NR-MVSNet: Learning Multi-View Stereo Based on Normal Consistency and Depth Refinement

Jingliang Li, Zhengda Lu, Yiqun Wang, Jun Xiao, and Ying Wang

***Abstract*— Multi-view Stereo (MVS) aims to reconstruct a 3D point cloud model from multiple views. In recent years, learning-based MVS methods have received a lot of attention and achieved excellent performance compared with traditional methods. However, these methods still have apparent shortcomings, such as the accumulative error in the coarse-to-fine strategy and the inaccurate depth hypotheses based on the uniform sampling strategy. In this paper, we propose the NR-MVSNet, a coarse-to-fine structure with the depth hypotheses based on the normal consistency (DHNC) module, and the depth refinement with reliable attention (DRRA) module. Specifically, we design the DHNC module to generate more effective depth hypotheses, which collects the depth hypotheses from neighboring pixels with the same normals. As a result, the predicted depth can be smoother and more accurate, especially in texture-less and repetitive-texture regions. On the other hand, we update the initial depth map in the coarse stage by the DRRA module, which can combine attentional reference features and cost volume features to improve the depth estimation accuracy in the coarse stage and address the accumulative error problem. Finally, we conduct a series of experiments on the DTU, BlendedMVS, Tanks & Temples, and ETH3D datasets. The experimental results demonstrate the efficiency and robustness of our NR-MVSNet compared with the state-of-the-art methods. Our implementation is available at https://github.com/wdkyh/NR-MVSNet.**

***Index Terms*— Multi-view stereo, deep learning, 3D model reconstruction, normal consistency, feature attention, plane sweep algorithm.**

I. Introduction

MULTI-VIEW Stereo is a major problem for computer vision, which has a wide range of applications in the real-world such as autonomous driving, augmented reality (AR), and robotics. The goal of MVS is to reconstruct the point clouds from multi-view images with known camera parameters. Similar to the Structure-from-Motion method [1], the core task is to measure the similarity between corresponding image patches and reconstruct 3D point clouds.

Until now, learning-based methods achieved prominent results compared to previous traditional MVS methods. Inspired by the success of Convolutional Neural Network (CNN) in many computer vision fields, MVSNet [2] first proposed an end-to-end architecture to predict the depth maps, which construct the cost volume based on plane sweep stereo [3]. However, its memory demand grows cubically with the model resolution since the 3D CNN module being applied to infer the multi-view correspondence. Recent works [4], [5], [6], [7], [8], [9] employ a coarse-to-fine approach that can obtain more low-frequency components and high-frequency details of images while being GPU memory friendly.

In the MVS task, they expect to preserve the distinctions between adjacent views and form an efficient cost volume for image correspondence matching. However, the coarse stage features become smoothed and some details are lost after several poolings and convolutional filters. Therefore, tiny translations between neighbor views cannot be caught, especially for the texture-less, repetitive-texture areas. As shown in Fig. 1, the reconstruction results of CasMVSNet [5], AA-RMVSNet [10] lose a lot of details and appear holes in the texture-less area. Furthermore, the predicted depth map in the coarse stage is unreliable as the cost volume cannot reserve the distinctions in smooth and similar features between different views. Meanwhile, we found an accurate input depth map can generate effective depth hypotheses which make a significant impact on the final results. The predicted depth map from the coarse stage is the input of the subsequent stage. Thus, improving the accuracy of the depth map in the coarse stage is a key issue for this work to solve the accumulative error problem.

Moreover, these methods [8], [11] construct the depth hypotheses based on the depth consistency of pixels locating on the same plane. However, these pixels don't have the same depth in the camera coordinate since their plane is not necessarily parallel to the camera imaging plane. Thus, these methods, such as PatchmatchNet [8], have not achieved satisfactory results as shown in Fig. 1. On the other hand, some traditional methods [12], [13] utilize the normal of neighbors to calculate the depth value of a given pixel. However, their results are unsatisfactory and time-consuming due to their

Manuscript received 27 September 2022; revised 26 January 2023 and 2 April 2023; accepted 16 April 2023. Date of publication 5 May 2023; date of current version 11 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant U2003109, Grant U21A20515, Grant 62102393, Grant 62206263, Grant 62271467, and Grant 62202076; in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDA23090304; in part by the Youth Innovation Promotion Association of the Chinese Academy of Sciences under Grant Y201935; in part by the Natural Science Foundation of Chongqing under Grant CSTB2022NSCQ-MSX0924; in part by the State Key Laboratory of Robotics and Systems [Harbin Institute of Technology (HIT)] under Grant SKLRS-2022-KF-11; and in part by the Fundamental Research Funds for the Central Universities and China Postdoctoral Science Foundation under Grant 2022T150639 and Grant 2021M703162. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Junhui Hou. *(Jingliang Li and Zhengda Lu are co-first authors.) (Corresponding author: Jun Xiao.)*

Jingliang Li, Zhengda Lu, Jun Xiao, and Ying Wang are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: xiaojun@ucas.ac.cn).

Yiqun Wang is with the College of Computer Science, Chongqing University, Chongqing 400044, China.

Digital Object Identifier 10.1109/TIP.2023.3272170

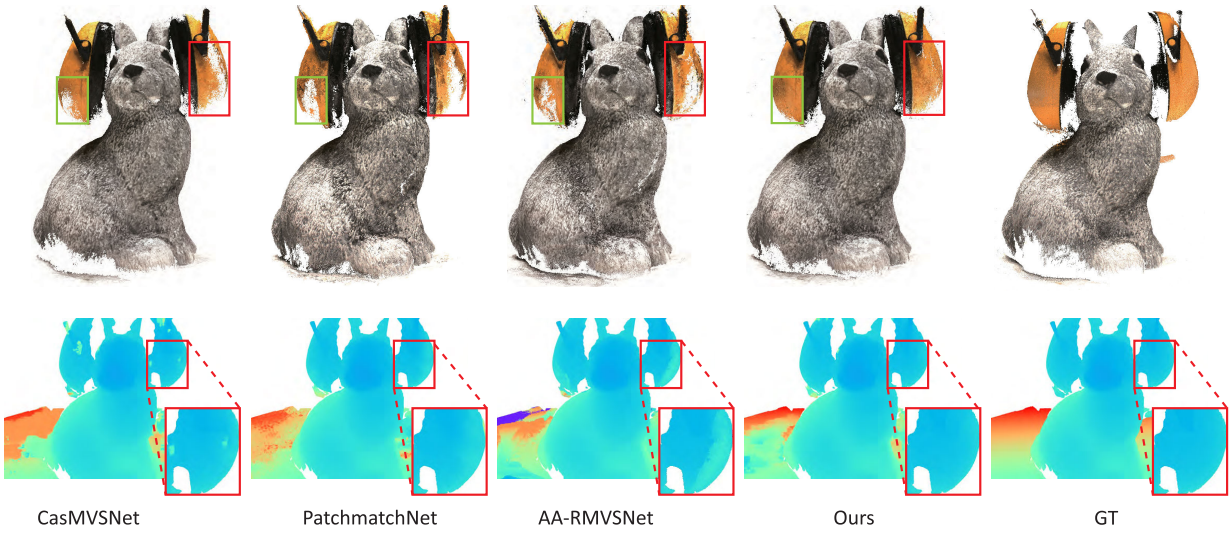

Fig. 1. Comparision with previous state-of-the-art learning-based methods [5], [8], [10] on scan33 of DTU [14]. The top row shows the point clouds, and the bottom row shows the corresponding depth maps. We show that our method can achieve smoother depth and a more complete reconstructed point cloud in the texture-less region.

uncertainty of initialization. Therefore, it is necessary to address how to collect more effective depth hypotheses and increase computational efficiency.

In this work, we propose the NR-MVSNet, a coarse-to-fine structure with the depth hypotheses based on the normal consistency (DHNC) module and the depth refinement with reliable attention (DRRA) module. As a consequence, the module could predict more accurate depth maps and reconstruct complete point clouds, especially in the texture-less and repetitive-texture areas, as shown in Fig. 1. Specifically, our DHNC module aims to find more effective depth hypotheses. For a reference pixel, we first adaptively propagate the locations of sampling neighbors, then calculate their normal to project the reference pixel to the tangent plane and generate the depth normal hypotheses. Finally, we merge them with the commonly used depth range uniform hypotheses to provide more effective depth hypotheses. On the other hand, we proposed the DRRA module applied in the coarse stage to improve the predicted depth accuracy by gathering the fusion of attentional reference features and cost volume features. First, we use the Transformer module to extract global relational features, and consider the occluded problem to omit the useless features. Meanwhile, we use the cost volume features attention branch to aggregate similar information of the source and reference features. Furthermore, we fuse the two attentional outputs and predict the residual depth value to refine the depth, which provides a more accurate depth map for the next stage.

In summary, the main contributions of this work are summarized as follows:

- We present a novel depth hypotheses module based on normal consistency which gathers the information from the neighbors and significantly improves the accuracy of depth hypotheses, especially in the texture-less and repetitive areas.
- We propose a depth refinement module that gathers the fusion of attentional reference features and cost volume features to improve the depth estimation accuracy in the coarse stage and solve the accumulative error problem.
- We perform the dense reconstruction experiments on MVS datasets including the DTU, Tanks & Temples, ETH3D, and BlendedMVS benchmark. Our method achieves competitive reconstruction results on these dataset.

II. Related Work

Our technique aims to improve the final depth estimation result of MVS with an effective strategy of depth hypotheses generation and attention method. We present a brief overview of the various techniques related to the MVS of traditional, learning-based, normal-based, and depth refinement methods.

A. Traditional Methods

In recent years, there has been an increasing amount of studies on Multi-view stereo. We refer to algorithms before the deep learning era as traditional MVS methods. Generally, traditional methods can be classified into four groups: voxel-based, surface-based, patch-based, and depth map-based methods. Voxel-based methods [15], [16] construct a fixed volume of an object or scene with millions of small voxels and then estimate if each voxel belongs to the surface. However, this representation requires high memory consumption when numerous voxels are not on the surface. Surface-based methods [17], [18] directly reconstruct a surface mesh and their results often look smoother than voxel-based approaches and lack detailed information. Patch-based [1], [12], [19] methods consider the surface as a collection of patches and first match these patches which contain more easily distinguished features, then propagate to the texture-less areas. These methods take advantage of spatial coherence but may lead to errors for texture-less or boundary areas. In contrast, the depth map [12], [13], [20], [21], [22] is the most flexible representation among all to estimate the depth of the object or scene and reconstruct point clouds through fusion methods. Similar to

these approaches, we also adopt the depth map representation. Though these works yield impressive results, they require photometric consistency and would achieve unsatisfactory matching results with hand-crafted features.

B. Learning-Based Methods

Deep CNN have achieved outstanding performance and significantly advanced the progress of various high-level vision tasks, e.g., object detection [23], [24], semantic segmentation [25], [26], and stereo matching [27]. Meanwhile, learning-based approaches have made impressive performance improvements for MVS vision tasks over traditional methods. MVSNet [2] first leveraged the plane sweep-based cost volume formulation followed by 3D CNN for regularization to predict the depth maps. After that, Learning-based MVS methods have developed rapidly. P-MVSNet [28] exploits a confidence metric and learns to aggregate it into a patch-wise matching confidence volume. However, 3D CNN in the MVSNet is time and memory-consuming. To handle the large-scale scenes, R-MVSNet [29] utilized a gated recurrent unit (GRU) [30] to replace 3D CNN. Fast-MVSNet proposed a sparse-to-dense framework for fast and accurate depth estimation. Recently, a novel coarse-to-fine strategy was widely used in the MVS task to reduce memory consumption. PVA-MVSNet [6] and CVP-MVSNet [4] form an image pyramid to construct a coarse-to-fine cost volume. CasMVSNet [5], UCS-Net [7] propose the cascade cost volumes based on a feature pyramid and estimate the depth maps in a coarse-to-fine manner. Predicting multi-view depth maps in a coarse-to-fine manner is a promising way to effectively reduce computational costs. UniMVSNet [9] designed a new loss function, which is more uniform and reasonable to combat the challenge of sample imbalance. TransMVSNet [31] used a Transformer to gather global context-aware information. There are also some other methods proposed to solve the MVS challenges, e.g. pixel-wise visibility [32], [33], [34], unreasonable depth hypotheses [35], [36] and inaccuracy on texture-less areas [8], [37].

C. Normal Based and Depth Refinement

In vision tasks, the normal map is a frequently utilized element, which can offer extra information. For traditional methods, Gipuma [12], Colmap [13] use the random plane parameters, depth, and normal for the initialization, then iteratively propagate and update them from neighboring pixels. The PatchMatch Stereo [38] also considers the slant planes problem using the normal of neighbors. However, their results are unsatisfactory because the fixed neighbors may also be inaccurate and time-consuming due to the uncertainty of initialization. For deep learning methods, Kusupati et al. [39] used an external branch to predict the normal map and construct the consistency between depth and normal to infer a more accurate depth map in the depth estimation task. These methods used a normal branch to supervise, which depends on the quality of the predicted normal and need additional parameters.

Depth refinement strategy exists long ago. MVSNet [2] and Cas-MVSNet [5] use a simple 2D CNN based refinement layers in the final stage. Distinct from these, we use a depth refinement module to predict a more accurate depth map in the coarse stage and help to calculate effective depth hypotheses. However, simple 2D CNN cannot extract the global features. MVSNet++ [40] uses a depth-based attention mechanism to generate smooth depth maps. LANet [37] proposed a long-range attention network to aggregate features which can guide more information for measuring similarity. EPP-MVSNet [35] designs an epipolar-assembling module. Inspired by these works, we proposed an attention-based branch to extract the global relational features and improve the accuracy of the depth map in the coarse stage.

III. Method

For the 3D reconstruction from a collection of multi-view images with known camera parameters, we propose the NR-MVSNet with the depth hypotheses based on normal consistency (DHNC) module and the depth refinement with reliable attention (DRRA) module. In this section, we first describe the overall architecture of the NR-MVSNet. Then we describe the details of our method including the DHNC module, the depth evaluation steps, and the DRRA module. Finally, we introduce the loss function used in the paper.

A. Network Architecture

Similarly with recently learning-based MVS networks, the architecture of our NR-MVSNet contains multi-stages with a coarse-to-fine strategy as illustrated in Fig. 2. Each stage has five common procedures including feature extraction, depth hypotheses generation, cost volume construction, regularization, and depth regression. Especially, we construct the DHNC module to generate more effective depth hypotheses. Meanwhile, we use the initial stage to obtain an initial depth map as the input of normal calculation in the first stage. In the initial stage, the learning parameters are share with the first stage and only use the depth range hypotheses generated by uniform sampling in the whole depth range. After that, we utilize the DRRA module to improve the accuracy of the initial depth map by gathering the fusion of attentional reference features and cost volume features.

More specifically, given the input of N images, we use I_1 and $\{I_i\}_{i=2}^{N}$ to denote the reference and the source images. First, we utilize a small FPN [41] to extract the multi-scale features $\{F_{i,k}\}_{k=1}^{3}$ at three resolutions for the input images. While the input image size is $H \times W \times C$, the resolutions of the corresponding features $\{F_{i,k}\}_{k=1}^{3}$ are $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{2} \times \frac{W}{2}$ and $H \times W$ respectively.

Then, we set the depth range hypotheses number as M_k^r and the depth normal hypotheses number as M_k^n in the k-th stage for the DHNC module. For the stages $\{k = 1, 2, 3\}$, we use the previous predicted depth map D_{k-1} to calculate the normal and generate the depth hypotheses $\{\tilde{\mathbf{D}}_k\}$. After that, we create the 3D cost volume by warping the source features to the reference view and then regularize it using 3D CNN, as shown in Fig. 2. Finally, we utilize a differentiable soft argmin process to regress the depth map D_k and reconstruct point clouds after depth estimation similar to MVSNet [2].

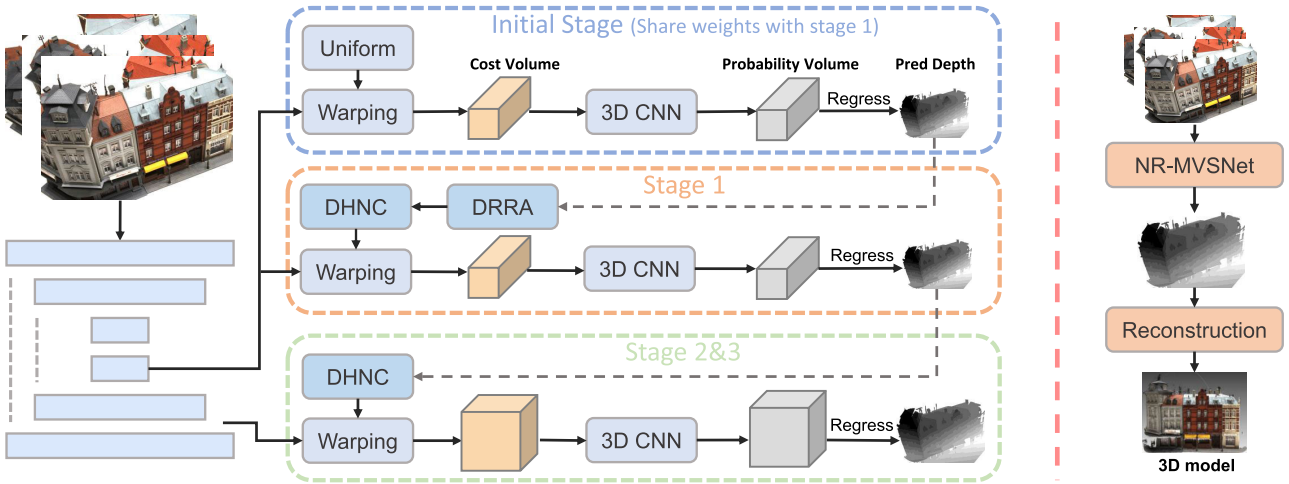

Fig. 2. The architecture of our NR-MVSNet. Reference and source images are first fed into a feature extraction module to form multi-scale features. Before the first stage, we use the initial stage to generate an initial depth map. The initial stage shares weights with the first stage. For the next stages, we then generate more accurate depth hypotheses using our DHNC module. After that, we build the cost, probability volume, and regress depth map. Especially, the initial depth map is fed into the DRRA module to update for a more accurate depth map. Finally, the predicted depth maps are used to reconstruct a 3D model with camera parameters.

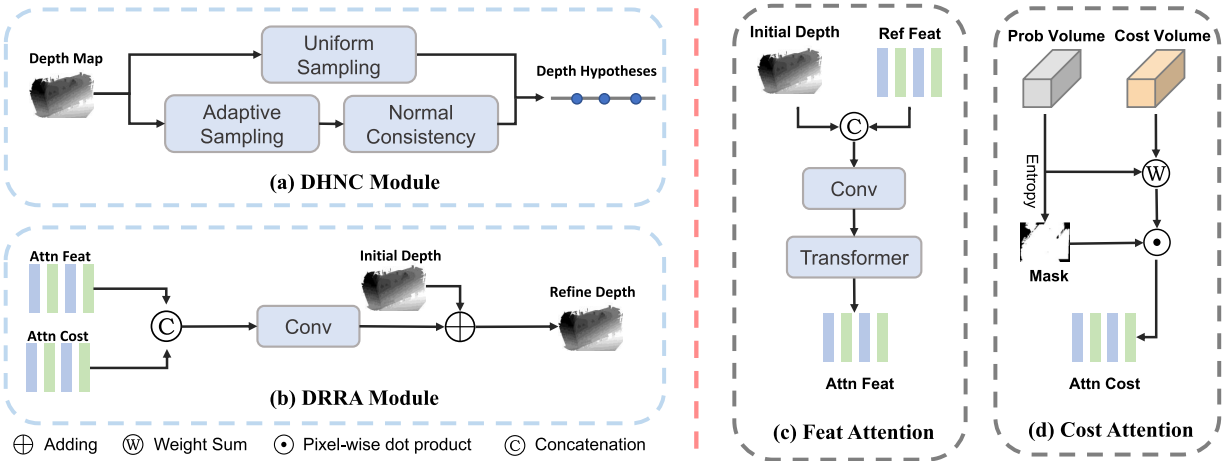

Fig. 3. Illustration of the submodules in our NR-MVSNet. (a) the DHNC module includes depth range hypotheses and depth normal hypotheses branch, (b) the DRRA module which fuses the attentional reference features and cost volume features, (c) the structure of attentional reference features, (d) the structure of attentional cost volume features.

B. Depth Hypotheses Based on Normal Consistency

Learning-based methods usually generate depth hypotheses using uniform sampling. However, the final predicted depth map may be less smooth because the correlation of neighboring pixels is not taken into consideration in this manner. To overcome such limitations, it is essential to provide more effective depth hypotheses through the depth hypotheses generation module. Some previous works [8], [11], [12] propagate the depth information by assuming the depths of pixels that lie on the same surface are consistent. However, pixels from the same physical plane do not all have the same depth value. Conversely, the normals of these pixels are consistent in the 3D space. Thus, we propose the DHNC module to leverage the normals of adaptive neighbors for more accurate depth hypotheses. The details of DHNC are shown in Alg. 1.

The pipeline of our DHNC module consists of two branches as shown in Fig. 3(a). In the first branch, we generate depth range hypotheses in the whole depth range. Given the previous depth prediction D_{k-1} at the pixel p, we dynamically adjust the depth range R_k based on the uncertainty estimation confidence [7]. The range R_k at pixel p and stage k is calculated as:

$$R_k^r = [D_{k-1} - \lambda U_{k-1}, D_{k-1} + \lambda U_{k-1}], \tag{1}$$

where λ is a scalar interval parameter, and U_{k-1} is an uncertainty value about the predicted depth D_{k-1} of the previous stage, which is learned from probability volume, according to [7].

Then, we calculate the depth range hypotheses $\tilde{\mathbf{D}}_k^r = \{\tilde{D}_{k,j}^r\}_{j=1}^{M_k^r}$ by uniform sampling in depth range R_k^r.

In the second branch, we find the depth normal hypotheses based on the normals of adjacent pixels. To gather the M_k^n depth normal hypotheses for pixel p in the reference image, we first calculate the normals of its neighboring pixels using depth value and camera parameters. However,

Algorithm 1 The Pesudo Code of DHNC at Stage k

Input: Predicted depth map D_{k-1} and uncertainty map U_{k-1} of previous stage, number of normal hypotheses M_k^n and range hypotheses M_k^r, scalar interval parameter λ.

Output: Depth hypotheses $\tilde{\mathbf{D}}_k$

```
Compute R_k according to Eq. (1)
Compute D̃_k^r by sampling uniformly in R_k
Obtain neighbors {p_j}_{j=1}^{M_k^n} of p_i by adaptively sampling.
for j = 1 to M_k^n do
    Compute normal n_j of p_j according to Eq. (2)
    Compute projection point p'_j by Eq. (3)
    Get p'_j depth value as D̃_{k,j}^n
end for
Compute D̃_k according to Eq. (4)
```

a static set of sampling neighbors lead to the effectless depth hypotheses since the pre-predicted depth may be inaccurate in some regions, especially for the texture-less areas, and some neighbors may cross the boundary. Therefore, we adaptively propagate the neighboring sampling locations based on the deformable convolution [42], similarly with PatchmatchNet [8], enabling us to effectively collect more suitable hypotheses, which also include the texture-less region.

For each pixel p, we first define a static neighborhood $\{p_j\}_{j=1}^{M_k^n}$, and then we adaptively update the position of the neighbors by learning a offset Δp_j, where $p_j := p_j + \Delta p_j$. Secondly, given the depth map of the reference image, we compute the 3D coordinates of the reference pixels from their 2D position relying on the reference camera intrinsic matrices. Then, we formulate the inference of pixel normal in 3D space following with [43]. The normal $\mathbf{n}$ of pixel p_j can be estimated in closed form as:

$$\mathbf{n} = \frac{\left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \mathbf{1}}{\left\| \left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \mathbf{1} \right\|_2}, \tag{2}$$

where $\mathbf{A}$ are the coordinates of the neighbors of pixel p_j. $\mathbf{1}$ is a vector with all 1 elements. Note that the surface normals calculation with a given depth map is a fix-weight method, without learnable parameters required.

Furthermore, we calculate the depth normal hypotheses of the pixel p based on the surface normals of its adaptive neighbors p_j, as shown in Fig. 4. Respectively, 3D coordinates and its normals of the neighbor's pixels p_j form the tangent planes, we project the 3D point of pixel p to these tangent planes and obtain new 3D points p'_j as follows:

$$p'_j = p + \mathbf{n_j} \frac{\mathbf{n_j} \cdot (p_j - p)}{\|\mathbf{n_j}\|_2}, \tag{3}$$

Finally, we use the projection point $\{p'_j\}$ and the reference camera intrinsic matrices to calculate the depth normal hypotheses $\tilde{\mathbf{D}}_k^n = \{\tilde{D}_{k,j}^n\}_{j=1}^{M_k^n}$.

The total depth hypotheses $\tilde{\mathbf{D}}$ can be computed as follows:

$$\tilde{\mathbf{D}}_k = \tilde{\mathbf{D}}_k^r \cup \tilde{\mathbf{D}}_k^n, \tag{4}$$

The total number of the depth hypotheses is $M_k = M_k^r + M_k^n$.

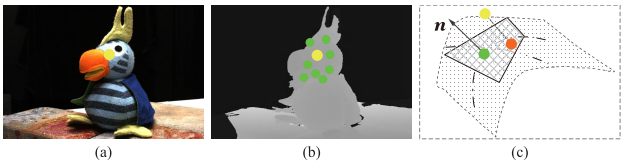

Fig. 4. Illustration of the DHNC. The yellow point represents the pixel location to be estimated. (a) reference image, (b) adaptive sampling locations (green), (c) point (yellow) projects to the tangent plane of its neighbor points (green) and obtain the projected point (orange) to calculate depth normal hypotheses.

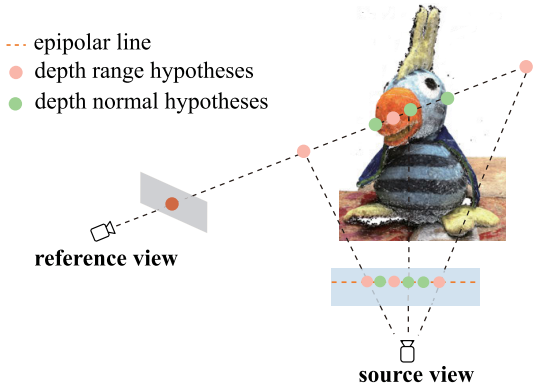

Fig. 5. Visualization of plane sweep. The point correspondence between reference and source images at different depth range hypotheses (pink points) and depth normal hypotheses (green). The dense interpolation of points along the epipolar line.

C. Coarse-to-Fine Depth Propagation

Similar to MVSNet [2], each stage in our NR-MVSNet also has three steps to evaluate the depth map, including cost volume construction, cost volume regularization, and depth regression. Note that, in this section, we omit the sub-indices denoting the stage in each parameter since these depth evaluation methods are identical in each stage.

First, we construct the cost volume by calculating the correlation between the reference and source features to measure their similarity. Specifically, We use the plane sweep stereo [3] to wrap the feature maps of source images into the fronto-parallel plane of reference view. As shown in Fig. 5, we compute the relationship between pixel p in the reference view and the corresponding pixel p_i in the i-th source view at hypothesized depth $\tilde{d}$ as follows:

$$p_i = K_i \cdot (R_1 \cdot (K_1^{-1} \cdot p \cdot \tilde{d}) + T_1), \tag{5}$$

where K_1, K_i denotes the intrinsic matrix of reference image I_1 and source image I_i, and R_1, T_1 denote the relative rotation and transformation parameters between the reference and source images.

After that, we obtain the relationship between the reference and source view pixels. Then, we gain the warped source feature maps of view i at depth $\tilde{d}$ via the differentiable bilinear interpolation. Next, we construct the cost volume $\mathbf{V}$ by the group-wise correlation [44] and aggregate over the views with a pixel-wise view weight [1] to estimate the visibility information.

Furthermore, the cost volume may contaminate noise and should be incorporated with smoothness constraint [2]. Thus,

Algorithm 2 The Pesudo Code of DRRA

Input: Initial depth map $\tilde{\mathbf{D}}$ and reference features F,cost volume $\mathbf{V}$ and probability $\mathbf{P}$.
Output: Updated depth map $\tilde{\mathbf{D}}^u$
1: Compute F^{coon} by Eq. (7)
2: Compute F^{attn} using our Transformer module
3: Obtain attention cost volume V' according to Eq. (8)
4: Entropy map $\mathbf{E}$ by Eq. (9)
5: Confidence map $\mathbf{U} = \mathbf{1} - \mathbf{E}$
6: Mask attention cost volume V^{attn} according to Eq. (10)
7: $\tilde{D}^u = \tilde{D} + \tilde{D}'$ using Eq. (11)

we also use a 3D U-Net which is commonly used in recent methods to process the cost volume at each stage and infer the probability volume $\mathbf{P}$, which aggregates epipolar information about similarity.

Finally, to solve the problem of training with the back-propagation, we use the generic approach in learning-based methods to weighted sum:

$$\mathbf{D}_k(p) = \sum_{j=1}^{M_k} \tilde{\mathbf{D}}_{k,j}(p) \cdot \mathbf{P}_{k,j}(p), \tag{6}$$

D. Depth Refinement With Reliable Attention

In the coarse stage, the coarse-level features cannot capture the small translations and deformations in neighbor views well, thus its cost volume is unreliable and the predicted depth map is inaccurate, which leads to a cumulative error in the subsequent stages. Besides, the accuracy of the predicted depth map is crucial for the normal calculation and remains to be improved in the coarse stage. Therefore, we propose the DRRA module after the initial stage as illustrated in Fig.3(b), which can gather the fusion of attentional reference features and cost volume features. The details of DRRA are shown in Alg. 2.

In the MVSNet, the global image structure is important for recalling more hard samples, such as pixels in texture-less and repetitive-texture surfaces. Therefore, we first calculate the attentional reference features to aggregate a large receptive field in the spatial domain. On the other hand, the probability volume represents the probability of the similarity in different hypotheses and also represents the importance of the cost volume in different depth hypotheses. Then they are helpful to regress a finer depth map and we calculate the attentional cost volume features to aggregate the similarity information of the source and reference features.

To get the attentional reference feature F^{attn}, we use a smaller transformer module (Mini-ViT [45]) to extract the global attentional features as illustrated in Fig.3(c). Compared with CNN, transformers [46] can more efficiently extract global relational features. However, transformer attention is expensive both in terms of memory and computational complexity, especially at higher resolutions. Thus, we utilize the smaller transformer module to obtain the global distributional information of the reference features.

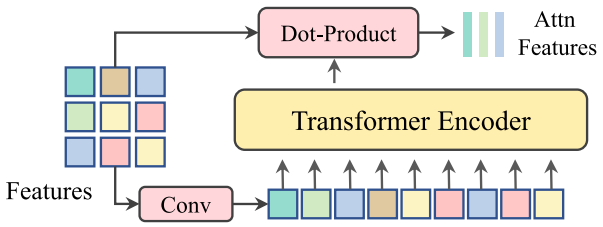

Fig. 6. Illustration of the Mini-ViT used in our method.

Let F and $\tilde{\mathbf{D}}$ be the reference features and the initial depth map, respectively. We first concatenate F and $\tilde{\mathbf{D}}$ and then use a convolutional layer to reduce the channels of the fused features F^{coon} and encode it as the input of the transformer, which is computed by

$$F^{coon} = f([F; \tilde{D}]), \tag{7}$$

where $[\cdot;\cdot]$ indicates the concatenation operation. $f(\cdot)$ is the convolution operation.

As shown in Fig. 6, the transformer takes a sequence of fixed size vectors as input, we first use a convolutional block with kernel size $k \times k$, stride k and number of output channels E, to encoder the fused features into a sequence in our transformer block. The result of this convolution is a tensor of size $H/k \times W/k \times E$. Then we reshape it into a spatially flattened tensor $F_s \in \mathbb{R}^{S \times E}$, where $S = (H \times W)/k^2$ serves as the effective input sequence length for the Transformer. Same as [47] and [48], we also add learned positional encodings to the patch embeddings before feeding them to the transformer. Then, the transformer block outputs a sequence of embedding features $F_e \in \mathbb{R}^{S \times E}$, which effectively contain more global information. Meanwhile, the fused features represent high-resolution and local pixel-level information. Therefore, we finally calculate the attention reference features F^{attn} by Dot-Product the fused features and the transformer outputs at the channel-dimension.

On the other hand, we introduce the cost volume attention approach, which can mine deeper for the similarity between reference and source views as shown in Fig. 3 (d). Let $\mathbf{P} \in \mathbb{R}^{M \times H \times W}$ be the probability and $\mathbf{V} \in \mathbb{R}^{C \times M \times H \times W}$ be the cost volume, where M is the number of the depth hypotheses, C indicates the number of channels of the similarity features. We first use element-wise summation between $\mathbf{V}$ and $\mathbf{P}$ to obtain the attentional cost volume features $\mathbf{V}' \in \mathbb{R}^{C \times H \times W}$, which can be computed by

$$\mathbf{V}'_{m,h,w} = \sum_{c=1}^{C} \mathbf{P}_{m,h,w} \odot \mathbf{V}_{c,m,h,w} \tag{8}$$

where $\odot$ denotes element-wise multiplication.

Furthermore, some visible pixels are occluded or in background, texture-less and repetitive-texture areas, their similar features in the cost volume are useless for the depth refinement. Thus, to measure the confidence of the similarity,

we compute the entropy map $\mathbf{E} \in \mathbb{R}^{H \times W}$ by

$$E_{h,w} = -\frac{1}{\log M} \sum_{m=1}^{M} \mathbf{P}_{m,h,w} \log(\mathbf{P}_{m,h,w}) \tag{9}$$

After that, we use $\log M$ to normalize the entropy to (0, 1], same as [49], and the confidence map $\mathbf{U}$ is computed by $\mathbf{U} = \mathbf{1} - \mathbf{E}$.

Note that, the similar feature with lower confidence scores may be wrong and we then incorporate a confidence threshold to dismiss the attention cost volume by

$$\mathbf{V}^{attn}_{c,h,w} = \begin{cases} \mathbf{V}'_{c,h,w} & \text{if } \mathbf{U}_{h,w} > \tau \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Here τ indicates the confidence threshold.

After obtaining the attentional reference features and cost volume features, we construct the high-dimensional features by fusing the two outputs and then use a convolutional layer to learn the residual depth values. The residual $\tilde{D}'$ are computed by

$$\tilde{D}' = f([F^{attn}; V^{attn}]), \tag{11}$$

E. Loss Function

We adopt a supervised learning strategy and construct the pyramid for ground truth depth as a supervisory signal. Similar to the existing MVSNet framework [2], we use the L_1 norm measuring to all stages and fuse them with different weights. The total loss can be defined as:

$$Loss = \sum_{k=1}^{N} \alpha_k \cdot L_k \tag{12}$$

where L_k refers to the loss at the k-th stage and α_k refers to its corresponding loss weight.

IV. Point-Cloud Reconstruction

The proposed NR-MVSNet estimates the depth maps of all views by taking each other as reference image. However, since not all the predicted depth values for each pixels are valid, it is necessary to first filter out outliers (i.e. background, occluded areas). Then, we apply a depth map fusion step to reconstruct point cloud, similar to other multi-view stereo methods [2], [12], [13].

A. Depth Map Filtering

First, we use the same depth map filtering strategy of MVSNet [2], including the photometric and geometric consistency. The photometric consistency measures the matching quality of different views. Given the probability volume $\mathbf{P}$ after 3D U-Net, we further transform the volume to a probability map P to measure the depth confidence. For each pixel, we first check the depth hypotheses value with the highest probability in probability volume, then add up its three adjacent probability values to form the probability map P. In our experiments, we regard pixels with probability lower than 0.8 as outliers and filter them.

The geometric constraint measures the depth consistency among multiple views. For a pixel p with predicted depth d in the reference view, we project it to pixel p' in another view by homography, and then reproject p' back to the reference view. If the reprojected coordinate p_{reproj} and the reprojected depth d_{reproj} satisfy $|p_{reproj} - p| < 1$ and $|d_{reproj} - d|/d < 0.01$, we regard the predicted depth d of pixel p is consistent. Finally, all depths should be consistent in at least three views.

B. Depth Map Fusion

By applying the above filtering mechanism, most of the outliers are removed and we can obtain clean depth maps. To further suppress reconstruction noises, we first determine the visible views for each pixel in the filitering step, and then take the average over all reprojected depths d_{reproj} as the final depth estimation. Finally, we directly reprojected the fused depth maps to generate the 3D point cloud.

V. Experiments

In this section, we first describe the details of our experiments, including the datasets, evaluation metrics, and implementation details. Then, we show the qualitative and quantitative comparisons of our network with state-of-the-art approaches. Finally, we perform ablation studies to validate the effectiveness of our proposed method on the DTU dataset.

A. Datasets and Evaluation Metrics

1) Datasets: We carry out experiments on DTU dataset [14] BlendedMVS [50], Tanks & Temples dataset [51] and ETH3D dataset [52]. These datasets are used for the performance evaluation and comparison of the proposed NR-MVSNet with our baseline method and existing state-of-the-art methods. The evaluation of Tank & Temples dataset and ETH3D dataset benchmark are conducted online by submitting generated point clouds to official websites. The DTU dataset is a large-scale indoor multi-view stereo dataset consisting of 124 different scenes from 49 or 64 views under 7 different lighting conditions with fixed camera trajectories, which are collected by a robot arm with a structured light scanner. We use the same training, validation, and evaluation sets as defined in MVSNet [2]. BlendedMVS dataset is a large-scale dataset for multi-view stereo and contains objects and scenes of varying complexity and scale. This dataset is split into 106 training scans and 7 validation scans. The Tanks & Temples dataset contains both indoor and outdoor scenes captured in more complex real scenarios, and it's divided into the intermediate and advanced sets. While the intermediate set contains 8 scenes with large-scale variations, the advanced set has 6 scenes. ETH3D contains 13 training and 12 test scenes, which include both indoor and outdoor challenging scenes.

2) Evaluation Metrics: The evaluation metrics on the four datasets are different. First, we adopt the accuracy and completeness of the distance metric to evaluate the quantitative results on the DTU dataset same as previous work [2], [14]. Among that, accuracy is measured as the distance from estimated point clouds to ground truth points, and completeness is measured as the distance from estimated point clouds to

the ground truth points. On the other hand, the F-score [51] is used as the evaluation metric to measure the accuracy and completeness of Tanks & Temples dataset and ETH3D dataset. Besides, we evaluate depth map estimations in DTU and BlendedMVS using depth-wise metric: EPE stands for L_1 distance between predicted depth map and ground truth, e1 and e3 represent the proportion of pixels with depth error larger than 1mm and 3mm.

B. Implementation Details

The baseline of our mehtods is CasMVSNet [5] with pixel-wise view weight [4]. We train the network on the DTU training set and fine-tune on the BlendedMVS dataset. As training, the resolution of the input image is 640×512, and the number of views N is 3. The numbers of depth range hypotheses and depth normal hypotheses are 48, 32, 8, and 16, 8, 0 for three stages, respectively. Before testing on Blended, Tanks & Temples and ETH3D benchmarks, we finetune our model on BlendedMVS for 10 epochs. We take 7 images as the input with the original size of 768 5̃76. The loss weights are set as $\alpha_1 = 1.0$, $\alpha_2 = 1.0$, $\alpha_3 = 1.0$ for three stages. And the loss weight for DRRA module is set as $\beta = 3.0$. The confidence threshold τ is set 0.3. We use the L_1 normal as the loss function to measure the absolute difference between the ground truth and the predicted depth on each stage. Finally, we implemented our network using Pytorch, utilize a batch size of 32 with four graphics cards, trained on Adam optimizer for 16 epochs with a learning rate of 0.001, and halved iteratively at the 10th, 12th, and 14th epochs.

All experiments were conducted on a server computer equipped with an Intel(R) Xeon(R) Gold 6130 CPU processor, 256GB of RAM, and eight NVIDIA TITAN RTX 24GB graphics cards.

C. Evaluation on DTU Dataset

In this section, we compare our proposed NR-MVSNet with the state-of-the-arts including both traditional and learning-based methods on the DTU dataset. Beside, we also apply our DHNC and DRRA modules to TransMVSNet [31] and UniMVSNet [9]. Note that, we use 5 neighboring views and set the image size to be 1152×864 in the testing set. The GPU memory only has a small increase compared with CasMVSNet [5] as shown in the ablation study section V-G.

We follow the evaluation metrics provided by the DTU dataset [14]. The quantitative results are shown in Tab. I. While we applied our proposed (DHNC and DRRA modules) to UniMVSNet [9] and TransMVSNet [31], these methods both achieve better results compared with their original methods. Meanwhile, TransMVSNet+Ours achieves the state-of-the-art result and our NR-MVSNet also achieves competitive performance.

We further demonstrate the quality of depth maps, which are the direct outputs by our NR-MVSNet and other methods CasMVSNet, PatchmatchNet [5], [8] on DTU dataset [14]. Since the DRRA module is used in the first stage, we not only evaluate the depth map in the final stage but also in the first and second stages. The quantitative results are illustrated in Tab. II,

TABLE I

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON DTU DATASET [14]. LOWER MEANS BETTER. (BOLD FIGURES INDICATE THE BEST, AND UNDERLINED FIGURES INDICATE THE IMPROVEMENT COMPARED WITH THE BASE METHOD

Method	Acc.	Comp.	Overall
Camp [20]	0.835	0.554	0.695
Furu [53]	0.613	0.941	0.777
Tola [21]	0.342	1.190	0.766
Gipuma [12]	**0.283**	0.873	0.578
SurfaceNet [54]	0.450	1.040	0.745
MVSNet [2]	0.396	0.527	0.462
R-MVSNet [29]	0.383	0.452	0.417
Point-MVSNet [55]	0.342	0.411	0.376
Fast-MVSNet [56]	0.336	0.403	0.370
CIDER [44]	0.417	0.437	0.427
CVP-MVSNet [4]	0.296	0.406	0.351
LANet [37]	0.320	0.349	0.335
UCS-Net [7]	0.338	0.349	0.344
CasMVSNet [5]	0.325	0.385	0.355
PatchmatchNet [8]	0.427	0.277	0.352
AA-RMVSNet [57]	0.376	0.339	0.357
UniMVSNet [9]	0.352	0.278	0.315
TransMVSNet [31]	0.321	0.289	0.305
Ours	0.331	0.285	0.308
UniMVSNet+Ours	0.332	**0.276**	0.304
TransMVSNet+Ours	0.316	0.278	**0.297**

TABLE II

QUANTITATIVE DEPTH RESULTS OF DIFFERENT METHODS ON DTU DATASET [14] (LOWER MEANS BETTER)

Stage	Method	EPE	$e1$	$e3$
1	CasMVSNet [5]	16.1	58.2	28.5
	PatchmatchNet [8]	26.3	66.9	34.4
	CasMVSNet + DRRA	15.6	50.3	25.6
	Ours	**15.3**	**49.4**	**25.0**
2	CasMVSNet [5]	15.2	41.3	23.3
	PatchmatchNet [8]	24.8	47.8	27.7
	CasMVSNet + DRRA	15.0	38.4	23.0
	Ours	**14.8**	**36.2**	**22.6**
3	CasMVSNet [5]	14.9	35.7	22.0
	PatchmatchNet [8]	24.7	43.3	24.9
	CasMVSNet + DRRA	14.6	34.8	21.5
	Ours	**14.5**	**32.8**	**21.3**

where EPE stands for L_1 distance, e1 and e3 represent the proportion of pixels with depth errors larger than 1mm and 3mm. First, since PatchmatchNet omits the 3D CNN regularization for the model lightweight, their depth maps are the worst in each stage. Then, the accuracy of the CasMVSNet result after adding the DRRA module is significantly improved. Finally, our NR-MVSNet achieves impressive results compared with other methods, which demonstrate its capability of yielding high-quality depth maps, especially in the first stage. Benefiting from the accuracy depth map in the first stage, more effective depth hypotheses are generated in the next stages. Therefore, our NR-MVSNet achieves competitive performance in each metric.

Fig. 7. Qualitative comparison of scan9 and scan49 in DTU dataset [14]. In each scan, the *top row:* generated point clouds of different methods and ground truth point cloud, and the *bottom row:* zoomed local region of red rectangle.

TABLE III

POINT CLOUD EVALUATION F-SCORES RESULTS ON THE INTERMEDIATE AND ADVANCED SUBSETS OF TANKS & TEMPLES DATASET [51].HIGHER SCORES ARE BETTER. THE **MEAN** IS THE AVERAGE SCORE OF ALL SCENES. (‡ DENOTES NO FINETUNING ON THE BLENDEDMVS [50] DATASET)

Method	Intermediate									Advanced						
	mean	Fam.	Fra.	Hor.	Lig.	M60	Pan.	Pla.	Tra.	**mean**	Aud.	Bal.	Cou.	Mus.	Pal.	Tem.
UCS-Net [7]‡	54.83	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	-	-	-	-	-	-	-
CasMVSNet [5]‡	56.42	76.36	58.45	46.20	55.53	56.11	54.02	58.17	46.56	31.12	19.81	38.46	29.10	43.87	27.36	28.11
PatchmatchNet [8]‡	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
AA-RMVSNet [57]	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
UniMVSNet [9]	64.36	**81.20**	**66.43**	53.11	63.46	**66.09**	**64.84**	62.23	57.53	38.96	**28.33**	44.36	**39.74**	**52.89**	33.80	34.63
TransMVSNet [31]	63.52	80.92	65.83	**56.94**	62.54	63.06	60.00	60.20	**58.67**	37.00	24.84	**44.59**	34.77	46.49	**34.69**	36.62
Ours	62.94	80.78	63.55	53.09	60.61	65.29	62.20	60.65	57.31	37.20	26.76	43.21	35.79	50.01	33.35	34.08
UniMVSNet+Ours	**65.17**	80.98	65.14	54.93	**68.70**	64.08	64.67	**65.60**	57.29	**39.43**	28.28	44.37	39.72	52.87	34.29	**37.06**
TransMVSNet+Ours	63.60	**81.20**	65.05	52.28	63.18	65.49	**64.84**	61.45	55.89	38.28	27.07	43.47	38.88	51.79	33.79	34.69

Fig. 7 shows the detailed visualization comparison results with CasMVSNet [5], PatchmatchNet [8] and AA-RMVSNet [10]. The details are highlighted in the red rectangles. In the first scan, we provides more complete 3D dense point clouds with preserved details, especially in texture-less area (see red rectangle). Meanwhile, we also show the advantage of our method in texture-less area as shown in Fig. 1. Moreover, our method can also keep more complete points and clearer texture structures for the complex scan with more details as shown in the second scan.

D. Evaluation on Tanks & Temples Dataset

Furthermore, we evaluate the generalization ability of our NR-MVSNet by fine-tuning the model trained on the DTU dataset [14] with the training set of the BlendedMVS dataset [50]. We reconstruct point clouds on the intermediate and advanced *Tanks & Temples* dataset [51] for comparison. The input image size is 1920×1024 and the number of views N is 7. Meanwhile, we use the camera parameters provided by MVSNet [2] as the input.

As shown in Tab. III, our NR-MVSNet method achieves competitive mean F-score in both *Intermediate* and *Advanced* subset. Meanwhile, we also compared the results of UniMVSNet and TransMVSNet applied with our proposed modules on *Tanks & Temples* dataset. Similarly, our modules are still effective for improving the results, although these two methods have achieved competitive results.

As shown in Fig. 8, we reconstruct some point clouds of Tanks & Temples dataset to demonstrate the quality of the reconstruction and the effectiveness of our method. Although our F-score is lower than CasMVSNet on the 'Panther' scene, we generate more points in the plane as shown in Fig.8(e).

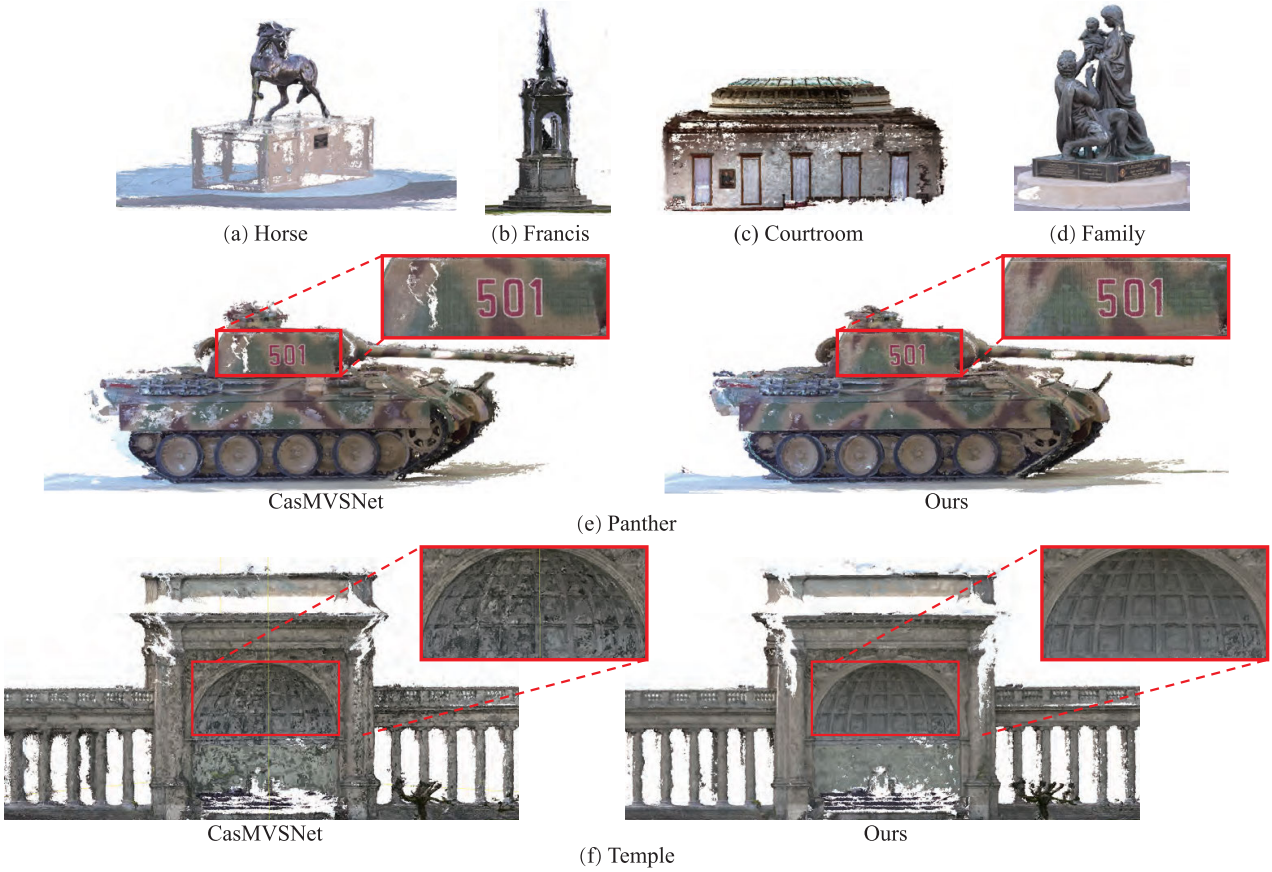

Fig. 8. Point cloud results of NR-MVSNet on the intermediate set of Tanks & Temples dataset [51]. In the first row, we show the reconstructed point clouds of our method. In the second and the third row, we compared the reconstructed results with CasMVSNet [5]. No fine-tuning for Ours and CasMVSNet.

On the other hand, we achieve the best F-score and preserve more details for the 'Temple' scene which includes lots of repetitive-texture regions (see Fig. 8 (f)).

E. Evaluation on ETH3D

In this section, we use the model finetuned on BlendedMVS to evaluate the generalization ability of our NR-MVSNet on the ETH3D dataset, the same as evaluation on Tanks & Templs dataset. The results of EPP-MVSNet and IterMVS are referenced from their papers, and other methods' results are evaluated by their pre-trained models. Furthermore, we found the reconstructed point clouds are parameter sensitive. We set the input image size to 1920 × 1280 and use the same fusion parameters, the number of views N to 7. The camera parameters and depth ranges are estimated with Colmap [1].

Quantitative results of ETH3D are shown in Tab. IV. Compared with other methods, our method achieves competitive results. Meanwhile, our proposed modules can also improve the reconstruction results of the recent state-of-the-art methods, by applying them to these methods. Especially for the TransMVSNet, our modules bring remarkable improvements at each metric and achieves best results at *Accuracy* on Training and Test dataset.

TABLE IV

QUANTITATIVE DEPTH RESULTS OF DIFFERENT METHODS ON ETH3D DATASET [52]. HIGHER MEANS BETTER

Method	Training			Test		
	F1	AC	CO	F1	AC	CO
PatchmatchNet [38]‡	64.21	64.81	65.43	73.12	69.71	77.46
EPP-MVSNet [35]	**74.00**	82.76	**67.58**	**83.40**	85.47	**81.79**
IterMVS [58]	71.69	79.79	66.08	80.06	84.73	76.49
UniMVSNet [9]	69.53	84.95	60.62	81.60	86.87	77.58
TransMVSNet [31]	64.74	79.05	57.83	75.39	84.02	70.07
Ours	68.79	82.89	61.08	80.23	86.00	76.05
UniMVSNet+Ours	70.59	84.62	62.30	81.88	86.84	78.15
TransMVSNet+Ours	69.25	**85.23**	60.38	80.84	**87.86**	75.86

F. Evaluation on BlendedMVS Dataset

Previous datasets apply evaluation metrics toward point clouds. Same as TransMVSNet [31], we further demonstrate the quality of depth maps, which are the direct outputs by networks, on the BlendedMVS validation dataset [50]. We set N = 5 and image resolution as 512 640 and apply the depth evaluation metrics where depth values are normalized to make depth maps with different depth ranges comparable.

Some quantitative results are illustrated in Tab. V. Compared with other methods, our NR-MVSNet achieves impressive results and also improves the results of the-state-of-art methods, which demonstrate the capability of yielding high-quality depth maps for our method.

TABLE V
QUANTITATIVE RESULTS TOWARDS PREDICTED DEPTH MAPS ON BLENDEDMVS VALIDATION DATASET [50]. LOWER MEANS BETTER

Method	EPE	$e1$	$e3$
CVP-MVSNet [4]‡	1.90	19.73	10.24
CasMVSNet [5]‡	1.43	19.01	9.77
EPP-MVSNet [35]	1.17	12.66	6.20
UniMVSNet [9]	0.94	6.89	3.29
TransMVSNet [31]	0.73	8.32	3.62
Ours	0.85	8.47	4.01
UniMVSNet+Ours	0.82	**6.27**	3.30
TransMVSNet+Ours	**0.71**	7.22	**2.43**

TABLE VI
COMPARISON OF DIFFERENT DESIGN CHOICES IN OUR NR-MVSNET STRUCTURE. OUR FINAL RESULTS IN A SIGNIFICANT BOOST IN PERFORMANCE. LOWER MEANS BETTER

Method	ACC.	Comp.	Overall
Baseline	0.364	0.336	0.350
Baseline+DHNC	0.348	0.314	0.332
Baseline+DRRA	0.339	0.298	0.319
Baseline+DHNC+DRRA (Ours)	**0.331**	**0.285**	**0.308**

TABLE VII
COMPARISON OF DIFFERENT BACKBONES WITH OUR DHNC AND DRRA MODULES ON OVERALL METRIC. LOWER MEANS BETTER

Method	Baseline	add DHNC	add DRRA	add Both
UCS-Net [7]	**0.344**	0.338	0.332	0.310
CasMVSNet [5]	0.355	0.341	0.328	0.315
PatchmatchNet [8]	0.352	0.346	0.322	0.321
PatchmatchNet w/o AP [8]	0.361	0.342	0.335	0.318
Ours	0.350	**0.332**	**0.319**	**0.308**

G. Ablation Study

To further demonstrate the contribution of our NR-MVSNet, we introduce some groups of ablation studies on the DTU dataset in this section.

1) Beneit of Each Component: Here, four experiments are executed to illustrate the improvement brought by each component of our method, as shown in Tab. VI. First, we remove the DHNC and DRRA modules in our NR-MVSNet as the baseline model. Then, we add these two modules to the baseline respectively for comparison. Finally, we achieved the best results than the baseline in each indicator after adding the DHNC and DRRA module.

Furthermore, we also apply our DHNC and DRRA modules to other learning-based coarse-to-fine methods. As shown in Tab. VII, these methods both achieve a significant improvement on *Overall* metric with our proposed modules. Notable, our DHNC modules provide less increase in PatchmatchNet [8] compared with other methods. First, PatchmatchNet also used deformable convolution to find the position of neighboring pixels (AP). Meanwhile, the module of AP conducts the depth hypotheses by assuming the depths of neighboring pixels that lie on the same surface are consistent. However, neighboring pixels don't have the same depth as the center pixel in the camera coordinate since their plane is not necessarily parallel to the camera imaging plane. Therefore, their AP module may conflict with our DHNC module and we replace AP with our DHNC module, which turns out that the DHNC module is more effective (0.352 → 0.342).

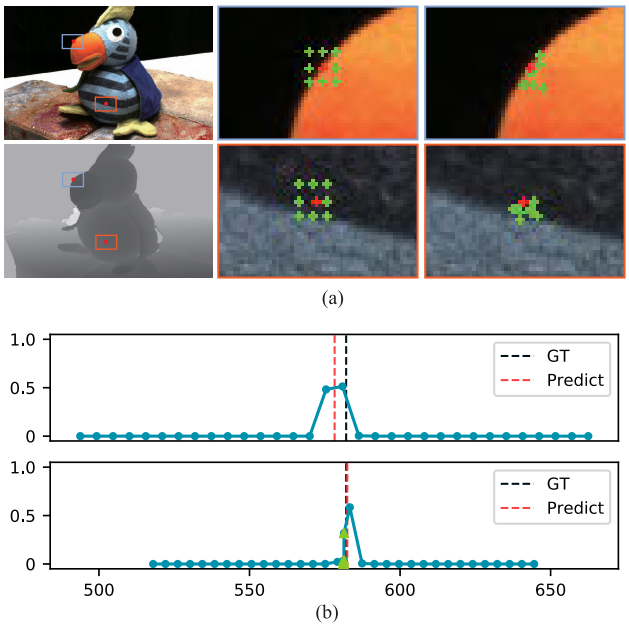

Fig. 9. (a) Visualization of adaptive sampling for two typical areas: object boundary (blue box) and texture-less (read box). The middle column shows the eight neighbors and the right column shows the adaptive locations. (b) Depth hypotheses comparison between CasMVSNet and ours on the second stage. The ordinate represents the predicted probability volume corresponding to the depth hypotheses value. Green triangles and cyan circles correspond to the depth normal and range hypotheses. *Top row:* CasMVSNet, only has depth range hypotheses [5], *Bottom row:* our NR-MVSNet, has depth range hypotheses and depth normal hypotheses.

As shown in Fig. 9(a), we visualize the locations of the depth normal hypotheses sampling in two typical situations. For the pixels at the object boundary or in the textureless region, the sampling points tend to the locations which have salient features and lie on the same surface. Beside, our DHNC module can generate more effective depth hypotheses as shown in Fig. 9(b) compared with the backbone method, CasMVSNet [5].

In addition, our DRRA module aims to refine the initial depth using the fusing features. Fig. 10 shows the predicted depth, corresponding normal and depth error map, before and after the DRRA module. Our DRRA module can better refine the depth map and the error map also indicates the refined depth is closer to the real depth. Furthermore, we can calculate a more smooth normal map benefiting from the high-quality depth map. Finally, our method has a more accurate depth and normal map for both the textured and the texture-less regions (see the front and top area of the input image in Fig. 10).

2) Number of Depth Hypotheses in DHNC: Furthermore, we also evaluate the influence of the different number of depth range hypotheses and depth normal hypotheses in each stage.

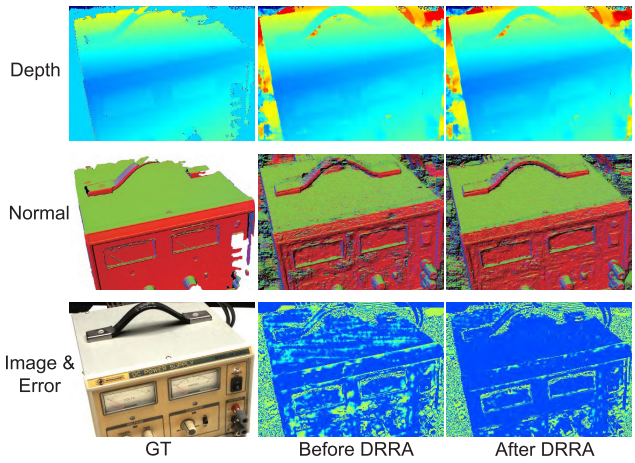

Fig. 10. Visualization of depth, normal and depth error map at before and after refinement on first stage. *Top row*: Depth map for ground truth, before DRRA, and after DRRA module. *Middle row*: Normal maps of the corresponding depths. *Bottom row*: Reference image and depth error maps between ground truth depth and predicted depths.

TABLE VIII

ABLATION STUDY OF DIFFERENT DEPTH HYPOTHESES NUMBER IN EACH STAGE.'RANGE' AND 'NORMAL' INDICATE THE DEPTH RANGE HYPOTHESES AND DEPTH NORMAL HYPOTHESES.THE NUMERICAL ORDER REPRESENTS 1, 2, AND 3 STAGES RESPECTIVELY.THE * INDICATES CORRESPONDING PARAMETER SETTINGS IN OUR PAPER

Range	Normal	Acc.	Comp.	Overall
40,24,8	8,4,0	0.343	0.312	0.328
48,24,8	8,4,0	0.339	0.311	0.325
48,32,8	8,4,0	0.336	0.295	0.316
40,24,8	16,8,0	0.338	0.310	0.324
48,24,8	16,8,0	0.335	0.302	0.319
* 48,32,8	16,8,0	0.331	**0.285**	0.308
40,24,8	16,8,8	0.338	0.309	0.324
48,24,8	16,8,8	0.332	0.305	0.319
48,32,8	16,8,8	**0.328**	0.286	**0.307**

Summarized in Tab. VIII, more depth range hypotheses and depth normal hypotheses number both bring the improvement on accuracy or completeness. Meanwhile, we also achieve the best quantitative results even if the total depth hypotheses number we used is less than or equal to other methods, such Cas-MVSNet (48,32,8), UCS-Net (64,32,8). Furthermore, we also add an ablation study with 8 normal hypotheses in the third stage. Due to the accuracy predicted depth map in the second stage, the result does not have quite improvement with 8 normal hypotheses in the third stage. Thus, we set the number of depth normal hypotheses in the third stage as 0.

3) Different Attention Approaches in DRRA: To quantitatively measure the effectiveness of the fusion features in the DRRA module, we first apply the reference features attention approach to refine the depth map, and then use the cost volume attention approach. Finally, we test both attention approaches. All of the experiments are based on the baseline network, without the DHNC module. As shown in Tab. IX, both attentional reference features and attentional cost volume features play a key role to extract the global distributional and local structural information and improve the accuracy of the predicted depth map.

TABLE IX

ABLATION STUDY OF THE DIFFERENCE ATTENTION APPROACHES IN DRRA MODULE

Method	Feat Attn	Cost Attn	ACC.	Comp.	Overall
1			0.364	0.336	0.350
2	✓		0.339	0.312	0.326
3		✓	0.342	0.320	0.331
4	✓	✓	**0.339**	**0.298**	**0.319**

TABLE X

COMPARISON OF THE RUNNING TIME (S PER VIEW) AND MEMORY CONSUMPTION BETWEEN OUR NR-MVSNET AND OTHER SOTA LEARNING-BASED METHODS

Method	Time(s).	Mem.(MB)
MVSNet [2]	1.210	10823
CasMVSNet [5]	0.492	5345
PatchmatchNet [8]	0.160	1629
AA-RMVSNet [57]	46.83	4897
Baseline+DRRA	0.55	5603
Baseline+DHNC	0.52	5547
Baseline+DHNC+DRRA (Ours)	0.57	5649

4) Memory and Run-Time Comparison: The computational cost of NR-MVSNet is compared with the mentioned learning-based methods by competing for memory consumption and run-time on DTU dataset. For a fair comparison, we use a fixed input size of 1152×864 and set the number of views to 5 on the evaluation set. As shown in Tab. X, our method has no increase in the complexity obviously comparable with CasMvsNet, our baseline method. Our DRRA only uses 102MB GPU memories and the DHNC module use 46MB GPU memories.

VI. CONCLUSION

In this work, we propose the NR-MVSNet, a coarse-to-fine structure, to reconstruct a 3D point cloud model from multiple views. We first present the DHNC module with the adaptively normal sampling to effectively collect more promising depth hypotheses and then propose the DRRA module to provide a more accurate depth map for the next stage. Finally, we construct the NR-MVSNet to generate more accurate depth estimation, especially in texture-less areas, which outperforms the SOTA approaches in quantitative metrics and visual effects. Thorough ablation studies demonstrate the benefits of the proposed DHNC and DRRA module for MVS. However, due to the difficulty of obtaining depth and point clouds, we will pay more attention to the self-supervised learning of MVS tasks in future work.

REFERENCES

[1] 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.

[2] 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. (ECCV)*, 2018, pp. 767–783.

[3] 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.

[4] 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.

[5] 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.

[6] H. Yi et al., "Pyramid multi-view stereo net with self-adaptive view aggregation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 766–782.

[7] 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.

[8] 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.

[9] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8645–8654.

[10] H. Xu and J. Zhang, "AANet: Adaptive aggregation network for efficient stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1959–1968.

[11] 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.

[12] 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.

[13] 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.

[14] 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. 10, pp. 1–16, Jan. 2016.

[15] S. N. Sinha, P. Mordohai, and M. Pollefeys, "Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[16] A. O. Ulusoy, M. J. Black, and A. Geiger, "Semantic multi-view stereo: Jointly estimating objects and voxels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4531–4540.

[17] Y. Furukawa and J. Ponce, "Carved visual hulls for image-based modeling," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2006, pp. 564–577.

[18] Z. Li, K. Wang, W. Zuo, D. Meng, and L. Zhang, "Detail-preserving and content-aware variational multi-view stereo reconstruction," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 864–877, Feb. 2016.

[19] 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.

[20] 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.

[21] 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.

[22] S. Bing Kang, R. Szeliski, and J. Chai, "Handling occlusions in dense multi-view stereo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2001, pp. 1–6.

[23] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[26] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. ICCV*, 2017, pp. 2961–2969.

[27] A. Kendall et al., "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 66–75.

[28] 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.

[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] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[31] Y. Ding et al., "TransMVSNet: Global context-aware multi-view stereo network with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8585–8594.

[32] 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. 1–10.

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

[34] 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. 1–4.

[35] 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.

[36] P. Yi, S. Tang, and J. Yao, "DDR-Net: Learning multi-stage multi-view stereo with dynamic depth range," 2021, *arXiv:2103.14275*.

[37] X. Zhang, Y. Hu, H. Wang, X. Cao, and B. Zhang, "Long-range attention network for multi-view stereo," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3782–3791.

[38] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo-stereo matching with slanted support windows," in *Proc. BMVC*, vol. 11, 2011, pp. 1–11.

[39] U. Kusupati, S. Cheng, R. Chen, and H. Su, "Normal assisted stereo depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2189–2199.

[40] P.-H. Chen, H.-C. Yang, K.-W. Chen, and Y.-S. Chen, "MVSNet++: Learning depth-based attention pyramid features for multi-view stereo," *IEEE Trans. Image Process.*, vol. 29, pp. 7261–7273, 2020.

[41] 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, pp. 2117–2125.

[42] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[43] D. F. Fouhey, A. Gupta, and M. Hebert, "Data-driven 3D primitives for single image understanding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3392–3399.

[44] Q. Xu and W. Tao, "Learning inverse depth regression for multi-view stereo with correlation cost volume," in *Proc. AAAI*, 2020, pp. 12508–12515.

[45] S. Farooq Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4009–4018.

[46] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–12.

[47] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[48] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[49] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab., Contrib. Theory Statist.*, vol. 1, 1961, pp. 547–562.

[50] 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.
[51] A. Knapitsch et al., "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, Jul. 2017.
[52] 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.
[53] 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. 2010.
[54] 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.
[55] 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.
[56] 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.
[57] 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.
[58] 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, pp. 8606–8615.

Jingliang Li received the bachelor's degree from Xidian University, Xi'an, China, in 2017, and the master's degree from the School of Artificial Intelligence, Xidian University, in 2020. He is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences, Beijing, China. His current research interests include 3D reconstruction and neural rendering.

Zhengda Lu received the bachelor's degree from Northwestern Polytechnical University, Xi'an, China, in 2016, and the Ph.D. degree from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2021. He is currently a Postdoctoral Researcher with the School of Artificial Intelligence, UCAS. His research interests include computer graphics, computer vision, and 3D reconstruction.

Yiqun Wang received the dual Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, and the University of Chinese Academy of Sciences. He is currently an Associate Professor with the College of Computer Science, Chongqing University. Before that, he was a Postdoctoral Research Fellow with the King Abdullah University of Science and Technology. His research interests include computer graphics and 3D vision.

Jun Xiao received the Ph.D. degree in communication and information system from the Graduate University of the Chinese Academy of Sciences, Beijing, in 2008. He is currently a Professor with the University of Chinese Academy of Sciences, Beijing. His research interests include computer graphics, computer vision, image processing, and 3D reconstruction. He is a Senior Member of CCF.

Ying Wang received the Ph.D. degree from the Beijing Institute of Technology in 1996. She is currently a Professor with the University of Chinese Academy of Sciences, Beijing. Her research interests include computer graphics, computer vision, and engineering computing.