Incorporating Co-Visibility Reasoning Into Surface Depth Measurement

Yimei Liu, Yuan Rao, Eric Rigall, Hao Fan, and Junyu Dong, *Member, IEEE*

***Abstract*— Multiview stereo (MVS) aims to measure the precise surface depth of a scene from observations at multiple photography angles and then densely reconstruct its 3-D geometry information. Learning-based MVS approaches have been dominantly popular for their robustness to low texture areas and non-Lambertian surfaces. However, most existing methods focus on estimating depth maps for input images by constructing global cost volumes and designing ingenious yet large variance-based 3-D-CNNs for cost volume regularization. Such approaches ignore the co-visible relationship embedded in multiple views, resulting in heavy computation, erroneous cost aggregation from invisible views, and finally inaccurate 3-D reconstruction results. In this article, we propose a co-visibility reasoning MVS network (CR-MVSNet) to explore the co-visible relationships hidden in multiple views for reliable multiview similarity measurement and efficient reconstruction. Precisely, the proposed co-visibility reasoning cost aggregation (CRCA) module includes the adaptive intercost volume aggregation via mining the uncertainty of co-visibility relationships in multiple views and the adaptive intracost volume aggregation by exploiting spatial contextual information. Moreover, the cost volumes are constructed via the proposed global-to-patch manner to speed up computation. Experimental results show that our approach achieves the best overall performance on the DTU, Tanks and Temples, and ETH3D-test datasets over recent state-of-the-art MVS algorithms. The consistently favorable results on three datasets with completely different depth ranges proved the superiority and generalizability of CR-MVSNet.**

***Index Terms*— 3-D reconstruction, deep learning, multiview stereo (MVS), surface depth measurement, visual measurement.**

I. Introduction

MULTIVIEW stereo (MVS) is an essential research topic for decades and has been used in many computer vision and industrial applications, such as photogrammetry, cartography, and augmented reality. The key technique of MVS is to estimate the depth maps corresponding to a set of images with precalculated camera parameters. Then, the 3-D geometry information of the observed scene can be obtained by fusing the depth maps of multiple views [1].

More specifically, MVS uses a reference image and several source images to infer the depth map of the reference image. The traditional MVS approaches suffer from common limitations when the observed scenes have reflective areas, low texture, or strong viewpoint variations. These situations make dense matching intractable and result in limited reconstruction quality [2], [3], [4], [5], [6], [7]. The learning-based MVS approaches have been dominantly popular for improving the problems above and showed notable performance on challenging MVS benchmarks [8], [9], [10], [11], [12], [13]. Concretely, most learning-based MVS methods [8], [9], [14], [15], [16] construct the initial pairwise cost volumes based on uniformly generated depth candidates in the global depth space, which exhibit source–reference matching quality. Then, several pairwise cost volumes of different source views are fused into a multiview cost volume using the variance-based metric [8], [15], [16], which implies multiview matching similarity clues. Finally, with a designed 3-D-CNN, the fused multiview cost volume can be regularized to a probability volume, which can then be regressed to a depth map. This popular pipeline exposes two critical problems. To begin, setting initially a large number of uniformly generated depth candidates can promote the accuracy of final estimation, but also degrades the algorithm efficiency exponentially. Second, using the variance-based fusing metric, each pairwise cost volume is treated equally. It presupposes that every pixel in the reference view is visible in all the source views, which is not always true in realistic situations. Then a motivating question is whether it is possible to juggle high-quality depth map estimation with algorithm efficiency.

Instead of constructing a global cost volume, several traditional patchmatch (PM) stereo-based methods [17], [18], [19], [20], [21] proposed recently showed great efficiency in depth map estimation tasks. Based on the proposed iterative pipeline: random initialization, propagation, and evaluation [17], the PM-based methods updated depth map iteratively to get the final output. However, most PM-based methods focus on designing ingenious traditional photometric and geometric consistency measure metrics for matching pixel patch but still with limited reconstruction ability in challenging areas [18], [20], [21]. PatchmatchNet, a learning-based PM method [10] which was proposed recently, has outperformed traditional PM-based methods in reconstruction quality. Compared with other global cost volume MVS methods, it exhibited significantly higher efficiency but still suffered from mismatches in challenging regions. It is mainly due to the strategy that solely relying on random initialization and propagation, the sampled depth candidates are not as diverse as generated evenly in

Manuscript received 28 September 2022; revised 10 December 2022; accepted 13 February 2023. Date of publication 24 March 2023; date of current version 28 March 2023. This work was supported in part by the National Key Research and Development Program under Grant 2019YFC1509100 and in part by the National Science Foundation of China under Grant 42106193 and Grant 41927805. The Associate Editor coordinating the review process was Dr. Xiaotong Tu. *(Corresponding author: Junyu Dong.)*

The authors are with the Department of Information Science and Technology, Ocean University of China, Qingdao 266100, China (e-mail: liuyimei@stu.ouc.edu.cn; dongjunyu@ouc.edu.cn).

Digital Object Identifier 10.1109/TIM.2023.3250231

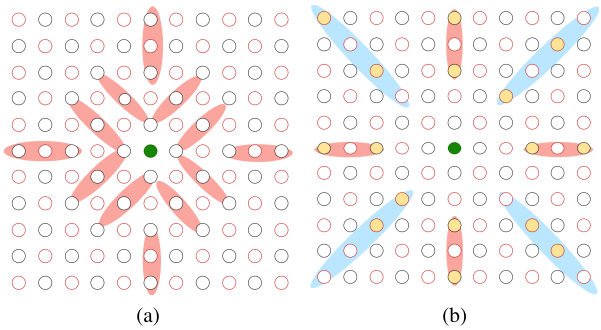


Fig. 1. (a) Traditional ACMM sampling pattern [18] and (b) our proposed sampling pattern. For a given pixel (in green), the sampling regions are in red and blue. In our proposed strategy, the region in red is sampled with fixed positions, whereas the region in blue is sampled based on previous cost measures. [18].

global depth space, thus resulting in a less robust multiview similarity measure.

To cope with the problems above in popular MVS pipeline and PM-based methods, we proposed a complementary framework to juggle final construction quality with method efficiency. Precisely, we propose CR-MVSNet, a co-visibility reasoning learning-based multiview stereo (MVS) network integrated with traditional PM idea. To begin, the initial pairwise cost volumes are constructed using iteratively generated fewer depth candidates, progressing from global depth space to predicted narrow depth range, and eventually patch-based propagation and local perturbation. Instead of a huge number of depth candidates distributed evenly across the global depth space, these global-to-patch generated depth candidates allow the network to focus on a small depth range rather than searching the entire depth space, and propagated depth candidates favored by structured region information are more favorable for accurate depth estimation [17], [22]. The computation time and GPU memory consumption decrease dramatically as the number of depth candidates decreases. Then, instead of completely relying on a 3-D-CNN for cost volume aggregation, a more efficient co-visibility reasoning cost aggregation (CRCA) module is proposed, including: 1) intercost volume aggregation by mining the uncertainty of co-visibility relationships embedded in multiple views and 2) intracost volume aggregation considers spatial contextual information for neighboring pixels. Moreover, a range prediction module and an adaptive propagation strategy [Fig. 1(b)] are put forward to provide more diverse depth candidates for cost volume construction.

Our main contributions are summarized as follows.

1) We propose a CRCA module which includes the adaptive intercost volume aggregation and the adaptive intracost volume aggregation. The proposed CRCA module considers both co-visibility relationships embedded in multiple views and spatial contextual information in a single view. It efficiently achieves accurate and robust multiview similarity measure.
2) We propose to construct cost volume in a global-to-patch manner, to juggle reconstruction quality with method efficiency. As the resolution of the processed images augmented, the cost volumes are constructed from global sparse samples to patch-based propagated samples. In this way, the traditional PM idea is introduced into a coarse-to-fine learning-based MVS framework. Moreover, an adaptive propagation strategy is proposed to provide more diverse depth candidates.
3) We verify the performance and generalization property of the proposed approach for the 3-D scene reconstruction tasks on three MVS benchmarks: DTU [11], Tank and Temples [23], and ETH3D high-res [24]. CR-MVSNet achieves top performance in terms of construction overall quality and method efficiency on three benchmarks.

II. Related Work

A. Multiview Stereo

Based on plane-sweeping stereo [25], many learning-based depth map estimation for reconstruction methods appeared in recent years [8], [9], [15], [16], [26], [27]. MVSNet [8] first proposed to construct a global cost volume based on the extracted feature maps instead of RGB information and then used a 3-D-CNN for cost volume regularization. To reduce the memory consumption of the network, R-MVSNet [9] proposed to use a 2-D GRU recurrent network instead of a large 3-D-CNN, but induced longer inference time. Afterward, Cas-MVSNet [16], CVP-MVSNet [27], and UCSNet [15] proposed to construct multiscale cost volumes based on decreasing depth candidates centered at previous estimation result. These three works estimated depth maps in a coarse-to-fine manner, reduced the memory consumption, and thus can support higher resolution reconstruction tasks. Nevertheless, these approaches all used the variance-based metric to fuse pairwise cost volumes, which assumed that a given pixel is visible in all the source views.

B. Visibility Estimation

Some traditional approaches used pairwise matching similarity statistics to set heuristic cost thresholds for multiview visibility estimation [17], [28]. Kang et al. [28] only used the best 50% pairwise matching similarities for depth map estimation to alleviate the influence of invisible views, whereas [17] changed to select the best K neighboring views for each pixel, each depth hypothesis. Besides, Xu et al. [18] proposed a multidepth hypothesis joint voting strategy to infer pixelwise visible view set.

Visibility information is taken into account implicitly in the majority of the learning-based MVS approaches. DeepMVS [29] used maxpooling of pairwise cost volumes to make the network learn to aggregate the cost of the most credible neighboring view, which can estimate high-quality depth maps efficiently. However, in this way, the aggregated cost volume is solely associated with the most credible source view. Vis-MVSNet [30] estimated depth map uncertainty of different source views with the concept of entropy, and then, by fusing independently predicted depth maps from different source views, the final depth map can be obtained. However, the robust multiview similarity measures produced by multiple

observations would still be lost more or less. Soon afterward, some learning-based MVS approaches proposed to consider visibility information explicitly. PVSNet [31] and PatchmatchNet [10] exploited designed 3-D-CNNs to predict the visibility maps of source views, incorporated with weighted pairwise cost volume fusion metric, and the weight of unmatchable views would be reduced. Although the results surpass the previous approaches in reconstruction accuracy and completeness, the adopted U-Net structured 3-D-CNNs are time-consuming and memory-consuming. Inspired by these works, we propose a more efficient CRCA module to adaptively aggregate the cost of multiple source views, weighted by the inferred visibility maps. In addition, a cross-entropy loss is used to supervise the aggregated multiview cost volume, to promote accurate depth regression.

C. PatchMatch Stereo-Based MVS

The key step of the PM stereo pipeline is the propagation strategy, which decides the set of depth candidates to propagate to the central pixel for evaluation. It is essential to method inference efficiency. Early traditional PM-based methods [2], [19], [32], [33] adopted the sequential propagation scheme. They propagated leftward/rightward in odd iteration steps and upward/downward in even steps. Soon afterward, Galliani et al. [17] adopted a checkerboard scheme for diffusion-like propagation to speed up computation. Different versions of diffusion-like propagation strategies have been developed [18], [34], [35], [36]; Xu et al. [18] proposed an adaptive checkerboard propagation strategy, guided by multi-depth hypothesis joint view selection and multiscale geometric consistency. Subsequently, Wang et al. [10] proposed another adaptive propagation strategy to learn different positions to propagate using a deformable convolution network. However, robust feature representation of input images is heavily relied on in such an implicit manner. In our method, we use partly fixed sparse points within a fixed window as a propagation strategy [Fig. 1(b)]. Through the proposed propagation strategy, larger regions are explored to adaptively propagate better depth candidates to the central pixel.

III. Methodology

A. Overall Architecture

The overall structure is illustrated in Fig. 2. CR-MVSNet first extracts the feature maps of input images at three resolutions (Section III-B). Then, the final depth map is estimated through four stages, using multiscale feature maps and RGB information accordingly. In the first three stages, intermediate multistage depth maps are predicted through the pairwise cost volumes' construction step and CRCA module, presented in Section III-C and III-D, respectively. The detailed procedure of each stage is illustrated in Fig. 3. We mainly introduce the proposed global-to-patch manner for pairwise cost volume construction (Section III-C1) and the proposed visibility map prediction module for efficient intercost volume aggregation (Section III-D1), which is an essential component of CRCA. In the last stage, as in previous MVS work, a spatial refinement

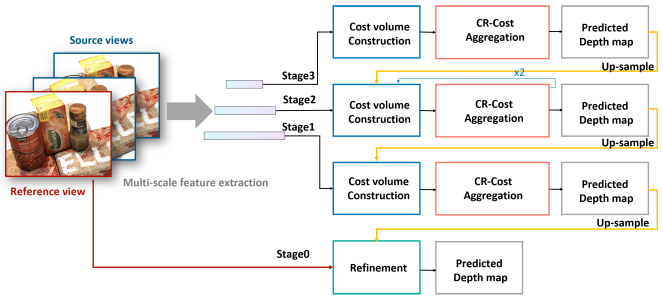


Fig. 2. Overview of CR-MVSNet. The proposed framework constructed cascade cost volumes to predict the high-resolution depth map from coarse to fine, from stage 3 to stage 0. At stage k, the resolution of the predicted depth map is $(W/2^k) \times (H/2^k)$, with input images of size $W \times H$.

module is used for fine-scale structures (Section III-F). Multiple designed loss functions are introduced in Section III-G to supervise the correctness of the intermediate results. Our multistage network estimates the high-resolution depth map from coarse to fine.

B. Multiscale Feature Extraction

To strengthen the robustness of the multiview similarity to illumination variation and low texture regions, we used a small feature pyramid network (FPN) [37] to convert RGB images into multiscale feature maps. The feature extraction network consists of three downsampling and two upsampling operations, and the output feature maps are at three resolutions, enabling the depth map inference at multiple resolutions in a coarse-to-fine manner. Given an input reference image $I_0 \in \mathbb{R}^{H\times W\times 3}$ and source images $\{I_i\}_{i=1}^{N-1} \in \mathbb{R}^{H\times W\times 3}$, the output three scales' feature maps $\{F_{i,3}\}_{i=0}^{N}$, $\{F_{i,2}\}_{i=0}^{N}$, $\{F_{i,1}\}_{i=0}^{N}$ are of resolution $(W/8) \times (H/8)$, $(W/4) \times (H/4)$, and $(W/2) \times (H/2)$, with 64, 32, and 16 channels, respectively.

C. Cost Volume Construction

The initial pairwise cost volumes are constructed via two steps: 1) depth candidates' generation in the proposed global-to-patch manner (Section III-C1) and 2) cost volumes construction by applying differentiable warping and groupwise correlation based on generated depth candidates (Section III-C5).

1) Global-to-Patch Depth Candidates' Generation: In our method, depth candidates are generated several times, using different manners in each stage, called the global-to-patch (GP) manner. The number of generated depth hypotheses decreases as the input feature resolution augments from stage 3 to stage 1. The proposed GP manner is for generating few but surrounding the actual surface depth candidates for efficient and accurate depth surface measurement. To the best of our knowledge, this is the first attempt to generate candidates using different strategies, from the global uniform manner to patch-based propagation manner, considering the uncertainty of previous estimates and the structured region information. Then, we present the detailed process for depth candidate generation in different stages.

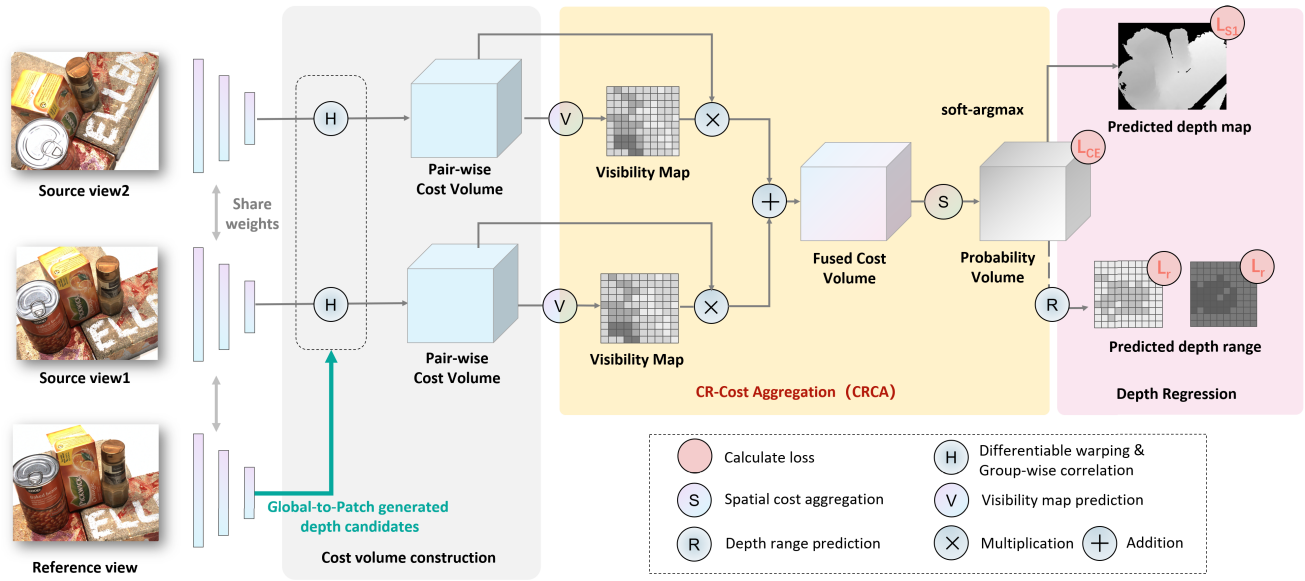


Fig. 3. Detailed process in a stage. For the initial coarsest stage 3, sparse depth candidates are generated from the global depth space. For the subsequent stages, depth candidates are progressively changed to be generated by patch-based propagation mode. Finally, local perturbation provides depth candidates around the previous estimates for local optimization. The proposed depth candidates' generation strategy is called the global-to-patch manner. The inferred visibility maps are used as fusing weights of pairwise cost volumes to eliminate the unreliable matching results caused by the invisibility of source views.

2) Sparse Depth Candidates' Generation: In stage 3, the depth candidates are generated randomly in the *inverse* depth space $[(1/d_{\min}), (1/d_{\max})]$, to encourage diversification and a more discriminative cost volume [10], [38]. We divide the *inverse* depth range evenly into D_s intervals and then generate a hypothesis randomly in each interval. Thus, D_s sparse depth candidates are generated uniformly in the *inverse* depth space.

3) Propagation and Local Perturbation: In the subsequent stages, the depth candidates are mainly generated by propagation and or local perturbation. For the propagation strategy, we adopt the checkerboard pattern as baseline [18], shown in Fig. 1(a). We note that when more depth candidates are propagated to the central pixel, the probability of sampling outliers with high matching costs rises. Different from [18], the proposed propagating pattern is shown in Fig. 1(b). In the intuition that the following local perturbation could generate similar candidates with the samples near the central pixel, we exclude them to reduce redundancy. We expand samples into eight extended strip regions. Each strip region in the up, down, left, and right direction contains two samples, and each other region contains four potential samples. Then we sample $D_n(t)$ fixed positions in up, down, left, and right directions as well as $D_n(t)$ good candidates in other strip regions depending on their previous multiview matching costs. As the candidates holding smaller multiview matching costs signify the reliable estimations, we finally get $2 - D_n(t)$ good candidates favored by the structured region information through the propagation. As for the local perturbation, we generated $D_p(t)$ candidates for each pixel p where the perturbation interval $R_t = 2^{-t}$, the variable t denotes the current iteration and the center of perturbation is set to be the estimation from the previous iteration. Adding the local perturbation, we get more diverse candidates than solely adopting propagation, which can refine the result locally.

4) Depth Candidates' Generation From Predicted Depth Range: Notably, in stage 3, after depth map regression of the sparse depth candidates, we obtain the depth map at the lowest resolution $[(W/8) \times (H/8)]$. At this step, this depth map is upsampled to enter the processing at stage 2. This upsampled depth map has limited accuracy. Irregular sampling induced by propagation may cause bias in depth regression, and the estimated depth is more likely to be influenced by similar hypotheses. Therefore, instead of performing a local perturbation operation, we evenly generated D_r depth samples from the predicted inverse depth range (see Section III-E), to provide diverse depth candidates centered at previous estimation.

5) Pairwise Cost Volumes' Construction: As in previous work [8], [10], the pairwise cost volumes are constructed using differentiable warping operation. The extracted feature map of the reference image is warped into source views based on the generated depth candidates. In particular, with intrinsic, rotations, and translations matrices $\{K_i, R_i, t_i\}_{i=0}^{N-1}$ of the corresponding view i, obtained from the classic structure from motion (SfM) algorithm, we compute the corresponding pixel coordinate in the source view $\{I_i\}_{i=1}^{N-1}$ for a pixel p in the reference image I_0 as follows:

$$p_{i,j} = K_i \cdot \left(R_i \cdot R_0^{-1}\left(d_j \cdot K_0^{-1} \cdot p - t_0\right) + t_i\right) \tag{1}$$

where d_j indicates the jth depth candidate for the pixel p, from which we can obtain the warped feature maps of source view i with d_j, denoted as $F_i(p_{i,j})$, corresponding to the differentiable bilinear interpolation in feature map of source view i. Warped source feature maps and reference feature map are denoted as $F_i(p_{i,j})$ and $F_0(p)$, respectively. To speed up computation, the feature channels C are then divided evenly into G groups, denoted as $F_i(p_{i,j})^g$ and $F_0(p)^g$. Subsequently, we calculate the similarity of feature pair for each depth

candidate via groupwise correlation, computed as

$$S_i(p,j)^g = \frac{G}{C}\cdot < F_0(p)^g, F_i\left(p_{i,j}\right)^g > \tag{2}$$

where $<\cdot,\cdot>$ is the inner product, $g = 0, 1, \ldots, G-1$, and the obtained $S_i(p,j)^g \in \mathbb{R}^{G\times D\times H\times W}$, $S_i(p,j)^g$ is actually a similarity volume (negative cost volume). This pairwise cost volume representation measures the matching quality over the set of depth candidates $\{d_j\}_{j=0}^{D-1}$ for reference view and source view i.

D. CR-Cost Volume Aggregation

In CR-MVSNet, the proposed Co-visibility Reasoning cost volume aggregation (CRCA) is used at each stage to get robust matching measures. The process consists of intercost volume aggregation and intracost volume aggregation. The former makes the network learn to distinguish unrelated views. The latter mitigates the ambiguity of similarity measures originating from noisy and low-textured areas.

1) Visibility Map Prediction Module For Efficient Intercost Volume Aggregation: After obtaining pairwise cost volumes $S_i(p,j)^g \in \mathbb{R}^{G\times D\times H\times W}$, a robust fusing mechanism is essential for obtaining a clear multiview cost volume. For this purpose, we proposed a visibility map prediction module (VM). By measuring the randomness of the pixelwise probability distribution, we explicitly estimate the visibility map of each source image I_i and use the inferred visibility maps as the pairwise cost volumes' fusing weight coefficients, and the goal of reducing matching disturb caused by the invisibility of source views can be achieved.

The constructed 4-D cost volumes $S_i(p,j)^g$ are first compressed into 3-D probability volumes $P_i(p,j) \in \mathbb{R}^{D\times H\times W}$ by a lightweight network composed of 3-D-CNNs and a sigmoid activation function. The elements in the tensor $P_i(p,j)$ represent the probability of the jth depth hypothesis at pixel $p_{i,j}$. We thus calculate the visibility map of source view i as follows:

$$\begin{aligned} V_i(p) &= f\left(\sum_{j=0}^{D-1}(1/D)\cdot\log\left(\frac{1/D}{P_i(p,j)}\right)\right) \\ &= f\left(\sum_{j=0}^{D-1}(-1/D)\cdot\log(D\cdot P_i(p,j))\right) \end{aligned} \tag{3}$$

where D is the number of depth candidates for each pixel. As shown in (3), we first compute the Kullback–Leibler divergence (KL-Divergence) between $P_i(p,j)$ and the uniform distribution at the depth dimension and then use a shallow 2-D-CNN layer f to consider the neighbors' information. KL-Divergence is commonly used to measure the similarity between probability distributions. In this equation, we use it to measure the randomness of a cost distribution based on a set of depth candidates for a single pixel. When the cost distribution is completely random, the KL-Divergence between the obtained and uniform distribution tends to be zero (minimal), which intuitively corresponds to an invisible view; conversely, when a cost distribution shows a unimodality form, KL-Divergence augments as the unimodality becomes shaper, corresponding to a visible view. For VM, we focused on whether the source view is visible for a single pixel. The accuracy of the depth corresponding to the peak distribution is supervised by the final regressed depth at the last step of each stage.

The fused cost volume for reference view is the average of the pairwise cost volumes weighted by their corresponding visibility maps, computed by

$$\widetilde{S}(p,j)^g = \sum_{i=1}^{N-1}\frac{V_i(p)}{\sum_{i=1}^{N-1}V_i(p)}\cdot S_i(p,j)^g. \tag{4}$$

The fused multiview cost volume $\widetilde{S}(p,j)^g \in \mathbb{R}^{G\times D\times H\times W}$ takes the visibility of each source view into account and improves the depth estimation accuracy. After the process of VM, a lightweight network containing three 3-D-CNN layers is applied to compresses the 4-D cost volume $\widetilde{S}_i(p,j)^g$ into a 3-D cost volume $C(p,j) \in \mathbb{R}^{D\times H\times W}$.

2) Intracost Volume Aggregation: After aggregated cost volume of different views, similar to other MVS methods [8], [16], we propose to aggregate spatial costs. Previous learning-based MVS methods usually apply a 3-D U-Net structure for spatial cost aggregation and regularization, which is time- and memory-consuming. Different from them, we follow the traditional MVS methods [6], [10] to aggregate costs over a spatial window, which is smaller than the propagation step. In this way, we can aggregate the cost over pixels located in the same 3-D space plane. For a spatial window having K_n pixels $\{p_k\}_{k=1}^{K_n}$, the aggregated spatial cost is defined as

$$\widetilde{C}(p,j) = \frac{1}{\sum_{k=1}^{K_n}w_k}\cdot\sum_{k=1}^{K_n}w_k\cdot C(p_k,j) \tag{5}$$

where $\{w_k\}_{k=1}^{K_n}$ are the weights of neighbors based on the feature similarity between the central pixel and its neighbors' samples.

E. Joint Estimation of Depth Map and Depth Probability Distribution

At the final phase of each stage, the conventional depth regression manner [8], [16] is used to obtain the depth prediction

$$\widehat{D}(p) = \sum_{j=0}^{D-1} d_j\cdot\widetilde{P}(p,j) \tag{6}$$

where the probability distribution over all the depth hypotheses $\{\widetilde{P}(p,j)\}_{j=0}^{D-1}$ is transformed from the fused negative cost volume $\widetilde{C}(p,j)$, by applying the softmax operation. Furthermore, in addition to the regressed depth map being supervised, we consider the constraint for the intermediate probability volume is more direct. For a set of depth hypotheses, many probability combinations can be weighted and summed to the same depth. The only implicit constraint for the regressed depth increases the difficulty of the model convergence. Thus, we proposed that the intermediate probability volume should be output and supervised as well. We supervise the output depth map and probability volume using the ground-truth

depth map and ground-truth distribution, the latter ground truth being a normalized Laplacian distribution centered at the ground-truth depth map. We promote the network to finally get a unimodal probability distribution centered at the ground-truth depth map, thus can regress a more accurate depth map.

After getting the regressed depth map, to provide diverse depth candidates centered at the previous estimates, inspired by Cheng et al. [15], we propose a depth range prediction module (RP) by calculating the variance coefficient cv of probability volume in the depth direction. The predicted depth range is used to transform the depth candidates' generation strategy from uniform to patch-based propagation manner. Compared with the standard deviation used in [15], cv is affected not only by standard deviation but also by the regressed depth value of the current stage. For a distant pixel (with a great depth value), finding consistent and accurate matching locations in the other source views is more challenging. Thus, a more confident estimate than a close pixel is indicated by the same standard deviation of the probability distribution of depth candidates. cv is fairer to represent the uncertainty of the prediction and more beneficial for large scenes with wider depth ranges. The following is a detailed description of the process. First, the variance $\widehat{V}(p)$ of the probability distribution is computed as

$$\widehat{V}(p) = \sum_{j=0}^{D-1} \widetilde{P}(p, j) \cdot \big(D(p, j) - \widehat{D}(p)\big)^2 \tag{7}$$

where $D(p, j)$ denotes the set of generated depth samples for cost volume construction, and $\widehat{D}(p)$ denotes the estimated depth map. Then, the corresponding cv is $\widehat{cv}(p) = (\widehat{V}(p))^{1/2}/\widehat{D}(p)$, and the depth range is calculated as

$$R(p) = \big[\widehat{D}(p) - \epsilon_r \Delta d \cdot \widehat{cv}(p), \widehat{D}(p) + \epsilon_r \Delta d \cdot \widehat{cv}(p)\big] \tag{8}$$

where Δd is the approximate depth range of scenes precalculated with camera parameters, and ϵ_r is a scale parameter used to adjust the size of the predicted depth range. The RP is only applied at the end of stage 3, to provide uniformly generated depth candidates in the predicted depth range for the next stage rather than local perturbation. Because the estimates at stage 3 have limited accuracy, local perturbation around the unreliable estimation will mislead the network. Contrarily, the depth candidates provided by RP possess adaptive depth intervals depending on their probability distribution. A shaper probability distribution tends to predict a smaller depth interval, which means a more confident estimation, and vice versa.

F. Refinement

For the highest resolution, the regressed depth map from probability volume provides a reliable estimation. Direct upsampling is not sufficient for accuracy, and the reconstruction edges may suffer from oversmoothing as a similar set of depth candidates is shared for a relatively large receptive field. Similar to previous works [8], [10], a depth residual network is used to refine the depth map estimation in detailed areas. Specifically, the original RGB image and the upsampled depth map are concatenated as an input with four channels. Then, the depth residual is learned by applying multiple 2-D-CNNs to the input. Finally, the learned depth residual is added back to the previous estimation to obtain the final output. Also, the upsampled depth map is prescaled to [0,1] to avoid of being biased for a specific depth scale and is converted back to its original scale after refinement.

G. Loss Function

The total loss L_{total} is composed of the losses among multiscale depth map estimations L_{s1}^k and their corresponding probability distributions L_{ce}^k $(k = 1, 2, 3)$, the loss of the depth range prediction L_r^3, and the loss of final depth map estimation from the refinement module L_{ref}^0, as shown in (9), where α_k and β are the balancing weights.

We adopted the smooth L1 loss and cross-entropy loss [34], [39], to respectively, supervise the depth estimations and probability distribution of each stage, as shown in (10) and (11). $\widetilde{P}^k(p, j)$ is the predicted probability distribution in (6), $P_{gt}^k(p, j)$ is the ground-truth probability distribution, which is a normalized Laplacian distribution centered at the inverse depth ground truth, and I denotes the number of valid pixels.

With $\epsilon > 0.5$, the inferred upper boundary is encouraged to be slightly greater than the ground truth (13), and the inferred lower boundary is encouraged to be slightly lower than the ground truth (14). The sum of these two equations is designed to supervise the predicted depth range L_r^3 (12)

$$L_{\text{total}} = \sum_{k=1}^{3} \alpha_k \big(L_{s1}^k + L_{ce}^k\big) + \beta L_r^3 + L_{ref}^0 \tag{9}$$

$$L_{s1}^k = \frac{1}{I} \sum_{n=1}^{I} l_{s1}\big(\widehat{D}^k(p), D_{gt}^k(p)\big) \tag{10}$$

$$L_{ce}^k = \frac{1}{I} \sum_{n=1}^{I} \sum_{j=0}^{D-1} -\widetilde{P}^k(p, j) \cdot log\, P_{gt}^k(p, j) \tag{11}$$

$$L_r^3 = L_{\text{upper}}\big(U(p) - D_{gt}^3(p)\big) + L_{\text{lower}}\big(L(p) - D_{gt}^3(p)\big) \tag{12}$$

$$L_{\text{lower}}(x) = \begin{cases} (1-\epsilon) l_{s1}(x), & \text{if } x > 0 \\ \epsilon l_{s1}(x), & \text{otherwise} \end{cases} \tag{13}$$

$$L_{\text{upper}}(x) = \begin{cases} \epsilon l_{s1}(x), & \text{if } x > 0 \\ (1-\epsilon) l_{s1}(x), & \text{otherwise.} \end{cases} \tag{14}$$

IV. Experiments

The proposed method is evaluated on the DTU [11], Tanks and Temples [23], and ETH3D high-res [24] datasets, and the proposed components are analyzed through ablation experiments.

Datasets: 1) the DTU dataset [11], captured under laboratory conditions, is an indoor MVS dataset. The dataset can be divided into training, testing, and validation sets following [40]; 2) the Tanks and Temples benchmark [23], acquired in realistic conditions, is an outdoor MVS dataset. It consists of intermediate and advanced datasets; and 3) ETH3D high-res benchmark [24] consists of images of scenes with viewpoints changed greatly. Moreover, in the ETH3D high-res benchmark,

the number of images contained in each scene varies greatly, from more than a dozen to hundreds. Therefore, it is a more challenging dataset compared with the previous. It is composed of the train dataset and test dataset.

Implementation Details: The proposed work is developed in the PyTorch framework and the model is trained on the DTU training set [11]. The input images' resolution is set to 640×512, and the number of input images is $N = 5$. For depth hypotheses' generation, at stage 3, the generated sparse depth number is $D_s = 48$, at stage 2, the depth generation and cost aggregation process are iteratively applied two times, where for the first and second iterations, the number of propagation hypotheses is, respectively, $D_n(0) = 8$, $D_n(1) = 4$, the number of perturbation hypotheses is, respectively, $D_p(0) = 16$, $D_p(1) = 8$, and the number of generated depth hypotheses in predicted range is $D_r = 10$; for spatial cost aggregation, $K_n = 9$ for all the stages, and for the depth range prediction module, we specify $\epsilon_r = 2$. The weights used in loss terms are set to $\alpha_3 = \alpha_2 = \alpha_1 = 0.7, \beta = 0.5$, $\epsilon = 0.9$. We train our model with the Adam optimizer for ten epochs and the learning rate is set to 0.001. The training batch size is set to 4, and the model is trained on two Nvidia GTX 1080Ti GPUs. After depth estimation, we first filtered depth maps with predicted photometric confidence maps and geometric consistency check and then fused them to point clouds for evaluation, as in previous MVS work [8].

A. Evaluation on DTU Dataset [11]

Following the evaluation metrics provided by the DTU dataset [11], we set $N = 5$ and image resolution as 1600×1200. As qualitative comparison shown in Fig. 4, our approach reconstructs a denser and smoother point cloud with finer details at structure edge areas and textureless areas, thanks to our efficient cost aggregation and global-to-patch depth generation strategy. As quantitative evaluation results reported in Table I, compared with [10], our approach maintains competitive completeness and improves accuracy in terms of reconstruction quality, thus having a higher overall reconstruction quality. We all contribute to the depth generation strategy. Wang et al. [10] use the adaptive propagation strategy to gather the candidates from pixels of the same surface, thus contributing to high completeness. Our proposed global-to-patch manner generates candidates not only through patch-based propagation strategy but also through the predicted depth range, which is calculated based on the uncertainty of the previous estimate. The candidates from the predicted depth range can prevent bias in regression caused by propagation from inaccurate estimates. Compared with other work [15], [16], [27], [38], [41] generating uniformly distributed depth candidates, our propagated candidates are irregular in the depth direction, which makes it more difficult to aggregate cost along the depth dimension using the 3-D-CNN, thus resulting in a lower accuracy quality. However, benefitting from the propagated depth candidates supported by the structured region information, the higher completeness quality helps us to surpass all the counterparts and get the highest overall reconstruction quality.

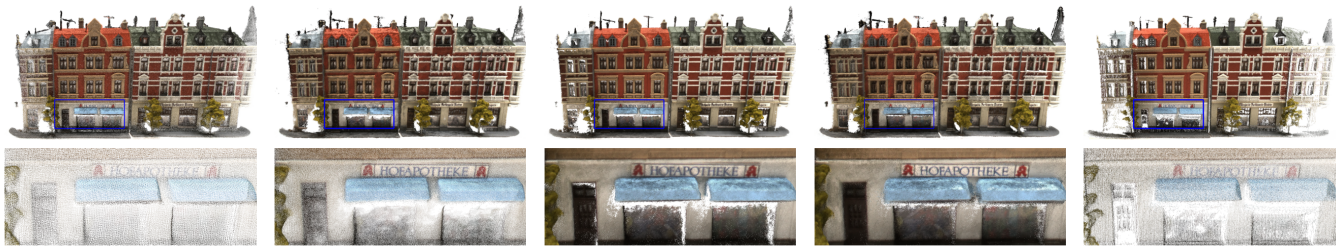
TABLE I
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON DTU'S EVALUATION SET [11]

Methods	Acc.(mm) ↓	Comp.(mm) ↓	Overall(mm) ↓
[2015] Gipuma [17]	**0.283**	0.873	0.578
[2018] MVSNet [8]	0.396	0.527	0.462
[2019] R-MVSNet [9]	0.383	0.452	0.417
[2020] CIDER [38]	0.417	0.437	0.427
[2020] Fast-MVSNet [43]	0.336	0.403	0.370
[2020] UCS-Net [15]	0.338	0.349	<u>0.344</u>
[2020] CasMVSNet [16]	0.325	0.385	0.355
[2020] PVA-MVSNet [42]	0.379	0.336	0.357
[2020] CVP-MVSNet [27]	<u>0.296</u>	0.406	0.351
[2021] PatchmatchNet [10]	0.427	**0.277**	0.352
[2022] IterMVS [41]	0.373	0.354	0.363
Ours	0.405	<u>0.278</u>	**0.342**

The accuracy and completeness are calculated by using the official MATLAB script provided by the DTU dataset. Numbers in bold and underlined represent the best and the second, respectively.

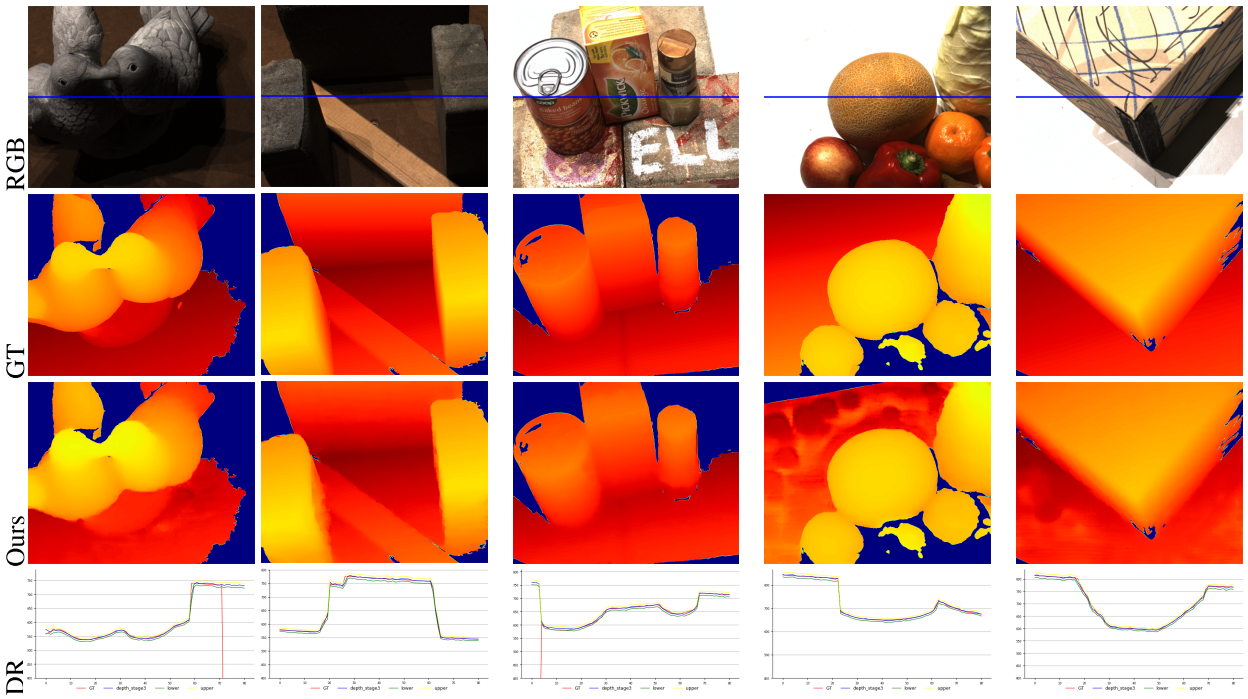
1) Evaluation of Depth Range Prediction: The predicted depth range can be visualized in Fig. 5, we note the predicted depth ranges cover the ground truth for most pixels, and the predicted depth ranges are almost in small depth spaces. This is made possible by the differentiable depth range prediction module and the end-to-end training process. Under the supervision of the depth range loss (13) and (14), the network learns to control the probability distribution to obtain the proper depth range. Because of this, we tested different $\epsilon_r = 1, 2, 3$ but got similar predicted depth ranges in the end. As a result, our network is not very sensitive to different ϵ_r. The proposed depth range prediction module helps the network focus on a small depth range instead of the global depth space in the subsequent stage for efficient and accurate depth map estimation.

2) Memory and Run-Time Comparison: The proposed CR-MVSNet is compared with MVSNet [8], CasMVSNet [16], and PatchmatchNet [10] in terms of memory and run-time performance. For a fair comparison, a fixed input size of 1152×864 is used to evaluate the computational cost on a single GPU of NVIDIA GeForce GTX 1080Ti. As shown in Table II, the proposed CR-MVSNet achieves 70.5% memory savings and 69.6% run-time reduction compared with MVSNet [8]. Compared with CasMVSNet [16], CasMVSNet achieves nearly the same overall accuracy as ours but has much more memory requirements. We all realized the depth map estimation in a coarse-to-fine manner, but the adopted global-to-patch manner for multiscale cost volume construction and CRCA module for cost volume aggregation are more efficient for depth map estimation. Compared with PatchmatchNet [10], PatchmatchNet has better memory efficiency and run-time but is limited in its reconstruction ability. We all used the patch-based propagation strategy for cost volume construction. In addition, we generated more diverse depth hypotheses from the predicted depth range in the proposed global-to-patch manner, and the proposed CRCA considered the co-visibility relationship in multiview, which contributes to higher construction quality. Overall, our method is a satisfying tradeoff for reconstruction quality and method efficiency.



(a) (b) (c) (d) (e)

Fig. 4. Visual comparison of reconstructed point clouds of scan 15 in the DTU dataset [11]. The zoomed local area shows that our result is more complete and denser. Compared with the ground truth obtained from structured light scanning, the depth map fusion manner can infer denser points for better visual experience. (a) R-MVS [9]. (b) PVA-MVS [42]. (c) UCS-Net [15]. (d) Ours. (e) Ground truth.



(a) (b) (c) (d) (e)

Fig. 5. Visualization of predicted depth range on the DTU dataset [11]. The predicted depth top boundary (yellow line) and bottom boundary (green line) are plotted in the last row (DR), corresponding to the pixels at the blue lines of the first row images. For most pixels, the predicted depth range can correctly cover the ground truth (red line), and the size of the predicted range is small for efficient cost volume construction. Higher ranges often happen at boundary pixels, which indicates a more uncertain estimation. (a) Scan 106. (b) Scan 126. (c) Scan 5. (d) Scan 64. (e) Scan 10.

TABLE II

COMPARISON OF GPU MEMORY, RUN-TIME, AND OVERALL RECONSTRUCTION QUALITY BETWEEN THE PROPOSED METHOD AND OTHER STATE-OF-THE-ART LEARNING-BASED METHODS ON THE DTU DATASET [11]

Methods	GPU Memory(MB)↓	Run-time(s)↓	Overall(mm)↓
[2018] MVSNet [8]	10632	1.435	0.551
[2020] CasMVSNet [16]	5667	0.459	0.355
[2021] PatchmatchNet [10]	**2323**	**0.417**	0.374
Ours	3135	0.436	**0.350**

B. Evaluation on Tanks and Temples [23]

In this experiment, we use the trained model on DTU without any fine-tuning to evaluate the generalization of our model. The image resolution for evaluation on the Tanks and Temples dataset is set to 1920 × 1056, and the number of input views is set to 7. The camera parameters and predetermined depth range are calculated with OpenMVG by Wang et al. [10]. As shown in Table III, our method performs the best among all the compared methods on the intermediate and more complex advanced dataset. The evaluated scenes in the Tanks and DTU datasets have different depth ranges and various view changes modes. Thanks to the proposed visibility prediction module for efficient intercost volume aggregation, the disturbance brought by unrelated views is weakened, thus making our model more stable and well adapted to these entirely different scenes. The comparison results indicate the better generalizability of our proposed CR-MVSNet than recent SOTA methods.

C. Evaluation on the ETH3D dataset [24]

As in the previous evaluation, we used the trained model on DTU without any fine-tuning. The image resolution for evaluation on the ETH3D high-res dataset is set to 2688 × 1792,

TABLE III

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE TANKS AND TEMPLES DATASET [23]

Methods	Intermediate(%) ↑	Advanced(%) ↑
[2016] COLMAP [44]	42.14	27.24
[2018] MVSNet [8]	43.48	-
[2019] R-MVSNet [9]	48.40	24.91
[2019] Point-MVSNet [14]	48.27	-
[2020] P-MVSNet [45]	55.62	-
[2020] Fast-MVSNet [43]	47.39	-
[2020] UCS-Net [15]	54.83	-
[2020] CasMVSNet [16]	56.42	31.12
[2020] CVP-MVSNet [27]	54.03	-
[2020] PVSNet [31]	56.88	33.46
[2021] PatchmatchNet [10]	53.15	32.31
[2021] PatchMatch-RL [46]	51.82	31.81
[2022] IterMVS [41]	56.22	33.24
Ours	**57.21**	**34.05**

The evaluation metric is the mean F score [23], which comprehensively considers the accuracy and completeness of the reconstructed point clouds. It is worth noting that most approaches give up evaluating the more challenging advanced set. Numbers in bold and underlined represent the best and the second, respectively.

TABLE IV

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON ETH3D HIGH-RES BENCHMARK [24]

Methods	Training		Testing	
	F_1 score(%)↑	Times(s)↓	F_1 score(%)↑	Times(s) ↓
[2015] MVE [47]	20.47	13278.69	30.37	10550.67
[2015] Gipuma [17]	36.38	587.77	45.18	689.75
[2016] COLMAP [44]	**67.66**	2690.62	73.01	1658.33
[2016] COLMAP (low) [44]	62.84	-	67.30	-
[2020] PVSNet [31]	67.48	-	72.08	829.56
[2021] PatchmatchNet [10]	64.21	**452.63**	73.12	**492.52**
Ours	64.83	557.53	**73.37**	612.36

The evaluation metric is the mean F_1 score [24]. Notably, PVSNet and PatchmatchNet [10], [31] are presently the rare competitive pure learning-based approaches on this challenging benchmark. Numbers in bold and underlined represent the best and the second, respectively.

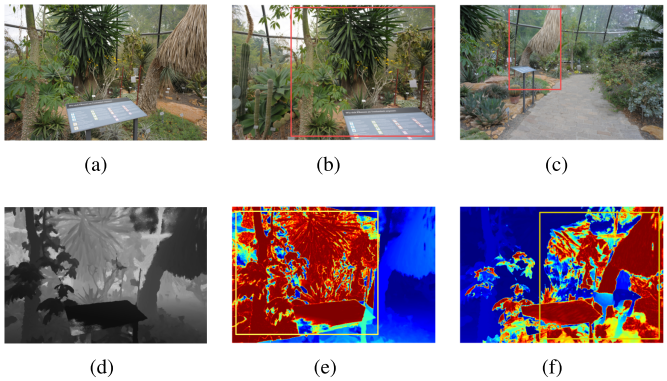


Fig. 6. (e) and (f) Visualization of the predicted visibility maps, which present the estimated co-visible regions between the reference view and source views, observed from the reference view. The red color indicates higher weight for intercost volume aggregation. (d) Depth map estimation result. (a) Reference image. (b) Source image 1. (c) Source image 2. (d) Depth map estimation. (e) Visibility map of source view 1. (f) Visibility map of source view 2.

and the number of input views is set to 7. The camera parameters and predetermined depth range are calculated using Colmap [44].

For this challenging benchmark that contains strong viewpoint variations, we compared the proposed approach with five state-of-the-art MVS methods in terms of reconstruction quality and inference time. The MVE [47], Gipuma [17], and COLMAP [44] are three traditional MVS methods proposed previously but with significant influence until now. Without considering visibility estimations, almost all the learning-based DTU-trained methods failed on this benchmark. PVSNet [31] and PatchmatchNet [10] are two rare counterparts that have been proposed in the past three years. The results are shown in Table IV. Our method is the second fastest after PatchmatchNet [10]. It is because we generated more depth candidates from the predicted depth range for evaluation. Thanks to the extra generated diverse depth hypotheses, we achieve better F_1 score results on both the training and testing sets than [10]. Compared with other methods, we witness a better performance than several traditional methods, such as MVE [47] and Gipuma [17]. Although the traditional COLMAP [44] and the learning-based PVSNet [31] perform better F_1 scores than ours on the training set, the opposite case occurs on the particularly challenging testing set. Note that the resolution of input images for COLMAP evaluation on the ETH3D high-res benchmark is set to be 3200×2130, which is higher than ours, 2688×1792. For this reason, COLMAP costs five times longer inference time than ours. When the resolution of input images is set to 2688×1792 for COLMAP, our reconstruction results are better than COLMAP with low-resolution input on both the training and testing sets. The results show that the proposed approach generalizes well to this challenging dataset. Compared with PVSNet, we all considered visibility estimation, but ours used the measured randomness of pixelwise probability distribution followed by lightweight 2-D-CNN layers to estimate visibility maps rather than PVSNet's ingenious yet large encoder–decoder structured 3-D-CNN. Besides, our spatial cost aggregation step is more lightweight than PVSNet's U-Net structured 3-D-CNN regularization network. All these contribute to a shorter inference time. Considering the performance of CR-MVSNet on the training and testing sets, our method is on par with PVSNet in terms of reconstruction quality and with a shorter inference time.

We exhibit our visibility map estimation results in Fig. 6. The visibility map results show the correctness of the co-visible area estimation result and can verify the effectiveness of the proposed CRCA module. The proposed CRCA module can learn to distinguish unrelated views and further get robust matching measure for high-quality depth estimation.

D. Ablation Study

The ablation experiments are mainly be conducted on the DTU dataset [11] to verify the improvement of the proposed components of our method. More concretely, we first verify the effectiveness of the main contributions: the proposed global-to-patch depth generation strategy and the proposed visibility map prediction module for efficient intercost volume aggregation. The latter is the vital component of the co-reasoning cost aggregation module (CRCA). Then, we verify the impact of the proposed loss terms, including the depth range loss bonded with the proposed depth prediction module, applied only at the coarsest stage, and the probability loss, applied at multistages.

TABLE V
ABLATED RESULTS OF GLOBAL-TO-PATCH DEPTH GENERATION (GP) AND VISIBILITY MAP PREDICTION MODULE (VM) ON THE DTU AND ETH3D-TRAINING DATASETS

Method	Depth Num.	DTU evaluation set [11]			ETH3D Training set [24]		
		Acc.(mm)↓	Comp.(mm)↓	Overall(mm)↓	Accuracy(%)↑	Completeness(%)↑	F_1 score(%)↑
GP + NVM	[48,34+12,8]	0.409	<u>0.279</u>	0.344	61.11	<u>59.82</u>	60.56
UF-B + NVM	[64,48,8]	<u>0.339</u>	0.351	0.345	62.93	49.53	55.21
UF-S + NVM	[48,32,8]	0.379	0.356	0.368	61.78	46.16	52.02
FP + NVM	[48,36,8]	0.497	0.332	0.415	55.89	50.21	52.17
GP + VM (Ours)	[48,34+12,8]	0.405	**0.278**	**0.342**	<u>69.56</u>	**62.31**	**64.83**
UF-B + VM	[64,48,8]	**0.337**	0.348	<u>0.343</u>	**70.12**	56.23	<u>61.12</u>
UF-S + VM	[48,32,8]	0.368	0.355	0.362	67.19	54.12	59.38
FP + VM	[48,36,8]	0.434	0.329	0.382	58.66	56.25	57.37

Methods are expressed in the format of "Depth candidates generation step" + "Inter-cost volume aggregation step". The number of evaluated depth candidates from stage three to one are presented afterward. For the depth candidates generation step: GP denotes the proposed Global-to-Patch manner for depth candidates generation, 34 and 12 are the number of evaluated depth candidates for two iterations in stage two; UF denotes the uniform depth candidates generation strategy, centered around the estimation of the previous stage, introduced by [15]; FP denotes patch-based propagation pattern with fixed positions, as in Fig.1a, proposed by [18]. For the inter-cost volume aggregation step, we compared the reconstruction results without/with the visibility map prediction module, denoted as NVM and VM, respectively. Numbers in bold and underlined represent the best and the second, respectively.

Finally, we compare the reconstruction results from different numbers of input images.

1) Global-to-Patch Depth Candidates' Generation and Visibility Map Prediction Module: We compared the performance of the proposed GP manner with other popular depth candidates generation strategies and the version without/with the visibility map prediction module (NVM/VM) on the DTU [11] and ETH3D-training set [24]. The results are shown in Table V.

For the GP depth candidates' generation, with similar number of evaluated depth candidates, the uniform generation strategy (UF-S) tends to get higher accuracy quality, while the propagation-based depth candidates' generation strategy (FP and GP) tends to get higher completeness quality. We infer the reason is that the cost-volume-adopted UF strategy is regular, which means the depth candidates are distributed uniformly in the inverse depth range. The 3-D-CNN operations perform better for cost aggregation along the depth dimension for most areas. However, for the textureless areas, since matching results are unreliable, 3-D-CNN operations become ineffective. Thus, a large number of depth estimation results are filtered out by geometry consistency check, thus resulting in lower completeness quality. The cost-volume-adopted propagation-based strategy is irregular, which weakens 3-D-CNN's fitting ability in most areas, but the generated depth candidates are favored by structured region information. Thus, it is more favorable for depth estimation in textureless areas. For the proposed GP manner, it is a tradeoff that uniformly generated depth candidates for low-resolution and progressively changed to propagation manner as the input image resolution augments. We note that the proposed GP improved the reconstruction accuracy of FP (from 0.434 to 0.405) and the reconstruction completeness of UF-S (from 0.355 to 0.278) on the DTU dataset. Consequently, GP obtains promising results with the highest overall quality and is on par with UF-B+VM, which constructs larger cost volumes with more depth candidates. The evaluation results on the ETH3D-training set indicate that the proposed GP manner can get comparable or even better accuracy than the UF-S. We infer that the propagated reliable candidates from neighbors play an essential role, particularly for the scenes in the ETH3D-training set, with greater and varied depth ranges. Despite constructing irregular cost volume, we can achieve better reconstruction accuracy with the propagated reliable candidates surrounding the actual surface than with the widely used uniform strategy.

For the proposed visibility map prediction module (VM), we witness a surge in improvement for reconstruction results by adopting VM on the ETH3D-training set, but a minor improvement on the DTU dataset. As introduced in Section IV, images in the ETH3D-training set have strong viewpoint variations, while the images in the DTU set have slight viewpoint changes. This can explain why the proposed VM performs differently on two datasets. Considering the consistently favorable results compared with other popular depth generation strategies and ablated VM results across two different datasets, we objectively conclude that the proposed GP+VM is a promising tradeoff solution with stronger generalization ability than its counterparts.

2) Depth Range Prediction Module and Multistage Probability Loss: After verifying the effectiveness of the proposed GP manner depth generation, we evaluate the depth range prediction module (RP) involved, which is jointly used with the depth range loss applied at the third stage. The version without RP is denoted as NRP. As shown in Table VI, the proposed RP and depth range loss jointly improved the estimation result in completeness quality. This experiment shows that sampling from the predicted range can provide more diverse and reliable depth hypotheses than solely relying on the neighbors' propagation manner. Apart from that, our sparse depth generation and propagation also contribute to faster convergence.

After verifying the effectiveness of the proposed GP manner depth generation, we further evaluate the depth range prediction module (RP) involved, which is constrained by the designed depth range loss term L_d. Compared with the reconstruction results without/with RP and L_d (second line and fourth line) in Table VI, we note that the proposed RP and depth range loss jointly improve the estimates in completeness while the accuracy stays almost the same. We infer that sampling from the predicted range can provide more diverse and reliable depth hypotheses than solely relying on the

TABLE VI
ABLATED RESULTS OF LOSS TERMS ON THE DTU DATASET [11]

Methods	Acc.(mm)↓	Comp.(mm)↓	Overall(mm)↓
L_{s1}	0.427	0.331	0.379
$L_{s1} + L_{ce}$	0.406	0.324	0.365
$L_{s1} + L_d$	0.429	0.308	0.369
$L_{s1} + L_d + L_{ce}$ (Ours)	**0.405**	**0.278**	**0.342**

L_{s1} denote the conventional applied smooth L1 loss for the output multi-stage depth maps. L_{ce} denote the additional applied cross-entropy loss for the output multi-stage probability volumes. L_d denote the designed depth range loss for the proposed depth range prediction module, only applied at the coarsest stage. We omit subindices denoting the stage for simplicity. Numbers in bold and underlined represent the best and the second, respectively.

TABLE VII
RESULTS OF DIFFERENT NUMBERS OF VIEWS ON THE DTU [11], TANKS AND TEMPLES [23], AND ETH3D HIGH-RES DATASETS [24]

N	DTU		Tanks	ETH3D	
	Acc.(mm)	Comp.(mm)	F score	Acc.(%)	Comp.(%)
3	0.426	0.285	52.14	49.26	42.67
5	**0.405**	**0.278**	55.66	53.99	56.67
7	0.411	0.281	**57.21**	**69.56**	**62.31**

propagation manner, especially for the textureless area, with inaccurate initial estimates.

Afterward, we verify the impact of the additional cross-entropy loss, which is applied at multistages to supervise pixelwise probability distribution. Comparing the results without/with L_{ce} (first line vs. second line, third line vs. fourth line), the additional cross-entropy loss term improves the accuracy and completeness of reconstruction generally. The direct constraint for the intermediate probability distribution is nontrivial for the superior performance of CR-MVSNet.

3) Number of Views: We compared our model from different numbers of input image views $N = 3,5,7$ for the DTU evaluation set [11], Tanks and Temples intermediate set [23], and ETH3D high-res training set [24]. As shown in Table VII, when the number of input images augments, the reconstruction quality improves in terms of completeness and accuracy. Using more views can alleviate partially views' invisible problem and significantly boost the reconstruction quality on the ETH3D high-res dataset, with strong viewpoint variations, whereas for the DTU dataset, with small viewpoint variations, too many input views may lead to matching ambiguity.

V. CONCLUSION

In this article, we present a co-visibility reasoning-based MVS network, namely, CR-MVSNet. Thanks to the proposed visibility maps prediction module that helps eliminate the interference brought by unrelated views, the robust multiview similarity measure can be obtained. The proposed global-to-patch manner can generate a small number of depth candidates surrounding the actual surface with high probabilities since considering the uncertainty of previous estimates and the structured region information, thus deriving accurate surface depth measurements with low memory requirements. The effectiveness of the proposed components is verified through ablation experiments. Experiments on three popular 3-D reconstruction benchmarks—DTU [11], Tanks and Temples [23], and ETH3D high-res [24]—demonstrate the competitive performance of our proposed method. In future work, we will focus on optimizing the efficiency of our network and extending it to movable platforms with limited computation and memory resources.

REFERENCES

[1] P. Suo, J. Sun, W. Tian, S. Sun, and L. Xu, "3-D image reconstruction in planar array ECT by combining depth estimation and sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[2] J. L. Schnberger, E. Zheng, M. Pollefeys, and J. Frahm, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. IEEE/Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2016, pp. 501–518.

[3] S. Li, K. Chen, M. Song, D. Tao, G. Chen, and C. Chen, "Robust, efficient depth reconstruction with hierarchical confidence-based matching," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3331–3343, Jul. 2017.

[4] T. Xue, L. Qu, and B. Wu, "Matching and 3-D reconstruction of multibubbles based on virtual stereo vision," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 6, pp. 1639–1647, Jun. 2014.

[5] D. Li, Z. Liu, S. Lu, and L. Chang, "A robust 3-D abrasion diagnosis method of pantograph slipper based on stereo vision," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 11, pp. 9072–9086, Nov. 2020.

[6] P. Kaewtrakulpong et al., "PatchMatch: A randomized correspondence algorithm for structural image editing," in *Proc. ACM SIGGRAPH*, vol. 28, 2009.

[7] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 519–528.

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

[9] 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.

[10] 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.

[11] 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.

[12] Y. Ju, M. Jian, S. Guo, Y. Wang, H. Zhou, and J. Dong, "Incorporating Lambertian priors into surface normals measurement," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[13] Y. Ju, B. Shi, M. Jian, L. Qi, J. Dong, and K.-M. Lam, "NormAttention-PSN: A high-frequency region enhanced photometric stereo network with normalized attention," *Int. J. Comput. Vis.*, vol. 130, no. 12, pp. 3014–3034, Dec. 2022.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] E. Zheng, E. Dunn, V. Jojic, and J.-M. Frahm, "PatchMatch based joint view selection and depthmap estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1510–1517.

[20] 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.

[21] W. Zhao, S. Liu, Y. Wei, H. Guo, and Y.-J. Liu, "A confidence-based iterative solver of depths and surface normals for deep multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6168–6177.

[22] M. Bleyer, C. Rhemann, and C. Rother, "PatchMatch stereo–stereo matching with slanted support windows," in *Proc. Brit. Mach. Vis. Conf.*, 2011.

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

[24] 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. 3260–3269.

[25] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1996, pp. 358–363.

[26] Y. Xue et al., "MVSCRF: Learning multi-view stereo with conditional random fields," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4312–4321.

[27] 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.

[28] S. B. Kang, R. Szeliski, and J. Chai, "Handling occlusions in dense multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2001, p. 1.

[29] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "DeepMVS: Learning multi-view stereopsis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2821–2830.

[30] J. Zhang, Y. Yao, S. Li, Z. Luo, and T. Fang, "Visibility-aware multi-view stereo network," 2020, *arXiv:2008.07928*.

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

[32] C. Bailer, M. Finckh, and H. Lensch, "Scale robust multi view stereo," in *Proc. IEEE/Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2012, pp. 398–411.

[33] J. Wei, B. Resch, and H. Lensch, "Multi-view depth map estimation with cross-view consistency," in *Proc. Brit. Mach. Vis. Conf.*, 2014, pp. 1–7.

[34] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, "DeepPruner: Learning efficient stereo matching via differentiable PatchMatch," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4384–4393.

[35] S. Liu, S. D. Mello, J. Gu, M. H. Yang, and J. Kautz, "Learning affinity via a spatial propagation neural network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019.

[36] Q. Xu and W. Tao, "Planar prior assisted patchmatch multi-view stereo," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 1–22.

[37] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017.

[38] Q. Xu and W. Tao, "Learning inverse depth regression for multi-view stereo with correlation cost volume," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12508–12515.

[39] C. Chen, X. Chen, and H. Cheng, "On the over-smoothing problem of CNN based disparity estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8997–9005.

[40] 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.

[41] F. Wang, S. Galliani, C. Vogel, and M. Pollefeys, "IterMVS: Iterative probability estimation for efficient multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022.

[42] H. Yi, Z. Wei, M. Ding, R. Zhang, and Y. W. Tai, "Pyramid multi-view stereo net with self-adaptive view aggregation," in *Proc. IEEE/Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2020.

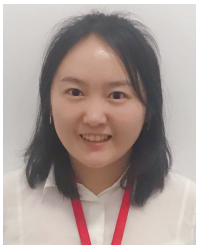
[43] Z. Yu and S. Gao, "Fast-MVSNet: Sparse-to-dense multi-view stereo with learned propagation and Gauss–Newton refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1949–1958.

[44] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4104–4113.

[45] K. Luo, T. Guan, L. Ju, H. Huang, and Y. Luo, "P-MVSNet: Learning patch-wise matching confidence aggregation for multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10452–10461.

[46] 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.

[47] S. Fuhrmann, F. Langguth, N. Moehrle, M. Waechter, and M. Goesele, "MVE—An image-based reconstruction environment," *Comput. Graph.*, vol. 53, pp. 44–53, Dec. 2015.



Yimei Liu received the B.Sc. degree from Xidian University, Xian, China, in 2013, and the M.E. degree from University Jean Monnet, Saint-Etienne, France, in 2015. She is currently pursuing the Ph.D. degree with the Ocean University of China, Qingdao, China.

Her research interests include computer vision and 3-D reconstruction.

Yuan Rao received the B.Sc. degree from the Ocean University of China, Qingdao, China, in 2017, and the M.E. degree from the Dalian University of Technology, Dalian, China, in 2019. He is currently pursuing the Ph.D. degree with the Ocean University of China.

His research interests include computer vision and autonomous robotics.

Eric Rigall received the degree in engineering from the Graduate School of Engineering, University of Nantes, Nantes, France, in 2018. He is currently pursuing the Ph.D. degree with the Vision Laboratory, Ocean University of China, Qingdao, China, supervised by Prof. Junyu Dong.

His research interests include radio frequency identification (RFID)-based positioning, image processing, machine learning, and computer vision.

Hao Fan received the B.Sc., M.E., and Ph.D. degrees from the Department of Computer Science and Technology, Ocean University of China, Qingdao, China, in 2012, 2014, and 2019, respectively.

He is currently a Lecturer in computer application technology with the Department of Computer Science and Technology, Ocean University of China. His research interests include computer vision, 3-D reconstruction, and underwater image processing.

Junyu Dong (Member, IEEE) received the B.Sc. and M.Sc. degrees from the Department of Applied Mathematics, Ocean University of China, Qingdao, China, in 1993 and 1999, respectively, and the Ph.D. degree in image processing from the Department of Computer Science, Heriot-Watt University, Edinburgh, U.K., in November 2003.

He joined the Ocean University of China, in 2004, where he is currently a Professor and the Head of the Department of Computer Science and Technology. His research interests include machine learning, big data, computer vision, and underwater image processing.