uncertainty estimation as a post-processing task of disparity estimation instead of handling both tasks jointly. Zhang *et al.* [40] infer pixel-wise matchability by the means of uncertainty from the entropy of probability volume. Inspired by this, Vis-MVSNet [41] measures the visibility of two-view images via estimating pair-wise matching uncertainty, where the uncertainty is estimated as done in [40]. Recently, to deal with ineffective supervision in the background, Xu *et al.* [42] introduce both aleatoric and epistemic uncertainty to filter out the invalid regions for self-supervised MVS. It is noteworthy that although methods in [41], [42] also introduce the uncertainty estimation into MVS, there are essential differences between our method and theirs. Vis-MVSNet [41] estimates uncertainty of the pair-wise depth map, which cannot tell the reliability of the final depth maps. Moreover, the uncertainty-based supervision in [41] is imposed on the intermediate output rather than the final output, so that it cannot attenuate adverse effect of the region with high uncertainty cross the entire depth estimation process like our method does. U-MVS [42] infers aleatoric uncertainty with an additional six-layer CNN applied to the input image, it mainly aims to identity the background area in the input image, while our method predict uncertainty from the cost volume directly to capture the uncertainty of the depth map.

Note that aforementioned methods infer aleatoric uncertainty only, except the method in [42]. Epistemic uncertainty usually requires sampling-based methods [7], [43]. As discussed in the Introduction (Section I), our work focuses on aleatoric uncertainty only, so epistemic uncertainty is out of scope in this paper and is not discussed further.

III. Method

Given a reference image X_0 and its $N-1$ source images $\{X_i\}_{i=1}^{N-1}$ with known camera parameters, we aim to predict the depth map D for the reference view. The framework of the proposed UGNet is illustrated in Fig. 2, the coarse-to-fine strategy is used in our method to estimate high-resolution depth maps. The whole network is trained with the introduced uncertainty-aware loss function, and the proposed uncertainty-based depth hypothesis sampling strategy is used to generate more compact cost volumes for finer stages.

This section will describe the proposed method in detail. We firstly give the detail of uncertainty-aware loss function in Section III-A; then, we will introduce the proposed uncertainty-based depth hypothesis sampling strategy in Section III-B; finally, the pipeline of the UGNet will be presented in Section III-C.

A. Uncertainty-Aware Optimization

The aleatoric uncertainty describes the noise ε inherent in the image, so that the observed GT depth D_{gt} can be assumed to be corrupted by the noise. We use the Laplacian prior to model it. In this case, the uncertainty is regarded as the

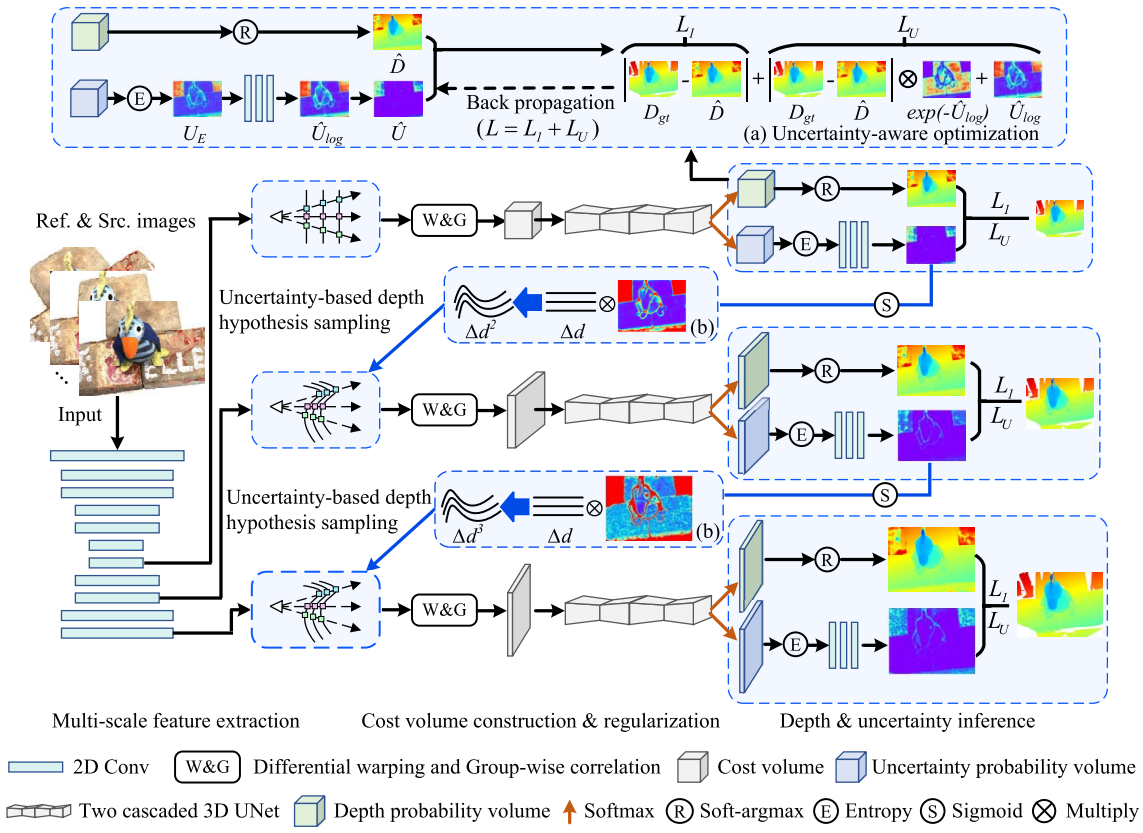


Fig. 2. The framework of UGNet, which adopts coarse-to-fine architecture to estimate the high-resolution depth map. (a) shows details of uncertainty-aware optimization (Section III-A), where the uncertainty-aware loss function $\mathcal{L}_U$ combined with $\mathcal{L}_1$ is utilized to guide the optimization of the network (notations are presented with Eq. (3) and Eq. (4)). (b) illustrates details of uncertainty-based hypothesis sampling strategy (Section III-B), which makes use of the estimated uncertainty map at the previous stage to adaptively determine the depth interval Δd^s ($s \in \{2, 3\}$) for the next stage based on the basic depth interval Δd (computed as Eq. (6)).

predictive variance, and the depth that the model intents to estimate is regarded as the mean. Therefore, the neural network for MVS incorporating the uncertainty can be formulated as:

$$D_{gt} = f(x, \theta) + \varepsilon, \tag{1}$$

where x denotes the input images, θ denotes the parameters of the model, and $f(\cdot)$ denotes the neural network for MVS. Since only aleatoric uncertainty is considered here, the ε is data-dependent, i.e., $\varepsilon = \varepsilon(x)$.

To enable the network to perceive the uncertainty, we assume $\varepsilon(x)$ follows the Laplacian distribution, and the Laplacian likelihood is introduced to model the uncertainty. So that the uncertainty-aware loss function derived from the negative log likelihood can be formulated as [7]:

$$\mathcal{L}_U = \frac{\sqrt{2}}{|\Omega|} \sum_{x \in \Omega} \frac{1}{\hat{U}} |D_{gt}(x) - \hat{D}(x)| + \log \hat{U}, \tag{2}$$

where Ω denotes valid pixels in D_{gt}, $|\Omega|$ denotes the number of valid pixels in Ω, and $\hat{U}$ denotes the predicted uncertainty. This function is composed of the residual regression weighted by uncertainty and the uncertainty regularization term. The former discourages the network to estimate low uncertainty for pixel with high residual error, while the latter tries to force the model to avoid estimating high uncertainty for all pixels. Note that the uncertainty is implicitly learned by this uncertainty-aware loss function in an unsupervised manner. Furthermore, the log uncertainty is predicted by the network in practice for numerical stability, namely, $\hat{U}_{log} = \log \hat{U}$, and the coefficient $\sqrt{2}$ is ignored for balance, so the uncertainty-aware loss function Eq. (2) for each stage is rewritten as:

$$\mathcal{L}_U^s = \frac{1}{|\Omega|} \sum_{x \in \Omega} |D_{gt}^s(x) - \hat{D}^s(x)| \exp(-\hat{U}_{log}^s) + \hat{U}_{log}^s, \tag{3}$$

where s denotes the s-th stage.

With the uncertainty-aware loss function above, the network needs to be able to predict the depth as well as uncertainty of each pixel. However, original neural network for MVS is used to estimate depth only, so the output of cost volume regularization module is one channel. Enlightened by [7], to estimate depth and uncertainty jointly, the output of cost volume regularization module is changed to two channels, the first one is for depth estimation, while the another one is for uncertainty estimation. To measure uncertainty more accurately, we form the uncertainty by using the entropy map of uncertainty probability volume with a shallow 2D CNN to increase the representation ability as suggested by [41]. Specifically, for the output channel of uncertainty estimation, the Softmax operation is first applied to get the uncertainty probability volume P_U, then entropy operation is used to obtain entropy map U_E. In order not to significantly increase

TABLE I
THE DETAILED LAYER CONFIGURATION OF THE SHALLOW NETWORK FOR UNCERTAINTY ESTIMATION. IF NOT SPECIFIED, EACH 2D CONVOLUTION LAYER IS FOLLOWED BY A BATCH NORMALIZATION (BN) AND A RECTIFIED LINEAR UNIT (RELU). S1 DENOTES THE CONVOLUTION STRIDE. "H × W" DENOTES THE SIZE OF THE INPUT ENTROPY MAP U_E. "∗" DENOTES NO BN AND RELU

Name	Layer Description	Output Size	Input
Conv1	Conv2D, 3×3, S1, 8	H×W×8	Entropy map U_E
Conv2	Conv2D, 3×3, S1, 8	H×W×8	Conv1
Conv3	Conv2D*, 3×3, S1, 1	H×W×1	Conv2+U_E

the computational complexity of the network, finally a shallow 2D CNN f_s is used to transform U_E to $\hat{U}_{log}$. The detailed layer configuration of this shallow 2D CNN is presented in Table I.

Using the uncertainty-aware loss alone to guide the optimization of the neural network for MVS will make the optimization process more biased to the regions that are easy to find correspondences across different views, such as rich-texture regions. This is because the $\hat{U}_{log}$ in the residual regression term of Eq. (3) can be regarded as an adaptive weight for depth residual loss, which can adaptively reduce the negative effects produced by high uncertainty regions when training and further makes the network more robust. Nevertheless, this will also force the network ignore the regions with high uncertainty regions, so that the characteristics of these regions cannot be fully learned by the network. For this reason, the mean absolute difference between the estimated depth and the ground truth is used together with the uncertainty-aware loss to guide the optimization of the neural network, which is given by:

$$\mathcal{L}^s = \mathcal{L}_1^s + \mathcal{L}_U^s = \frac{1}{|\Omega|}\sum_{x\in\Omega} |D_{gt}^s(x) - \hat{D}^s(x)| + \mathcal{L}_U^s. \quad (4)$$

At this point, the total loss function for the entire network is formulated as:

$$\mathcal{L} = \sum_{s=1}^{S} \lambda^s \mathcal{L}^s, \quad (5)$$

where λ^s is a weight coefficient for stage s.

B. Uncertainty-Based Depth Hypothesis Sampling

Appropriate depth search range and depth hypothesis interval can greatly benefit the depth estimation. The good depth search range not only can excluded invalid range which helps to improve the efficiency for correspondence search, but also can discard the depth range that may cause matching ambiguity. A well defined depth hypothesis interval is able to boost the representation ability of the features. In the coarse-to-fine architecture [4], [13], [14], the depth search range is progressively narrowed according to the estimated depth at previous stage and the predefined methods such as using a fixed factor to control the depth search range [13], determining depth search range according to the depth probability volume [4], or finding the optimal mean pixel distance on the epipolar line [14]. However, these methods either cannot adaptively find an optimal search range for each pixel or may cause a too wide/narrow search range. To solve these problems, we proposed a novel depth search range refinement method by taking advantage of the estimated uncertainty.

The uncertainty map accurately describes the quality of the predicted depth map. The region with high uncertainty in the estimated uncertainty map is highly consistent with the region with large depth errors in the estimated depth map, and vice versa. Intuitively, the region with accurate depth should have a relative narrow depth search range in the next stage, while region with erroneous depth should have wider depth search range to rectify it at next stage. The uncertainty map can be regarded as a clue to indicate whether the estimated depth is accurate or not. Based on this insight, we use the uncertainty map to adaptively determine the per pixel depth search range. To be specific, the depth hypothesis planes are uniformly sampled from the entire depth range in the first stage. In the finer stage, the basic depth interval Δd is pre-defined as [13] does, namely,

$$\Delta d = R^1/(H^1 * 4), \quad (6)$$

where R^1 denotes the depth search range at the 1-st stage which is predetermined, H^1 denotes the number of depth hypothesis plane at the 1-st stage. We further determine the depth interval for finer stage by

$$\Delta d^s = \text{Sigmoid}(U^{s-1}) * \Delta d, \quad (7)$$

where Δd^s denotes the depth interval at the s-th stage ($s \in \{2, 3\}$), so that the depth search range R^s for the s-th stage is $\Delta d^s * H^s$. Sigmoid$(\cdot)$ denotes the Sigmoid function. Since the scale of uncertainty for each pixel varies a lot, we use Sigmoid to normalize it. At this point, we are able to leverage the estimated uncertainty to assign rational depth interval for each pixel.

C. Uncertainty Guided Multi-View Stereo Network

Fig. 2 shows the framework of UGNet, which adapts the coarse-to-fine depth estimation strategy. Due to that as the number of stages increases, the performance of the network first remarkably increases and then stabilizes, which has been verified by [13]. Considering the efficiency and performance of the model, we set the number of pyramid layers to 3. As illustrated, we first extract multi-scale image features $\{F_i^s\}_{i=0}^{N-1}$ ($s = 1, 2, 3$) for all images by using the multi-scale feature extraction network [41], [44] with shared weights. Table II gives the detailed layer configuration of the multi-scale feature extraction network, where "Out1", "Out2" and "Out3" are the three outputs of the network, representing the extracted features $\{F_i^1\}_{i=0}^{N-1}$, $\{F_i^2\}_{i=0}^{N-1}$ and $\{F_i^3\}_{i=0}^{N-1}$, respectively. It can be seen that the multi-scale feature extraction network is a 2D UNet, which is mainly composed of an encoder and a decoder with skip connections. To enhance the capability of feature representation, the encoder and decoder are composed of multiple residual blocks.

Then, for each stage, all feature maps of source views are warped into a set of fronto-parallel planes of reference view at the sampled depth d to form the $N-1$ feature

TABLE II
THE DETAILED LAYER CONFIGURATION OF THE MULTI-SCALE FEATURE EXTRACTION NETWORK. IF NOT SPECIFIED, EACH 2D CONVOLUTION/DECONVOLUTION LAYER IS FOLLOWED BY A BATCH NORMALIZATION (BN) AND A RECTIFIED LINEAR UNIT (RELU). S1/2 DENOTES THE CONVOLUTION STRIDE. "H × W" DENOTES THE SIZE OF THE INPUT IMAGE X_i. "⋆" DENOTES NO RELU, "∗" DENOTES NO BN AND RELU, "+" DENOTES ADDITION, AND "[]" DENOTES CONCATENATION

Name	Layer Description	Output Size	Input
Conv1	Conv2D, 5×5, S2, 16	½H×½W×16	Image X_i
Conv2_0	Conv2D, 3×3, S1, 32	½H×½W×32	Conv1
Conv2_1	Conv2D⋆, 3×3, S1, 32	½H×½W×32	Conv2_0
Conv2_2	Conv2D⋆, 1×1, S1, 32	½H×½W×32	Conv1
Relu2_1	ReLU	½H×½W×32	Conv2_1+Conv2_2
Conv2_3	Conv2D, 3×3, S1, 32	½H×½W×32	Relu2_1
Conv2_4	Conv2D⋆, 3×3, S1, 32	½H×½W×32	Conv2_3
Relu2_2	ReLU	½H×½W×32	Relu2_1+Conv2_4
Conv3_0	Conv2D, 3×3, S2, 64	¼H×¼W×64	Relu2_2
Conv3_1	Conv2D⋆, 3×3, S1, 64	¼H×¼W×64	Conv3_0
Conv3_2	Conv2D⋆, 1×1, S2, 64	¼H×¼W×64	Relu2_2
Relu3_1	ReLU	¼H×¼W×64	Conv3_1+Conv3_2
Conv3_3	Conv2D, 3×3, S1, 64	¼H×¼W×64	Relu3_1
Conv3_4	Conv2D⋆, 3×3, S1, 64	¼H×¼W×64	Conv3_3
Relu3_2	ReLU	¼H×¼W×64	Relu3_1+Conv3_4
Conv4_0	Conv2D, 3×3, S2, 128	⅛H×⅛W×128	Relu3_2
Conv4_1	Conv2D⋆, 3×3, S1, 128	⅛H×⅛W×128	Conv4_0
Conv4_2	Conv2D⋆, 1×1, S2, 128	⅛H×⅛W×128	Relu3_2
Relu4_1	ReLU	⅛H×⅛W×128	Conv4_1+Conv4_2
Conv4_3	Conv2D, 3×3, S1, 128	⅛H×⅛W×128	Relu4_1
Conv4_4	Conv2D⋆, 3×3, S1, 128	⅛H×⅛W×128	Conv4_3
Relu4_2	ReLU	⅛H×⅛W×128	Relu4_1+Conv4_4
Conv5_0	DeConv2D∗, 3×3, S2, 64	¼H×¼W×64	Relu4_2
Conv5_1	Conv2D⋆, 3×3, S1, 64	¼H×¼W×64	[Relu3_2, Conv5_0]
Conv5_2	Conv2D, 3×3, S1, 64	¼H×¼W×64	Conv5_1
Conv5_3	Conv2D⋆, 3×3, S1, 64	¼H×¼W×64	Conv5_2
Relu5	ReLU	¼H×¼W×64	Conv5_1+Conv5_3
Conv6_0	DeConv2D∗, 3×3, S2, 32	½H×½W×32	Relu5
Conv6_1	Conv2D⋆, 3×3, S1, 32	½H×½W×32	[Relu2_2, Conv6_0]
Conv6_2	Conv2D, 3×3, S1, 32	½H×½W×32	Conv6_1
Conv6_3	Conv2D⋆, 3×3, S1, 32	½H×½W×32	Conv6_2
Relu6	ReLU	½H×½W×32	Conv6_1+Conv6_3
Out1	Conv2D⋆, 3×3, S1, 32	⅛H×⅛W×32	Relu4_2
Out2	Conv2D⋆, 3×3, S1, 32	¼H×¼W×32	Relu5
Out3	Conv2D⋆, 3×3, S1, 32	½H×½W×32	Relu6

volumes $\{C_i^s\}_{i=1}^{N-1}$ upon the frustum of the reference view by differentiable homography:

$$H_i^s(d) = K_i^s T_i T_0^{-1} K_0^{s-1}, \tag{8}$$

where K_i^s, T_i and K_0^s, T_0 denote the intrinsic and extrinsic camera parameters of the $i\text{-}th$ source view image and the reference image at stage s, respectively.

Next, all feature volumes $\{C_i^s\}_{i=1}^{N-1}$ are aggregated together with the reference feature F_0^s to build a 3D cost volume C^s by computing the group-wise correlation similarity [30], [41], [45]. Note that the channel number of cost volume for all stages are set to 8 in our method. Afterwards, the cost volume regularization network [41] is applied to regularize the raw cost volume to obtain pixel-wise depth probability distribution and pixel-wise uncertainty probability distribution. The detailed layer configuration of cost volume regularization network is presented in Table III, which is composed of two cascaded 3D U-Nets. Similar to multi-scale feature extraction network, the encoder and decoder parts of the 3D U-Net also mainly consist of multiple residual blocks to enhance the capability of feature representation. It is worth noting that the structure of the cost volume regularization networks is the same for each stage. Finally, the depth map D^s is regressed from the expectation of the depth probability distribution, and the uncertainty map U^s is regressed from the entropy of uncertainty probability distribution with a shallow 2D CNN.

During training, the uncertainty-aware loss function is used to guide the optimization of the network where the uncertainty is learned implicitly without supervision, and the depth search range for each scale is getting more and more narrow by using the proposed uncertainty-based depth hypothesis sampling strategy, the largest-scale depth map D^3 whose resolution is the half of the input image will be the final output.

IV. EXPERIMENTS

In this section, the datasets used in the experiments are introduced first; We then describe the implementation details of our method; Next, an ablation study to quantify the improvements by integrating the proposed components is demonstrated; Finally, benchmarking on different datasets followed by comparison of memory consumption and run-time are presented.

A. Dataset

There are four datasets used in the experiments: DTU dataset [16], BlendedMVS [15] dataset, Tanks and Temples dataset [17], and ETH3D high-res benchmark [18]. DTU [16] is a large-scale indoor dataset, which includes more than 100 scenes captured in a laboratory environment. We divide DTU dataset into the training set, validation set and evaluation set in the experiments as Ji *et al.* [21] does, and the data are preprocessed as [3] does. The BlendedMVS [15] contains 113 scenes, covering a variety of different scenes, including buildings, street scenes, sculptures and small objects. It is split into the training set and the validation set, no evaluation set provided. Tanks and Temples [17] contains outdoor and indoor scenes captured in realistic environments, which is divided into the training set and the test set. For the test set, it is further divided into *intermediate* subset and *advanced* subset. ETH3D [18] is also composed of both realistic outdoor and indoor scenes, but it is more challenge compared with Tanks and Temples since viewpoints in ETH3D vary a lot while data in Tanks and Temples are presented as video sequences. For all datasets, we use the method in [3] to calculate view selection scores between each reference view and each of the other views, and use the N best views for training and testing.

B. Implementation Details

1) Training: We train our model on the training set of DTU [16] for ablation study on evaluation set of DTU and DTU Benchmarking, and the model trained on the training set of BlendedMVS [15] is mainly used for ablation study on training set of ETH3D, Tanks and Temples and ETH3D Benchmarking. During training, the resolution of all input images are set to 640×512, the resolution of the output depth map is half of the input image, the number of views N is 6 for training. Our coarse-to-fine architecture is composed of

TABLE III

THE DETAILED LAYER CONFIGURATIONS OF THE COST VOLUME REGULARIZATION NETWORK. IF NOT SPECIFIED, EACH 3D CONVOLUTION/DECONVOLUTION LAYER IS FOLLOWED BY A BATCH NORMALIZATION (BN) AND A RECTIFIED LINEAR UNIT (RELU). S1/2 DENOTES THE CONVOLUTION STRIDE. "D × H × W" DENOTES THE SIZE OF THE INPUT COST VOLUME C^s. "$\star$" DENOTES NO RELU, "$*$" DENOTES NO BN AND RELU, "+" DENOTES ADDITION, AND "[]" DENOTES CONCATENATION

Name	Layer Description	Output Size	Input
Conv1_0	Conv3D, 3×3×3, S1, 8	D×H×W×8	Cost Volume C^s
Conv1_1	Conv3D*, 3×3×3, S1, 8	D×H×W×8	Conv1_0
Relu1	ReLU	D×H×W×8	C^s+Conv1_1
Conv2_0	Conv3D, 3×3×3, S2, 16	½D×½H×½W×16	Relu1
Conv2_1	Conv3D*, 3×3×3, S1, 16	½D×½H×½W×16	Conv2_0
Conv2_2	Conv3D*, 1×1×1, S2, 16	½D×½H×½W×16	Relu1
Relu2	ReLU	½D×½H×½W×16	Conv2_1+Conv2_2
Conv3_0	DeConv3D*, 3×3×3, S2, 8	D×H×W×8	Relu2
Conv3_1	Conv3D*, 3×3×3, S1, 8	D×H×W×8	[Relu1, Conv3_0]
Conv4	Repeat Conv1_0-Conv3_1	D×H×W×8	Conv3_1
Conv5	Conv3D*, 3×3×3, S1, 8	D×H×W×2	Conv4
Depth probability volume	Softmax	D×H×W×1	1*st* channel of Conv5
Uncertainty probability volume	Softmax	D×H×W×1	2*nd* channel of Conv5

3 stages, the number of depth hypothesis planes H^s for each stage is set to 32, 16, 8.

Our network is implemented by PyTorch [46]. Adam [47] is used as the optimizer to train the network for 10 epochs. We set the initial learning rate to 0.001 and decrease the learning rate by half at 6-*th*, 8-*th*, 9-*th* epoch to avoid falling into local optima. The λ^s of each stage is 0.5, 1.0 and 2.0. The batch size is set to 2 on one GeForce RTX 2080Ti GPU device.

2) Testing: The proposed method is tested on validation set and evaluation set of DTU dataset [16], the validation set of BlendedMVS [15], Tanks and Temples dataset [17] and ETH3D high-res benchmark [18]. The resolution of input image is 1600×1184 for DTU, 768×567 for BlendedMVS, 1920×1056 for Tanks and Temples, 1920×1280 for ETH3D. The number of view N is 7 for DTU and BlendedMVS, 5 for ETH3D, and 10 for Tanks and Temples. We first use the proposed model to predict depth maps for all input views, then use the probability map generated as [3] does to filter depth maps, and dynamic consistency checking algorithm [31] is used to reconstruct the 3D point cloud from depth maps. Due to the resolution of images in BlendedMVS dataset is relative small, the number of depth hypothesis planes H^s for each stage are set to 32, 16, 8 on BlendedMVS. The number of depth hypothesis planes H^s for each stage on the DTU and Tanks and Temples databases are 64, 32, 16, and the number of depth hypothesis planes H^s for each stage on the ETH3D database are 96, 48, 24.

3) Evaluation Metrics: The *Accuracy* (Acc.) and *Completeness* (Comp.) of the distance metric are used to measure the quality of reconstructed point clouds for DTU dataset, while the accuracy and completeness of the percentage metric are adopted for Tanks and Temples dataset and ETH3D high-res benchmark. We calculate the average of the mean accuracy and the mean completeness as the *overall* score for DTU dataset and F_1 score for other two datasets.

In order to quantify the quality of the estimated uncertainty map, we use the Receiver Operating Characteristic (ROC) curve analysis as suggested by previous methods [35]–[37]. The ROC curve plots the error rate of uncertainty map as a function of pixels sampled from the depth map in order of increasing uncertainty. Based on the ROC curve, the Area Under the Curve (AUC) can be computed to measure the quality of the uncertainty map. In this case, to plot the ROC curve, the criterion that is used to judge whether the estimated depth is correct or not needs to be defined first. To this end, we use the 1 depth-wise pixel and 3 depth-wise pixel as thresholds like previous methods [33], [41], namely, the criterion is computed as $\frac{|D_{gt}(x)-\hat{D}(x)|}{\Delta d} < \tau$ where $\tau = 1$ *or* 3. And the error rate ϵ of depth map can be formulated as

$$\epsilon = \frac{\text{Count}_{x\in\Omega}(\frac{|D_{gt}(x)-\hat{D}(x)|}{\Delta d} > \tau)}{\text{Count}(\Omega)} \times 100\%, \tag{9}$$

where $\text{Count}(\Omega)$ denotes the number of valid pixels in D_{gt}, $\text{Count}_{x\in\Omega}$ denotes the number of pixels satisfies $\frac{|D_{gt}(x)-\hat{D}(x)|}{\Delta d} > \tau$, and Δd denotes the basic depth interval computed by Eq. (6). In this context, the optimal AUC is computed as

$$AUC_{opt} = \int_{1-\epsilon}^{1} \frac{p-(1-\epsilon)}{p}\, dp \ = \epsilon + (1-\epsilon)\ln(1-\epsilon), \tag{10}$$

where p is the percentage of pixels sampled from the depth map. Note that the closer the AUC is to the optimal AUC, the more accurate the estimated uncertainty is.

C. Ablation Study

The ablation study is performed to validate the effectiveness of each component in the proposed method. In addition, the analysis of number of views used for each dataset is also conducted in this section. Experiments on DTU are conducted on the model trained with the DTU training set and tested on the DTU evaluation set which contains 22 different scenes. Experiments on ETH3D and Tanks and Temples are conducted by the model trained with the BlendedMVS training set and tested on the training sets of ETH3D and Tanks and Temples. The experimental setting is the same as described in Section IV-B. The baseline model is obtained without uncertainty-aware loss function (ULF) and uncertainty-based depth hypothesis sampling (UDHS), namely, the model is trained by only using $\mathcal{L}_1$ in Eq. (4), and the depth hypothesis

TABLE IV

ABLATION EXPERIMENTS ON DTU EVALUATION SET BY USING THE DISTANCE METRIC [*mm*] (LOWER IS BETTER) AND ETH3D TRAINING SET BY USING F_1 SCORE (%) AT THRESHOLD $2cm$ (HIGHER IS BETTER), WHICH VALIDATE THE IMPROVEMENT OF DIFFERENT COMPONENTS OF OUR APPROACH

Model	ULF	UDHS	DTU			ETH3D
			Acc. ↓	Comp. ↓	Overall ↓	F_1 ↑
Baseline			0.357	0.372	0.365	66.35
Model-A	✓		0.369	0.347	0.358	67.47
UGNet	✓	✓	0.334	0.330	0.332	72.78

TABLE V

QUANTITATIVE RESULTS ON DTU VALIDATION SET AND BLENDEDMVS VALIDATION SET IN TERMS OF DEPTH AND UNCERTAINTY MAPS USING THE ERROR RATE ϵ (%) AND AUC, RESPECTIVELY

Dataset	$\tau = 1$			$\tau = 3$		
	ϵ	AUC	AUC_{opt}	ϵ	AUC	AUC_{opt}
DTU	19.69	0.062	0.035	11.70	0.035	0.014
BlendedMVS	8.06	0.009	0.005	4.12	0.003	0.001

TABLE VI

COMPARISON RESULTS OF DIFFERENT DEPTH HYPOTHESIS SAMPLING STRATEGIES ON DTU EVALUATION SET. MODEL-B AND MODEL-C DENOTE OUR BACKBONE WITH HYPOTHESIS SAMPLING STRATEGY IN CVP-MVSNET [14] AND UCSNET [4], RESPECTIVELY. (LOWER IS BETTER)

Model	Acc. ↓	Comp. ↓	Overall ↓
Model-B	0.415	0.375	0.395
Model-C	0.303	0.394	0.348
UGNet (Ours)	0.334	0.330	0.332

sampling strategy in CasMVSNet [13] is used here where the coefficients for depth interval for each stage are set as 4, 2 and 1.

1) Uncertainty-Aware Loss Function: In order to validate the effectiveness of uncertainty-aware loss function, we present the experimental results without and with the ULF based on baseline model in the first and second rows in Table IV, respectively. It can be seen that with the uncertainty-aware loss, the performance of reconstructed point clouds are greatly improved. This is because that using the uncertainty-aware loss to guide the training of network, it can force the model to learn to reduce the effect from the high-uncertainty areas in the data as well as erroneous areas in the label. In this way, the network can be more robust to noise data, and further contributes to estimate more accurate depth map. Fig. 3 illustrates the qualitative comparisons of depth maps estimated by Baseline and Model-A, it can be seen that depth maps estimated by Baseline are noisy, especially in low-texture areas. Compared with the Baseline, the depth map estimated by Model-A is smoother, which is able to deal with the depth estimation on low-texture areas better. With the help of the uncertainty introduced by Eq. (3), it makes the model more robust to noisy data, such as low-texture areas. The depth residual regression part is annealed by the learned uncertainly $\hat{U}$ during training as shown in Eq. (3), it allows the network to adapt the residual's weighting, so that it can learn to attenuate the adverse impact of noisy data.

We further quantify the quality of estimated depth map and uncertainty map using the error rate and AUC, respectively. Quantitative results on DTU validation set and BlendedMVS validation set are given in Table V. Note that results in Table V are achieved by averaging the error rate ϵ, AUC and AUC_{opt} on the entire validation sets of DTU and BlendedMVS, the results on DTU and BlendedMVS are obtained using the UGNet trained on their own training sets. As shown, the error rates of estimated depth maps with $\tau = 1$ and 3 are 19.69% and 11.70% on DTU dataset, 8.06% and 4.12% on BlendedMVS dataset, which indicates that the proposed UGNet can predict depth accurately. Moreover, the AUC is very close to AUC_{opt} on both DTU and BlendedMVS datasets, which indicates that the estimated uncertainty is also accurate. In addition, Fig. 1 illustrates the consistency between depth error maps and uncertainty maps, which also reflects the accuracy of estimated uncertainty maps.

2) Uncertainty-Based Depth Hypothesis Sampling: Experiments are further performed to evaluate the effectiveness of proposed uncertainty-based depth hypothesis sampling strategy, we compare the experimental results of our uncertainty-based depth hypothesis sampling strategy with that of the Model-A in Table IV which uses the depth hypothesis sampling strategy of CasMVSNet [13]. As shown in the second and third rows in Table IV, our method achieves better results campared to Model-A in terms of the reconstruction *accuracy*, *completeness* and *overall* quality. It is worth noting that the proposed uncertainty-based depth hypothesis sampling strategy greatly improves the performance on ETH3D dataset, which enables our method to handle wide baseline data. This further shows the importance of reasonable depth hypothesis sampling. In addition, as shown in Fig. 3, the introduction of the proposed uncertainty-based depth hypothesis sampling can further improve the quality of estimated depth maps. This is because the estimated uncertainty can accurately reflect the accuracy of the estimated depth map, so that the depth hypothesis can be adaptively and accurately determined according to the currently estimated depth map.

Furthermore, we compare our proposed uncertainty-based depth hypothesis sampling strategy with other depth hypothesis sampling strategies used in CVP-MVSNet [14] and UCSNet [4] by combined them with our backbone, namely, except that the depth hypothesis sampling strategy is different, other is the same as the setting of the proposed UGNet. Comparison results of different depth hypothesis sampling strategies on DTU evaluation set are presented in Table VI, where Model-B and Model-C denote our backbone with hypothesis sampling strategy in CVP-MVSNet [14] and UCSNet [4], respectively. It is worth noting that Table IV already gives the results of our backbone with hypothesis sampling strategy in CasMVSNet [13]. We can observe from Table IV and Table VI that our proposed uncertainty-based depth hypothesis sampling strategy outperforms that of other methods.

3) Analysis of Number of Views: To decide the number of views used for training and testing, a series of experiments are conducted in this section. First of all, experiments are conducted on the DTU dataset to decide the number of

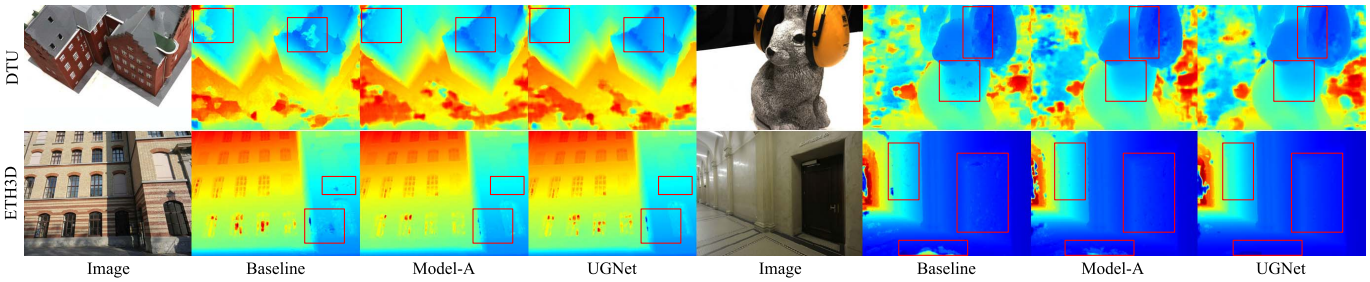


Fig. 3. Qualitative comparisons of depth maps estimated by Baseline, Model-A and UGNet on DTU and ETH3D datasets, respectively.

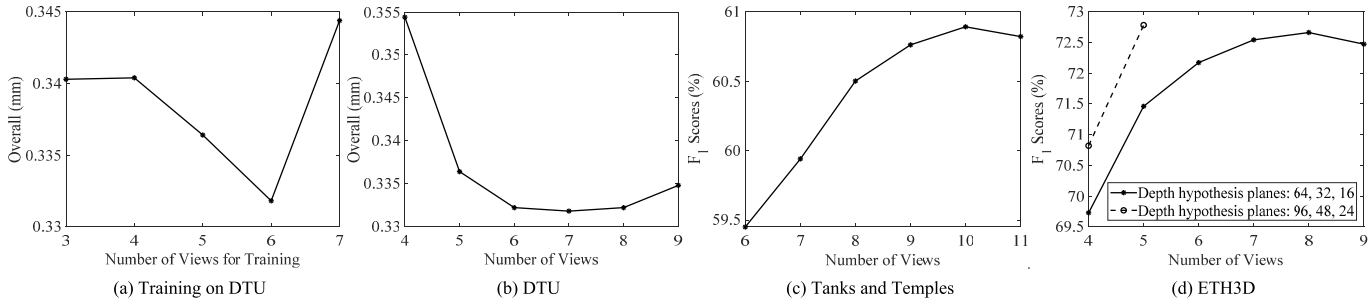


Fig. 4. Experiments results for the analysis of number of views used for training and testing on various datasets. (a) illustrates the experiments results with the change of number of views used for training on DTU dataset. (b)-(d) illustrate the experiments results with the change of number of views used for testing on DTU, Tanks and Temples and ETH3D datasets, respectively. Note that lower is better for the metric of Overall, while higher is better for the metric of F_1 scores.

views used for training. Fig. 4 (a) gives the results of these experiments, note that when testing the number of views is set to 7. As shown, when the number of views increases, the performance of the network first improves and then decreases sharply. When the number of views is 6, the performance of the model reaches the best. This means that when the number of input views increases from 3 to 6, it is conducive for the network to learn more useful information from more images and better deal with the invisibility between the reference view and source views. However, when the number of views exceeds 6, more images far from the perspective of the reference view is introduced, which means more noises is introduced. This exceeds the ability of the network to deal with noises, which will seriously degrade the performance of the network.

Fig. 4 (b)–Fig. 4 (d) illustrate the experiments results with the change of the number of views for testing on DTU, Tanks and Temples and ETH3D datasets. It is worth noting that the width of the baseline between views and the resolution of the images used in the ETH3D dataset are wider and larger than those in other datasets, so the performance is better when testing with more depth hypotheses. However, due to the limitation of memory, if the depth hypothesis planes H^s for each stage are 96, 48, 24, the maximum number of views allowed to be tested is 5. If the memory allows more views to be tested, better performance can be obtained on the ETH3D dataset. It can be seen from Fig. 4 (b)–Fig. 4 (d) that the optimal number of views used for testing varies according to the specific dataset. On the Tanks and Temples dataset, the optimal number of views is greater than the other two datasets. This is because the Tanks and Temples dataset is presented as video sequences, and the change of perspective between views is smaller than other datasets. Therefore, as more views are introduced, relatively less noise is introduced. Besides, for DTU dataset, it mainly aims to reconstruct the object in the image, while ETH3D dataset aims to reconstruct the whole scene in the image. As a result, more views can introduce more useful information when doing scene reconstruction. On the contrary, more views mean more noises will be introduced for object reconstruction. Consequently, the optimal number of views used for testing on the ETH3D dataset is slightly more than that of the DTU dataset. In this case, if the input data is presented as video sequence like Tanks and Temples dataset, we recommend that the number of views can be set

TABLE VII

QUANTITATIVE RESULTS OF RECONSTRUCTED POINT CLOUDS ON DTU EVALUATION SET BY USING THE DISTANCE METRIC [*mm*] (LOWER IS BETTER). OUR METHOD OUTPERFORMS ALL METHODS IN TERMS OF OVERALL QUALITY

	Method	Acc. ↓	Comp.↓	Overall ↓
Geometric	Camp [48]	0.835	0.554	0.695
	Furu [20]	0.613	0.941	0.777
	Tola [49]	0.342	1.190	0.766
	Gipuma [25]	**0.283**	0.873	0.578
	COLMAP [27]	0.411	0.657	0.534
Learning	MVSNet [3]	0.396	0.527	0.462
	CIDER [30]	0.417	0.437	0.427
	R-MVSNet [29]	0.383	0.452	0.417
	D^2HC-RMVSNet [31]	0.395	0.378	0.386
	PointMVSNet [50]	0.342	0.411	0.376
	Vis-MVSNet [41]	0.369	0.361	0.365
	AA-RMVSNet [32]	0.376	0.339	0.357
	AttMVS [51]	0.383	0.329	0.356
	CasMVSNet [13]	0.325	0.385	0.355
	EPP-MVSNet [34]	0.413	0.296	0.355
	PatchmatchNet [52]	0.427	**0.277**	0.352
	CVP-MVSNet [14]	0.296	0.406	0.351
	UCSNet [4]	0.338	0.349	0.344
	AACVP-MVSNet [33]	0.357	0.326	0.341
	UGNet (Ours)	0.334	0.330	**0.332**

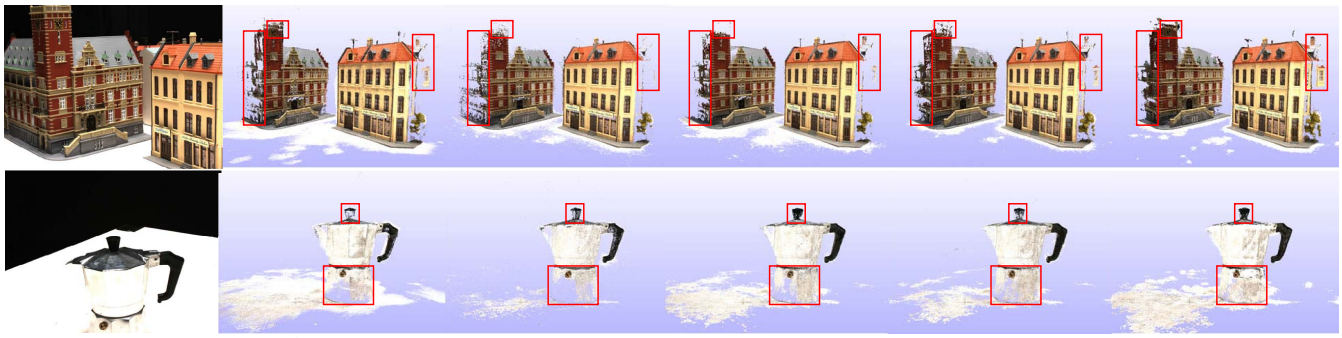


Fig. 5. Qualitative comparisons of reconstructed point clouds with D^{2}HC-RMVSNet [31], CVP-MVSNet [14], CasMVSNet [13] and UCSNet [4] in terms of Scan29 and Scan77 in DTU dataset.

TABLE VIII
QUANTITATIVE RESULTS ON TANKS AND TEMPLES DATASET WITH F_1 SCORES (%). (HIGHER IS BETTER)

	Method	Intermediate									Advanced						
		Mean	Fam.	Franc.	Horse	Light.	M60	Pan.	Play.	Train	Mean	Audi.	Ballr.	Courtr.	Museum	Palace	Temple
Geometric	COLMAP [27]	42.14	50.41	22.25	25.63	56.43	44.83	46.97	48.53	42.04	27.24	16.02	25.23	34.70	41.51	18.05	27.94
	ACMH [26]	54.82	69.99	49.45	45.12	59.04	52.64	52.37	58.34	51.61	33.73	21.69	32.56	40.62	47.27	24.04	36.17
	OpenMVS [53]	55.11	71.69	51.12	42.76	58.98	54.72	56.17	59.77	45.69	34.43	**24.49**	37.39	38.21	47.48	27.25	31.79
	ACMM [26]	57.27	69.24	51.45	46.97	63.20	55.07	57.64	60.08	54.48	34.02	23.41	32.91	**41.17**	48.13	23.87	34.60
Learning	MVSNet [3]	43.48	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69	-	-	-	-	-	-	-
	R-MVSNet [29]	48.40	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38	24.91	12.55	29.09	25.06	38.68	19.14	24.96
	PatchMatch-RL [54]	51.81	60.37	43.26	36.43	56.27	57.30	53.43	59.85	47.61	31.78	24.28	40.25	35.87	44.13	22.43	23.73
	PatchmatchNet [52]	53.15	66.99	52.64	43.24	54.87	52.87	49.54	54.21	50.81	32.31	23.69	37.73	30.04	41.80	28.31	32.29
	CVP-MVSNet [14]	54.03	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54	-	-	-	-	-	-	-
	UCSNet [4]	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	-	-	-	-	-	-	-
	CasMVSNet [13]	56.84	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	31.12	19.81	38.46	29.10	43.87	27.36	28.11
	AACVP-MVSNet [33]	58.39	78.71	57.85	50.34	52.76	59.73	54.81	57.98	54.94	-	-	-	-	-	-	-
	D^{2}HC-RMVSNet [31]	59.20	74.69	56.04	49.42	60.08	59.81	59.61	60.04	53.92	-	-	-	-	-	-	-
	Vis-MVSNet [41]	60.03	77.40	60.23	47.07	63.44	62.21	57.28	60.54	52.07	33.78	20.79	38.77	32.45	44.20	28.73	37.70
	AttMVS [51]	60.05	73.90	62.58	44.08	64.88	56.08	59.39	**63.42**	56.06	31.93	15.96	27.71	37.99	**52.01**	29.07	28.84
	AA-RMVSNet [32]	61.51	77.77	59.53	51.53	64.02	64.05	59.47	60.85	54.90	33.53	20.96	40.15	32.05	46.01	29.28	32.71
	EPP-MVSNet [34]	61.68	77.86	60.54	**52.96**	62.33	61.69	60.34	62.44	55.30	35.72	21.28	39.74	35.34	49.21	30.00	**38.75**
	UGNet (Ours)	**63.12**	**79.61**	**63.35**	50.32	**66.36**	**64.80**	**60.84**	62.25	**57.41**	**37.12**	23.28	**43.49**	36.04	50.59	**31.81**	37.54

to about 10, while if the perspective of data changes relatively large such as DTU and ETH3D datasets, the number of views can be set to about 7.

D. Benchmark Performance

We compare the proposed UGNet with other state-of-the-art methods on the DTU dataset, Tanks and Temples dataset, and ETH3D high-res benchmark without any fine-tuning.

1) Results on DTU: Quantitative results of reconstructed point clouds on DTU evaluation set are demonstrated in Table VII. As shown, our method achieves the best performance in terms of *overall score* among all traditional and learning-based methods listed in Table VII. We also compare our method with state-of-the-methods D^{2}HC-RMVSNet [31], CVP-MVSNet [14], CasMVSNet [13] and UCSNet [4] in terms of the reconstructed point clouds on Scan29 and Scan77, qualitative comparisons are visualized in Fig. 5. As illustrated, thanks to the introduced uncertainty, compared with other methods, our method is able to reconstruct a more complete point cloud in the low-texture region of Scan 77 and a more delicate point cloud on Scan29.

2) Results on Tanks and Temples: We test our model on Tanks and Temples dataset without any fine-tuning, quantitative results are shown in Table VIII. Obviously, the proposed method shows very promising results on both *Intermediate* set and *Advanced* set, which validates the generalization ability of our method. Specially, our UGNet is superior to traditional methods and methods also using coarse-to-fine architecture (i.e., CVP-MVSNet [14], UCSNet [4], CasMVSNet [13], AACVP-MVSNet [33], Vis-MVSNet [41] and EPP-MVSNet [34]) on both *Intermediate* set and *Advanced* set. Moreover, the proposed method outperforms recent AA-RMVSNet [32] on both *Intermediate* set and *Advanced* set which uses a hybrid RNN-CNN to regularize the cost volume. Some reconstructed point clouds on Tanks and Temples dataset are illustrated in Fig. 6, it shows that our reconstructed point clouds are complete and rich in details.

3) Results on ETH3D: We further evaluate our method on more challenging ETH3D high-res benchmark without any fine-tuning, comparisons of point clouds on ETH3D high-res benchmark are given in Table IX. It can be observed that our method also shows competitive results with traditional methods (MVE [55], Gipuma [25], PMVS [20] and Colmap [27]) and recent state-of-the-art learning-based methods (PVSNet [56], PatchMatch-RL [54] and Patchmatch-Net [52]), which validate that our method is also able to handle with data that have wide baselines. Fig. 7 illustrates some reconstructed point clouds on ETH3D high-res benchmark, which further shows that our method can achieve dense and accurate reconstruction.

Fig. 6. Visualization of reconstructed point clouds of tanks and temples dataset.

Fig. 7. Visualization of reconstructed point clouds of ETH3D high-res benchmark.

TABLE IX

COMPARISONS OF RECONSTRUCTED POINT CLOUDS ON ETH3D HIGH-RES BENCHMARK USING F_1 SCORE (%) AT THRESHOLD $2cm$. (HIGHER IS BETTER)

	Method	Training set	Test set
Geometric	MVE [55]	20.47	30.37
	Gipuma [25]	36.38	45.18
	PMVS [20]	46.06	44.16
	Colmap [27]	67.66	73.01
Learning	PVSNet [56]	67.48	72.08
	PatchMatch-RL [54]	67.78	72.38
	PatchmatchNet [52]	64.21	73.12
	UGNet (Ours)	**72.78**	**80.83**

E. Memory and Run-Time Comparison

We demonstrate the efficiency of the proposed method compared with state-of-the-art approaches in terms of GPU memory usage and run-time. The memory consumption and running time are tested with one GeForce 2080Ti GPU when the resolution of images is 1920×1056 and the number of views is 5. The summary of these results are listed at Table X, where the F_1 scores are obtained on the *intermediate* set of Tanks and Temples dataset. Obviously, compared with learning-based methods CasMVSNet [13] and Vis-MVSNet [41], the memory consumption of our method is the least. Although the run-time of CasMVSNet [13] is less than our method, its memory consumption is about 1.5 times that of ours, and our run-time is still within an acceptable range. In addition, compared with traditional method ACMM [26], although the memory usage of learning-based methods is much larger, in terms of running time, learning-based methods use much less time. Moreover, with the development of deep learning technology in the field of MVS, the performance of learning-based methods has greatly exceeded the traditional methods which are limited by traditional handcrafted features.

TABLE X
COMPARISON RESULTS OF MEMORY CONSUMPTION, RUNNING TIME AND THE RECONSTRUCTION QUALITY ON THE INTERMEDIATE SET OF TANKS AND TEMPLES DATASET. NOTE THAT THE MEMORY CONSUMPTION AND RUNNING TIME ARE TESTED WHEN THE RESOLUTION OF IMAGES IS 1920 × 1056 AND THE NUMBER OF VIEWS IS 5

Method	Memory (MB)	Time (s)	F_1 scores (%)
ACMM [26]	2145	5.3092	57.27
CasMVSNet [13]	10271	0.8540	56.84
Vis-MVSNet [41]	7543	2.1482	60.03
UGNet (Ours)	6695	1.3223	63.12

V. CONCLUSION

We present an Uncertainty Guided multi-view stereo Network (UGNet) for depth estimation in this paper. In practical applications, the perception of network output reliability is very important. However, the existing learning-based methods for MVS lack the reliability measurement of the estimated depth map. To this end, we introduce an uncertainty-aware loss function to guide training process of the network. In this way, our method can estimate the depth map and uncertainty map at the same time rather than treating uncertainty map as a post-processing step as previous methods. With the uncertainty-aware loss function, the bad impact of high uncertainty regions and the erroneous labels in the training set can be attenuated during training, and the accuracy of the estimated depth can be further boosted. Moreover, we further propose an uncertainty-based depth hypothesis sampling strategy to adaptively determine the depth search range for finer stages, which helps to generate more rational depth intervals compared with other methods and build more compact cost volume without redundancy. Experimental results conducted on various benchmark datasets verify the effectiveness of our method.

The uncertainty maps give clues to indicate whether the estimated depths are accurate or not. For future work, we will explore using the uncertainty map to perform adaptive depth refinement on the high uncertainty pixels of the final depth map predicted by the multi-view depth estimation network, so as to obtain a more accurate depth map. In addition, existing learning-based MVS methods [3], [4], [13], [14], [30], [33], [34], [41], [51] and the proposed method heavily rely on the strong regularization ability of 3D CNNs when performing cost volume regularization, which is one of the main reasons for the relatively low efficiency of these methods. This is because the computational cost of 3D CNNs is relatively high, while the computational efficiency of 2D CNNs is much higher than that of 3D CNNs. Therefore, we will explore how to combine 3D CNNs with 2D CNNs to improve the efficiency of depth map estimation with guaranteed performance in the future.

REFERENCES

[1] Y. Lee and H. Kim, "A high-throughput depth estimation processor for accurate semiglobal stereo matching using pipelined inter-pixel aggregation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 411–422, Jan. 2022.

[2] H. Kim, J. Y. Guillemaut, T. Takai, M. Sarim, and A. Hilton, "Outdoor dynamic 3-D scene reconstruction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 11, pp. 1611–1622, Nov. 2012.

[3] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 767–783.

[4] S. Cheng *et al.*, "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2524–2534.

[5] I. Hacking *et al.*, *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability, Induction and Statistical Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[6] H. Wu, D. Zeng, Y. Hu, H. Shi, and T. Mei, "Dual spoof disentanglement generation for face anti-spoofing with depth uncertainty learning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Dec. 7, 2021, doi: 10.1109/TCSVT.2021.3133620.

[7] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5574–5584.

[8] Z. Li, W. Zuo, Z. Wang, and L. Zhang, "Confidence-based large-scale dense multi-view stereo," *IEEE Trans. Image Process.*, vol. 29, pp. 7176–7191, 2020.

[9] A. Kuhn, C. Sormann, M. Rossi, O. Erdler, and F. Fraundorfer, "DeepC-MVS: Deep confidence prediction for multi-view stereo reconstruction," in *Proc. Int. Conf. 3D Vis. (3DV)*, Nov. 2020, pp. 404–413.

[10] Y. Wang, T. Guan, Z. Chen, Y. Luo, K. Luo, and L. Ju, "Mesh-guided multi-view stereo with pyramid architecture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2039–2048.

[11] H. Zhang, X. Ye, S. Chen, Z. Wang, H. Li, and W. Ouyang, "The farther the better: Balanced stereo matching via depth-based sampling and adaptive feature refinement," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Dec. 6, 2021, doi: 10.1109/TCSVT.2021.3133353.

[12] H. Dai, X. Zhang, Y. Zhao, H. Sun, and N. Zheng, "Adaptive disparity candidates prediction network for efficient real-time stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3099–3110, May 2022.

[13] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2495–2504.

[14] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4877–4886.

[15] Y. Yao *et al.*, "BlendedMVS: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1790–1799.

[16] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 153–168, 2016.

[17] A. Knapitsch *et al.*, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, Jul. 2017.

[18] T. Schops *et al.*, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2538–2547.

[19] M. Lhuillier and L. Quan, "A quasi-dense approach to surface reconstruction from uncalibrated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 418–433, Mar. 2005.

[20] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2008.

[21] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "SurfaceNet: An end-to-end 3D neural network for multiview stereopsis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2307–2315.

[22] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 365–376.

[23] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," *Int. J. Comput. Vis.*, vol. 35, no. 2, pp. 151–173, 1999.

[24] S. Roth and S. R. Richter, "Matryoshka networks: Predicting 3D geometry via nested shape layers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1936–1944.

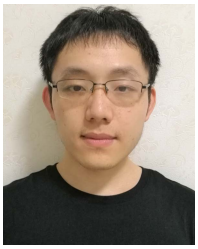
[25] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 873–881.

[26] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5483–5492.

[27] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 501–518.

[28] Q. Xu and W. Tao, "Planar prior assisted patchmatch multi-view stereo," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12516–12523.
[29] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5525–5534.
[30] Q. Xu and W. Tao, "Learning inverse depth regression for multiview stereo with correlation cost volume," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12508–12515.
[31] J. Yan *et al.*, "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 674–689.
[32] Z. Wei, Q. Zhu, C. Min, Y. Chen, and G. Wang, "AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6187–6196.
[33] A. Yu *et al.*, "Attention aware cost volume pyramid based multi-view stereo network for 3d reconstruction," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 448–460, May 2021.
[34] X. Ma, Y. Gong, Q. Wang, J. Huang, L. Chen, and F. Yu, "EPP-MVSNet: Epipolar-assembling based depth prediction for multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5732–5740.
[35] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2121–2133, Jan. 2012.
[36] M. Poggi and S. Mattoccia, "Learning from scratch a confidence measure," in *Proc. Brit. Mach. Vis. Conf.*, 2016, p. 4.
[37] F. Tosi, M. Poggi, A. Benincasa, and S. Mattoccia, "Beyond local reasoning for stereo confidence estimation with deep learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 319–334.
[38] S. Kim, S. Kim, D. Min, and K. Sohn, "LAF-Net: Locally adaptive fusion networks for stereo confidence estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 205–214.
[39] M. Mehltretter and C. Heipke, "Aleatoric uncertainty estimation for dense stereo matching via CNN-based cost volume analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 171, pp. 63–75, Jan. 2021.
[40] J. Zhang *et al.*, "Learning stereo matchability in disparity regression networks," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 1611–1618.
[41] J. Zhang, Y. Yao, S. Li, Z. Luo, and T. Fang, "Visibility-aware multi-view stereo network," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2020, pp. 1–12.
[42] H. Xu *et al.*, "Digging into uncertainty in self-supervised multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6078–6087.
[43] M. Mehltretter, "Uncertainty estimation for end-to-end learned dense stereo matching via probabilistic deep learning," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. V-2-2020, pp. 161–169, Aug. 2020.
[44] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 234–241.
[45] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3273–3282.
[46] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[48] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2008, pp. 766–779.
[49] E. Tola, C. Strecha, and P. Fua, "Efficient large-scale multi-view stereo for ultra high-resolution image sets," *Mach. Vis. Appl.*, vol. 23, no. 5, pp. 903–920, 2012.
[50] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1538–1547.
[51] K. Luo, T. Guan, L. Ju, Y. Wang, Z. Chen, and Y. Luo, "Attention-aware multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1590–1599.
[52] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "PatchmatchNet: Learned multi-view patchmatch stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14194–14203.
[53] OpenMVS. *Open Multi-View Stereo Reconstruction Library*. Accessed: May 20, 2015. [Online]. Available: https://github.com/cdcseacave/openMVS
[54] J. Y. Lee, J. DeGol, C. Zou, and D. Hoiem, "PatchMatch-RL: Deep MVS with pixelwise depth, normal, and visibility," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6158–6167.
[55] S. Fuhrmann, F. Langguth, and M. Goesele, "MVE: A multi-view reconstruction environment," in *Proc. Eurographics Workshop Graph. Cultural Heritage*, Oct. 2014, pp. 11–18.
[56] Q. Xu and W. Tao, "PVSNet: Pixelwise visibility-aware multi-view stereo network," 2020, *arXiv:2007.07714*.

Wanjuan Su received the B.S. and M.S. degrees from the School of Automation, China University of Geosciences, Wuhan, China, in 2017 and 2020, respectively. She is currently pursuing the Ph.D. degree with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan. She was a Visiting Student with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, from 2018 to 2019. Her current research interests include stereo vision, pattern recognition, and image processing.



Qingshan Xu received the B.S. degree in automation from Central South University (CSU), Changsha, China, in 2016, and the Ph.D. degree from the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2021, under the supervision of Prof. W. Tao. From October 2020 to May 2021, he was a Visiting Student with the Computer Vision and Geometry Group (CVG), ETH Zürich, Switzerland, with Prof. M. Pollefeys. His research interests include multi-view stereo, image-based 3D reconstruction, and deep learning with geometry.

Wenbing Tao (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2004. He was a Research Fellow with the Division of Mathematical Sciences, Nanyang Technological University, Singapore, from 2008 to 2009. He is currently a Full Professor with the School of Artificial Intelligence and Automation, HUST. He has authored numerous papers and conference papers in the area of computer vision and 3D reconstruction. His current research interests include computer vision, image segmentation, object recognition, and 3D reconstruction.