RayMVSNet++: Learning Ray-Based 1D Implicit Fields for Accurate Multi-View Stereo

Yifei Shi, *Member, IEEE*, Junhua Xi, Dewen Hu, *Senior Member, IEEE*, Zhiping Cai, *Member, IEEE*, and Kai Xu, *Senior Member, IEEE*

***Abstract*—Learning-based multi-view stereo (MVS) has by far centered around 3D convolution on cost volumes. Due to the high computation and memory consumption of 3D CNN, the resolution of output depth is often considerably limited. Different from most existing works dedicated to adaptive refinement of cost volumes, we opt to directly optimize the depth value along each camera ray, mimicking the range (depth) finding of a laser scanner. This reduces the MVS problem to ray-based depth optimization which is much more light-weight than full cost volume optimization. In particular, we propose RayMVSNet which learns sequential prediction of a 1D implicit field along each camera ray with the zero-crossing point indicating scene depth. This sequential modeling, conducted based on transformer features, essentially learns the epipolar line search in traditional multi-view stereo. We devise a multi-task learning for better optimization convergence and depth accuracy. We found the monotonicity property of the SDFs along each ray greatly benefits the depth estimation. Our method ranks top on both the DTU and the Tanks & Temples datasets over all previous learning-based methods, achieving an overall reconstruction score of 0.33 mm on DTU and an F-score of 59.48% on Tanks & Temples. It is able to produce high-quality depth estimation and point cloud reconstruction in challenging scenarios such as objects/scenes with non-textured surface, severe occlusion, and highly varying depth range. Further, we propose RayMVSNet++ to enhance contextual feature aggregation for each ray through designing an attentional gating unit to select semantically relevant neighboring rays within the local frustum around that ray. This improves the performance on datasets with more challenging examples (e.g., low-quality images caused by poor lighting conditions or motion blur). RayMVSNet++ achieves state-of-the-art performance on the ScanNet dataset. In particular, it attains an AbsRel of 0.058m and produces accurate results on the two subsets of textureless regions and large depth variation.**

***Index Terms*—Multi-view stereo, implicit fields, deep neural networks.**

Manuscript received 30 September 2022; revised 23 March 2023; accepted 6 July 2023. Date of publication 17 July 2023; date of current version 3 October 2023. This work was supported in part by the National Key Research and Development Program of China under Grants 2018AAA0102200 and 2018YFB1305100, in part by the NSFC under Grants 62325211, 62132021, and 62002379, and in part by the Natural Science Foundation of Hunan Province of China under Grants 2023JJ20051 and 2023JJ20048. Recommended for acceptance by R. P. Wildes. *(Yifei Shi and Junhua Xi contributed equally to this work.) (Corresponding authors: Dewen Hu; Kai Xu.)*

Yifei Shi and Dewen Hu are with the College of Intelligence Science and Technology, National University of Defense Technology, Changsha, Hunan 410073, China (e-mail: yifei.j.shi@gmail.com; dwhu@nudt.edu.cn).

Junhua Xi, Zhiping Cai, and Kai Xu are with the College of Computer Science, National University of Defense Technology, Changsha, Hunan 410073, China (e-mail: hjh17@nudt.edu.cn; zpcai@nudt.edu.cn; kevin.kai.xu@gmail.com).

Digital Object Identifier 10.1109/TPAMI.2023.3296163

I. Introduction

LEARNING-BASED multi-view stereo has gained a surge of attention since the seminal work of MVSNet [74]. The core idea of MVSNet and many followup works is to construct a 3D cost volume in the frustum of the reference view through warping the image features of several source views onto a set of fronto-parallel sweeping planes at hypothesized depths. 3D convolutions are then conducted on the cost volume to extract 3D geometric features and regress the final depth map of the reference view.

Existing methods are often limited to low-resolution cost volume since 3D CNN is both computation and memory consuming. Several recent works proposed to upsample or refine cost volume aiming at increasing the resolution of output depth maps [10], [19], [72]. Such refinement, however, still needs to trade off between depth and spatial (image) resolutions. For example, CasMVSNet [19] opts to narrow down the range of depth hypothesis to allow high-res depth estimation, matching the spatial resolution of input RGB. 3D convolution is then confined within the narrow band, thus degrading the efficacy of 3D feature learning.

In fact, depth map is view-dependent although cost volume is not. Since the target is depth map, refining the cost volume seems neither economic nor necessary. There could be a large portion of the cost volume invisible to the view point. We advocate direct optimization of the depth value along each camera ray, mimicking the range (depth) finding of a laser scanner. This allows us to reduce the MVS problem to a ray-based depth optimization one which is, individually, a much more light-weight task than full cost volume optimization. We formulate the "range finding" of each camera ray as learning a 1D implicit field along the ray whose zero-crossing point indicates the scene depth along that ray (Fig. 1, top row). To achieve that, we propose RayMVSNet which learns sequential modeling of multi-view features along camera rays based on recurrent neural networks.

Technically, RayMVSNet contains two critical designs to facilitate learning accurate ray-based 1D implicit fields. *First*, the sequential prediction of 1D implicit field along a camera ray is essentially conducting an epipolar line search [2] with cross-view feature matching whose optimum corresponds to the point of ray-surface intersection. To learn this line search, we propose *Epipolar Transformer*. Given a camera ray of the reference view, it learns the matching correlation of the pixel-wise 2D features of each source view based on attention mechanism.

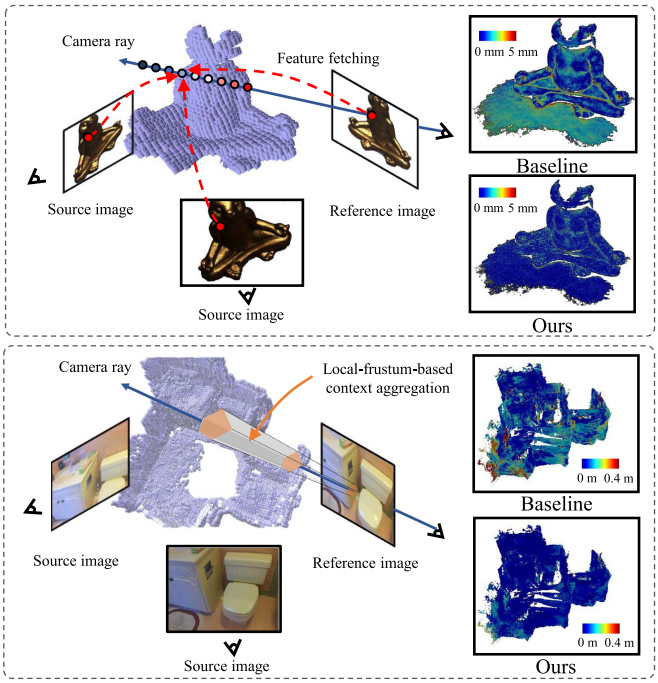

Fig. 1. RayMVSNet performs multi-view stereo via predicting 1D implicit fields on a camera ray basis. Top: The sequential prediction of 1D field is light-weight and the monotonicity of ray-based distance field around surface-crossing points facilitates robust learning, leading to more accurate depth estimation than the purely cost-volume-based baselines such as MVSNet [74]. Bottom: By aggregating more contextual feature with an extra local frustum-based attentional gating unit, RayMVSNet++ is able to achieve more accurate and robust depth predictions in challenging scenarios, such as poor lighting conditions or motion blur.

The transformer features of all views, together with (low-res) cost volume features, are then concatenated and fed into an LSTM [22] for implicit field regression. Fig. 3 visualizes how epipolar transformer selects reliable matching features from different views.

Second, we confine the sequential modeling for each camera ray within a fixed-length range centered around the hypothesized surface-crossing point given by the vanilla MVSNet. This makes the output 1D implicit field along each ray monotonous, which is normalized to $[-1, 1]$. Such restriction and normalization lead to significant reduction of learning complexity and improvement of result quality. We devise two learning tasks: 1) sequential prediction of signed distance at a sequence of points sampled in the fixed-length range and 2) regression of the zero-crossing position on the ray. A carefully designed loss function correlates the two tasks. Such multi-task learning approach yields highly accurate estimation of per-ray surface-crossing points.

Learning view-dependent implicit fields has been well-exploited in neural radiance fields (NeRF) [41] with great success. Recently, NeRF was combined with MVSNet for better generality [6]. Albeit sharing conceptual similarity, our work is completely different from NeRF. *First*, NeRF (including MVSNeRF [6]) is designed for novel view synthesis, a different task from MVS. *Second*, the radiance field in NeRF is defined and learned in continuous 3D space and camera rays are used only in the volume rendering stage. In our RayMVSNet, on the other hand, we explicitly learn 1D implicit fields on a camera ray basis. *Third*, while NeRF is usually trained to fit a given scene, RayMVSNet naturally generalizes to novel scenes.

RayMVSNet was published in CVPR 2022 [67] where we demonstrated state-of-the-art performance of RayMVSNet on two public datasets over all learning-based methods. RayMVSNet achieves an overall reconstruction score of 0.33mm on DTU and an F-score of 59.48% on Tanks & Temples. In particular, RayMVSNet is able to produce high-quality depth estimation and point cloud reconstruction results in challenging scenarios such as objects/scenes with non-textured surface, severe occlusion, and highly varying depth range. Notably, since all rays share weights for the LSTM and the epipolar transformer, the RayMVSNet model is light weight. Moreover, the computation for each ray is highly parallelizable.

The ray-based solution, however, has an inherent limitation of insufficient context aggregation; it does not account for the interaction between neighboring rays. This may lead to degraded performance on larger and more complex scenes (such as those from ScanNet [11]) where context is more essential. In this paper, we propose RayMVSNet++, an augmented version of RayMVSNet, by enhancing the ray-based feature aggregation with *local-frustum-based context aggregation*. For each ray, we extract its features in the frustum centered around the ray learn. This amounts to select semantically relevant neighboring rays in the frustum and aggregate the contextual information from those rays. In particular, an attentional gating unit with the Gumbel-Softmax trick [25] is designed to make the selection of neighboring rays end-to-end trainable. This leads to more accurate and robust depth predictions, especially in the challenging scenarios such as poor lighting conditions or motion blur which cannot be well handled by existing methods.

RayMVSNet++ outperforms prior works (including RayMVSNet) on ScanNet, achieving an AbsRel of 0.058m. We also demonstrate that RayMVSNet++ is able to produce accurate results on two subsets of ScanNet containing textureless regions and exhibiting large depth variation.

Our work makes the following contributions (those which are newly introduced in RayMVSNet++ are marked with bullet symbol of "$*$"):

- A novel formulation of deep MVS as learning ray-based 1D implicit fields.
- An epipolar transformer designed to learn cross-view feature correlation with attention mechanism.
- A multi-task learning approach to sequential modeling and prediction of 1D implicit fields based on LSTM.
- A challenging test set focusing on regions with specular reflection, shadow or occlusion based on the DTU dataset [1] and associated extensive evaluations.
- A local-frustum-based context aggregation that extends the receptive field of the ray-based model, leading to more accurate and robust predictions.
- New experiments on the ScanNet dataset to comprehensively evaluate the performance in challenging senarios.

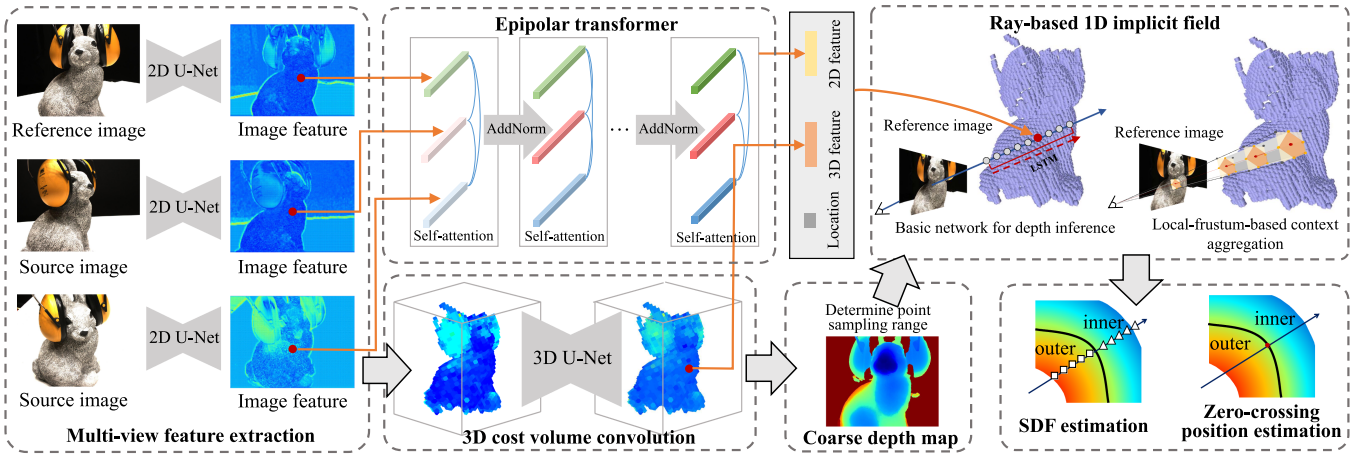

Fig. 2. Method overview. Given multiple overlapping RGB images, the multi-view image features are extracted by a 2D U-Net. The coarse depth map is then estimated by a coarse 3D cost volume. 2D multi-view image features are then correlated and aggregated by epipolar transformer. At last, the ray-based 1D implicit field, which includes a local-frustum-based context aggregation module, is learnt on each camera viewing ray to simultaneously estimate the SDF of the sampled points and the location of the zero-crossing point.

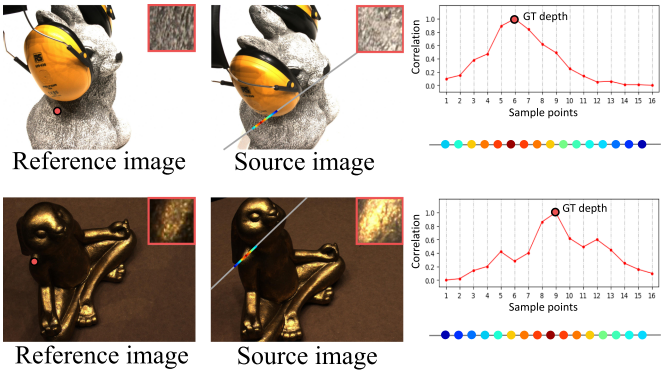

Fig. 3. Effects of epipolar transformer. Given a point in the reference image, epipolar transformer automatically selects reliable matching feature on the epipolar line of the source image. Note that it finds the matching feature correctly despite the influences of light changing (top row) and specular reflection (bottom row). The visualized point-pair correlations are deduced from the Softmax($\mathbf{QK}^T$) in Formulation (1).

II. Related Work

Learning-Based MVS: Recent advances have made remarkable progress on learning-based MVS. Hartmann et al. [20] first propose to learn the multi-patch similarity from two views by a Siamese convolutional network. SurfaceNet [26] and DeepMVS [23] warp the multi-view images into the 3D cost volume and adopt 3D neural networks to estimate the geometry. LSM [28] introduces differentiable projection operation to enable the end-to-end 3D reconstruction from multi-view images. MVSNet [74] proposes a differentiable homography and leverages 3D cost volume in a learning pipeline. MVSNet aggregates contextual information by a 3D convolutional network, especially on regions with complex illumination, specularity, and occlusion. However, the high computation and memory consumption restrict the output depth resolution, limiting its scalability in large scenes.

To reduce the requirements, many follow-up works have been developed. R-MVSNet [75] proposes to regularize the 2D cost maps along the depth direction so the memory consumption could be greatly reduced. Point-MVSNet [7] first computes the coarse depth with a low-resolution cost volume and then uses a point-based refinement network to generate the high-resolution depth map. CasMVSNet [19] adopts a cascade cost volume to gradually narrow the depth range and increase the cost volume resolution. Similar ideas are later explored to reduce the memory cost of 3D convolutions and/or increase the depth quality, such as coarse-to-fine depth optimization [10], [39], [68], [69], [71], [72], [79], attention-based feature aggregation [38], [66], [78], [84], and patch matching-based method [37], [62]. Unlike these works, RayMVSNet optimizes the depth on each camera viewing ray instead of the 3D volume, which is more light-weight.

Multi-view feature aggregation is one of the most crucial components in learning-based MVS. Previous works adopted various solutions to learn mutual correlations [85], avoiding the influences of incorrect matches caused by occlusion. Popular solutions include the visibility-based aggregation [8], [80], the attention-based aggregation [66], [73], [77], etc. RayMVSNet follows the attention-based aggregation route. Nevertheless, it learns feature aggregation at each 3D point, instead of the entire image or volume, thus greatly reducing the memory consumption.

Our method is also relevant to [42], [55] in terms of reconstructing 3D object by estimating the SDFs. While their methods focus on reconstructing the global TSDF volume of large-scale scenes and generating 3D surfaces with good completeness, our method estimates local SDFs on each camera ray individually resulting in more accurate depth estimation.

Learning MVS With Transformers: Since the pioneering work of [58], Transformers have significantly advanced the research of natural language processing [29], [30], [33]. More recently, Transformers show great potential in vision tasks, such as image classification [4], [15], object detection [4], scene segmentation [12], panoptic segmentation [35], pose estimation [47], and visual localization [54], thanks to the superb capabilities of modeling long-range dependencies. There are also a bunch of

works that utilize Transformers to capture the long-range relations in solving MVS problems. Most of them aggregate context from the extracted 2D image features and solve the cross-view matching problem. For example, MVSTR [87], LANet [84] and TransMVSNet [14] unitize the attention mechanisms to extract dense features with global contexts. PA-MVSNet [82] and AACVP-MVSNet [78] introduce self-attention layers for hierarchical features extraction, which is able to capture multi-scale matching clues for the subsequent depth inference task. AttMVS [38] introduces an attention-enhanced matching confidence volume to improve the matching robustness. To reduce the searching cost, recent researches [21], [73] have been focusing on leveraging the geometric prior of epipolar line by restricting attention associations within the epipolar line, which makes the learning more efficient. Our method also utilizes the epipolar geometric prior. However, it is different from the previous works as the proposed epipolar transformer essentially learns feature fusion at a 3D point by aggregating multi-view image features, while previous methods learn the matching of 2D pixels from two images. This leads to different network architectures.

Learning Implicit Representation: Many works have attempted learning shape representation based on implicit fields. Implicit field shows promising results on facilitating a variety number of problems, such as shape reconstruction [13], [43], [81], [86] and rendering [41], [56]. It achieves high quality shape reconstruction by allocating a value to every point in 3D space and extracting the shape surface as an iso-surface. DeepSDF [45] proposes to predict the magnitude of 3D point to indicate the distance to the surface boundary and a sign to determine whether the point is inside or outside of the shape. IM-Net [9] and Occupancy Network [40] learn the implicit fields to estimate the point-wise occupancy probability with a binary classifier. To improve the effectiveness and generalization on complex scenes, latest studies propose to enhance implicit field by introducing extra inputs [46], [70], adopting advanced learning techniques [16], [44], [53], [57] and decomposing the scene into local regions [5], [18], [27], [56]. In particular, PIFu [48] proposes an implicit representation that locally aligns pixels of 2D images with the global context of their corresponding 3D object. The method is able to infer both the object surface and texture from single or multiple input images.

NeRF [41] represents complex scenes by learning a view-dependent implicit neural radiance field, achieving high-resolution realistic novel view synthesis. Aside from the reasons mentioned in the introduction, our method is different from NeRF in the following aspects. First, NeRF learns the radiance field by MLPs. In contrast, our method tackles the problem of cross-view feature correlation with sequential modeling. Second, our model generalizes to untrained scenes, while NeRF generally does not. To increase the NeRF's generality on untrained scenes, a series of methods have been proposed, such as NeuralRay [36], TransNeRF [61]. In particular, IBRNet [64] learns multi-view image-based rendering with a ray transformer, bringing great cross-scene generality. Despite the similarity in the concept of inference on the camera ray, our task is different from theirs, resulting in different network designs and training schemes.

Since NeRF is designed for view synthesis, it has inferior abilities on approximating the scene's geometry, due to the shape-radiance ambiguities [83]. Recent works have investigated incorporating the geometric priors or clues, such as the depth prior [65] and the TSDF [63], [76], to enhance the scene reconstruction performance while maintaining the quality of view synthesis. Our method is also different from these methods, as these methods are trained with both appearance and geometry supervision while our method only requires the latter.

III. Method

Overview: RayMVSNet++ estimates the depth maps from multiple overlapping RGB images. Similar to [74], at each time, it takes one reference image I_1 and $N-1$ source images $\{I_i\}_2^N$ as input, and infers the depth map of the reference image. RayMVSNet++ starts from building a light-weight 3D cost volume and estimating a coarse depth map (Section III-A). Then, epipolar transformer is proposed to learn the matching correlation of the pixel-wise 2D features of each view using attention mechanism (Section III-B). The transformed features are fed into the 1D implicit field, implemented by an LSTM, along each camera viewing ray to estimate the signed distance functions (SDFs) of the hypothesized points as well as the zero-crossing position (Section III-C). In particular, a local-frustum-based context aggregation is introduced to aggregate more context from the semantically relevant neighboring rays. The method overview is illustrated in Fig. 2.

A. 3D Cost Volume and Coarse Depth Prediction

We first feed the multi-view images $\{I_i\}_1^N$ to a 2D U-Net to extract image features $\{\mathbf{F}_i^I\}_1^N$. The width and height of the image features are the same to those of the input images. Hence, $\{\mathbf{F}_i^I\}_1^N$ preserve the fine appearance feature of local details, facilitating the high-resolution depth estimation. By leveraging the 2D multi-view image features and the camera parameters, we build a variance-based 3D cost volume V, and extract the 3D volumetric features $\mathbf{F}^V$ via a 3D U-Net [74]. Since 3D convolution is memory-consuming, the resolution of V in our work is set to be smaller than that in the previous works [10], [19], [72]. The coarse depth maps are estimated from the 3D volumetric features, which are then used for determining the modeling range of the ray-based 1D implicit fields.

B. Epipolar Transformer

We cast a set of rays $\mathbf{R} = \{\mathbf{r}_i\}_1^M$ from the camera's viewing direction of the reference image, where M is the number of pixels in the reference image. Our goal is to estimate the location of the zero-crossing point on each ray, so we can obtain the depth map of the reference view. Compared to methods that estimate depth on the 3D cost volume, the ray-based method maintains the following advantages. *First*, since the depth map is view-dependent, ray-based depth optimization is more straightforward and light-weight. *Second*, all the ray-based 1D implicit fields share an identical spatial property, i.e., the monotonicity of the SDFs along the ray direction. As a result, the learning

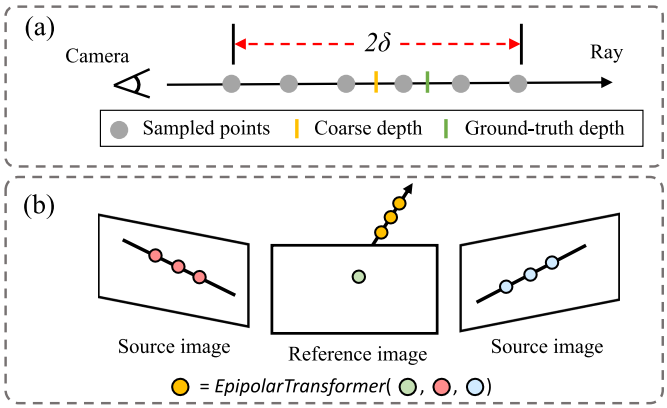

Fig. 4. (a) The hypothesized points are sampled around the predicted coarse depth to narrow down the search space of the zero-crossing position. (b) Epipolar transformer learns the matching correlation of the pixel-wise 2D features and aggregates these features using an attention mechanism.

would be simplified and well regularized, leading to efficient network training and more accurate results.

Zero-Crossing Hypothesis Sampling: We perform a point sampling to generate the zero-crossing point hypothesis on each ray. Ideally, one could generate as many points as possible on each ray. However, most of the points are far from the surface, providing less informative information for the depth estimation. To facilitate efficient training, as shown in Fig. 4(a), we adopt the coarse depth map predicted in Section III-A and uniformly sample K points $P = \{p_k\}_1^K$ on the ray in the range of $\pm\delta$ around the estimated coarse depth.

Attention-Aware Cross-View Feature Correlation: The next step is to aggregate feature for the hypothesized points based on the multi-view image features. A naive way to achieve this is to fetch the features from multi-view images based on the view projection, and take the variance. However, image feature could be easily influenced by image defects, such as specular reflection and light changing. Naive variance considers all image features equally, which might incur unreliable features and provide incorrect cross-view feature correlation. To alleviate this problem, we propose *Epipolar Transformer* to learn cross-view feature correlation with attention mechanism (Fig. 4(b)).

To be specific, the network architecture of epipolar transformer contains four self-attention layers, each followed by two AddNorm layers and one feed-forward layer. Suppose $\mathbf{X} = \text{Concat}(\mathbf{F}^I_{1,p}, \ldots, \mathbf{F}^I_{N,p})$, where $\text{Concat}(\cdot)$ is the concatenation operation, $\{\mathbf{F}^I_{i,p}\}_1^N$ are the fetched multi-view image features at 3D point p. The self-attention layer of epipolar transformer is:

$$\mathbf{S} = \text{SelfAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\mathbf{Q}\mathbf{K}^T)\mathbf{V}, \quad (1)$$

where $\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathbf{Q}}$, $\mathbf{K} = \mathbf{X}\mathbf{W}^{\mathbf{K}}$, $\mathbf{V} = \mathbf{X}\mathbf{W}^{\mathbf{V}}$ are the query vector, the key vector and the value vector respectively. $\mathbf{W}^{\mathbf{Q}}$, $\mathbf{W}^{\mathbf{K}}$, $\mathbf{W}^{\mathbf{V}}$ are the learned weights. Examples to demonstrate the effects of first self-attention layer in epipolar transformer are visualized in Fig. 3. The AddNorm layer of epipolar transformer is:

$$\mathbf{Z} = \text{AddNorm}(\mathbf{X}) = \text{LayerNorm}(\mathbf{X} + \mathbf{S}), \quad (2)$$

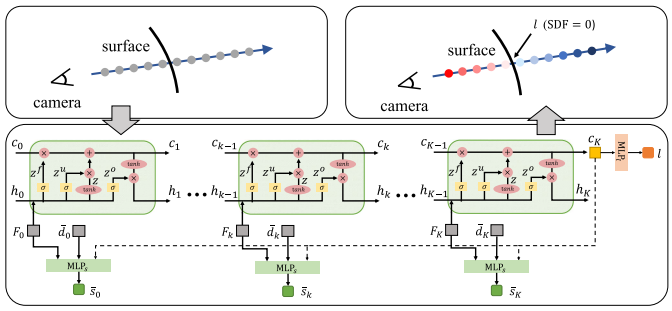

Fig. 5. Basic network architecture of the ray-based 1D implicit field. The hypothesized points are fed into an LSTM sequentially, to estimate the position of the zero-crossing point as well as the SDFs.

where $\text{LayerNorm}(\cdot)$ is the layer normalization operation. The output of epipolar transformer is the attention-aware denoised multi-view feature $\mathbf{F}^A_p = \{\mathbf{F}^A_{1,p}, \ldots, \mathbf{F}^A_{N,p}\}$.

To further improve the feature quality, we concatenate the attention-aware feature with the 3D volume feature $\mathbf{F}^V_p$ fetched from the 3D cost volume processed in Section III-A:

$$\mathbf{F}_p = \text{Concat}(\mathbf{F}^A_{\mu,p}, \mathbf{F}^A_{\sigma,p}, \mathbf{F}^A_{1,p}, \mathbf{F}^V_p). \quad (3)$$

where $\mathbf{F}^A_{\mu,p}$ and $\mathbf{F}^A_{\sigma,p}$ are the mean and variation of the elements in $\mathbf{F}^A_p$ [24], [74]. $\mathbf{F}^A_{1,p}$ is the attention-aware feature at 3D point p in the reference image.

C. Ray-Based 1D Implicit Field

LSTM Versus Alternative: Given the features of the hypothesized points, the ray-based 1D implicit fields are learned with an LSTM [22]. Crucially, we leverage two attributes of LSTM. *First*, the mechanism of sequential processing inherently facilitates the learning of the SDF monotonicity along the ray direction. *Second*, the property of time invariance increases the network robustness by allowing the zero-crossing position to appear at any place (time-step) on the ray. An alternative to performing sequential inference is to use transformer [58]. However, we experimentally found that replacing LSTM with transformer would not make the performance improve (see Table VII). The reason might be that transformer, which is designed for modeling non-local relations, does not explicitly encode relative or absolute position information [50], making it less suitable to our zero-crossing position searching problem.

Basic Network Architecture: The network architecture of the 1D implicit field is shown in Fig. 5. The LSTM first aggregates the hypothesized points sequentially, and generates the ray feature $\mathbf{c}_K$. Specifically, the formulations of an LSTM unit at time-step k are:

$$\begin{aligned}
\mathbf{z} &= \tanh(\mathbf{W}[\mathbf{F}_k, \mathbf{h}_{k-1}] + b),\\
\mathbf{z}^f &= \sigma(\mathbf{W}^f[\mathbf{F}_k, \mathbf{h}_{k-1}] + b^f),\\
\mathbf{z}^u &= \sigma(\mathbf{W}^u[\mathbf{F}_k, \mathbf{h}_{k-1}] + b^u),\\
\mathbf{z}^o &= \sigma(\mathbf{W}^o[\mathbf{F}_k, \mathbf{h}_{k-1}] + b^o),\\
\mathbf{c}_k &= \mathbf{z}^f \circ \mathbf{c}_{k-1} + \mathbf{z}^u \circ \mathbf{z},\\
\mathbf{h}_k &= \mathbf{z}^o \circ \tanh(\mathbf{c}_k),
\end{aligned} \quad (4)$$

where $\mathbf{F}_k$ is the feature of point p_k, $\mathbf{h}_k$ and $\mathbf{h}_{k-1}$ are the hidden state of point p_k and p_{k-1} respectively, $\mathbf{z}$ is the cell input activation vector, $\mathbf{z}^f$ is the activation vector of the forget gate, $\mathbf{z}^u$ is the activation vector of the update gate, $\mathbf{z}^o$ is the activation vector of the output gate, $\mathbf{c}_k$ is the cell state vector, $\mathbf{W}$, $\mathbf{W}^f$, $\mathbf{W}^u$, $\mathbf{W}^o$ are the weight matrices, b, b^f, b^u, b^o are the weight vectors, $\circ$ is the element-wise multiplication, $\sigma(\cdot)$ is the sigmoid function. The LSTM is initialized with $\mathbf{c}_0 = 0$ and $\mathbf{h}_0 = 0$.

For each hypothesized point p_k, we use the ray feature $\mathbf{c}_K$, the point-wise feature $\mathbf{F}_k$ and its depth value d_k (indicating the location on the ray) to estimate its SDF s_k using an MLP. Instead of using the true depth value d_k and estimating the true SDF s_k, we use the normalized depth value $\overline{d}_k = k/K$ and the normalized SDF $\overline{s}_k = s_k/s_{\max} \in [-1, 1]$, where $s_{\max}$ is the maximal absolute SDF value on the ray. Such normalization leads to a significant reduction of learning complexity and improvement of the result quality. The formulation of the SDF prediction is:

$$\overline{s}_k = \text{MLP}_s([\mathbf{c}_K, \mathbf{F}_k, \overline{d}_k]). \tag{5}$$

The above network predicts the SDFs of the hypothesized points on the ray. However, post-processing, e.g., ray casting, is still needed to find the zero-cross position. We extend our method to estimate the zero-cross position explicitly with another MLP. Taking the ray feature $\mathbf{c}_K$ as input, the MLP predicts the zero-crossing location l on the ray in the normalized 1D coordinate:

$$l = \text{MLP}_l(\mathbf{c}_K). \tag{6}$$

Local-Frustum-Based Context Aggregation: The basic network architecture described above focuses on the inference along each ray direction. This method would work in scenarios where the images are clearly captured under satisfactory conditions, e.g., in good lighting and without motion blur. This is because in such scenarios the features aggregated along the ray direction are able to provide sufficient information to infer the underlying geometry. Nevertheless, there is a flurry of data [11], [52] that does not meet these requirements, making the depth estimation either inaccurate or infeasible. As such, specific mortifications should be taken to allow the method to tolerate those disadvantages.

We tackle this problem by proposing a simple yet effective method: consider the interaction between neighboring rays and aggregate more contextual feature to boost ray-based inference. To achieve this, based on the basic network architecture above, we introduce a local-frustum-based context aggregation module that adaptively aggregates contextual features from neighbouring rays. By involving more context, the ray feature $\mathbf{c}_K$ and 3D point feature $\mathbf{F}_k$ in formulation (5) and formulation (6) are expected to be more informative and thus result in more accurate depth estimation.

To achieve this, we first extract the features of each ray individually by the above LSTM. By projecting the ray features to the corresponding pixels in the reference image, we generate a feature map whose width and height are the same as those of the reference image. For any pixel in the feature map, we set its receptive field as a square with width t, and the center is the pixel. Suppose $\mathbf{c}_K^{\text{cen}} \in \mathbb{R}^\kappa$ is the extracted feature of the center pixel. κ is the feature-length. $\{\mathbf{c}_K^\theta\}, \theta \in (1, \Theta)$ are the extracted feature of the neighbouring pixels, where $\Theta = (t+1)^2 - 1$ is the number of neighbouring pixels.

A naive solution to aggregate context in the square is using average-pooling or max-pooling. However, as not all neighboring rays are equally important to the central ray, the naive pooling would involve irrelevant features and therefore debilitate the network training. To address this problem, we introduce an *attentional gating unit* (see Fig. 6) that dynamically selects semantically relevant neighboring rays within the local frustum and adaptively aggregates their features, conditioned on the extracted ray-wise features.

We first consider the variance of $\mathbf{c}_K^{\text{cen}}$ and $\{\mathbf{c}_K^\theta\}, \theta \in (1, \Theta)$, and generate a tensor $\hat{R}^s = \{\mathbf{c}_K^\theta - \mathbf{c}_K^{\text{cen}}\}, \theta \in (1, \Theta)$. $\hat{R}^s \in \mathbb{R}^{\kappa \times \Theta}$ is taken as the input to the gating unit G, estimating the soft gating decisions $G^s \in \mathbb{R}^\Theta$ that indicates how relevant are each ray to the central ray:

$$G^s = \sigma(G(\hat{R}^s) + g), \tag{7}$$

where g is the Gumbel noise, the gating unit G is implemented as a 1D MLP in our method. Note that although we do not use any semantic supervision directly, we found that most of the selected pixels have the same semantic labels as the center pixel (see Fig. 20). The reason is that both $\mathbf{c}_K^{\text{cen}}$ and $\{\mathbf{c}_K^\theta\}$ are high-level features which already contain semantic information.

Then, a Gumbel-Softmax module H [25] turns soft decisions G^s into hard decisions $H^s \in \{0, 1\}^\Theta$ by replacing the softmax with an argmax during the forward pass and retaining the softmax during the backward pass [32], [60]:

$$H^s = H(G^s). \tag{8}$$

The hard decision H^s is a binary mask that indicates which ray is semantically relevant to the center ray. The Gumbel-Softmax module provides a mechanism that outputs a binary mask in the forward pass and also allows the gradient to be back-propagated in the backward pass. As is shown in Fig. 7, the gating attentional unit is end-to-end trainable.

Having determined the semantically relevant neighboring rays, we then aggregate context on the activated positions in the mask. We consider contextual feature aggregation from two aspects.

For *ray feature aggregation*, we take the features from the activated positions, take the average, and add it to the initial central ray feature:

$$\mathbf{c}_K^a = \frac{\sum_{\theta=1}^{\Theta}(R^s \circ H^s)}{\|H^s\|_0} \oplus \mathbf{c}_K, \tag{9}$$

where $\mathbf{c}_K^a \in \mathbb{R}^\kappa$ is the aggregated feature of the central ray, $R^s = \{\mathbf{c}_K^\theta\}, \theta \in (1, \Theta)$ are the features of the neighbouring rays, $\circ$ is the element-wise multiplication, $\oplus$ is the element-wise addition, $\|H^s\|_0$ is the number of activated pixel in H^s.

For *sample points feature aggregation*, we adopt the same mask and aggregate feature at the k-th layer of the frustum:

$$\mathbf{F}_k^a = \frac{\sum_{\theta=1}^{\Theta}(P_k^s \circ H^s)}{\|H^s\|_0} \oplus \mathbf{F}_k, \tag{10}$$

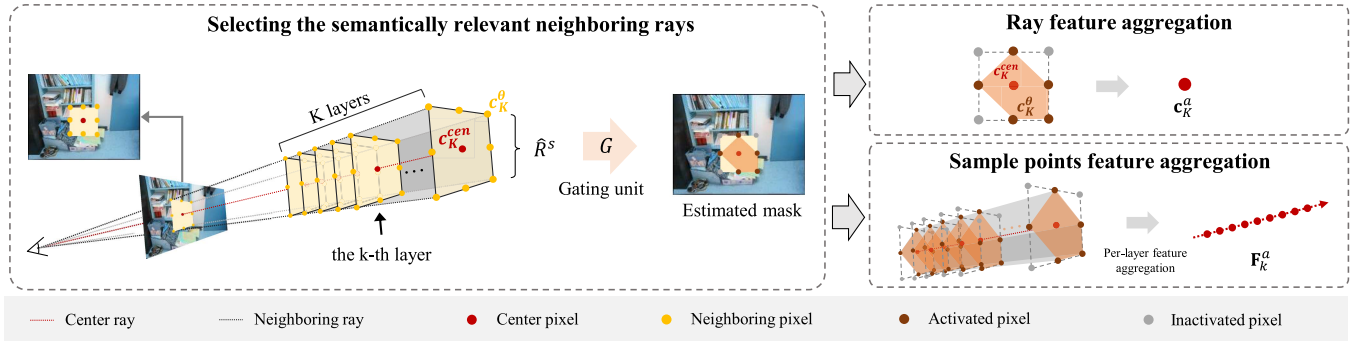

Fig. 6. Attentional gating unit for the local-frustum-based context aggregation. For each camera rays of the reference image, the method estimates a mask that denotes the semantically relevant neighbouring rays. Based on the mask, ray feature and sample points feature, with more contextual information, are aggregated respectively.

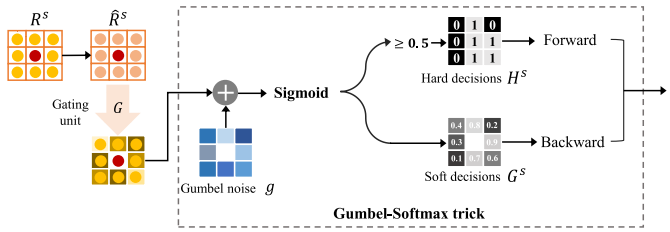

Fig. 7. Training to estimate the mask with the Gumbel-Softmax trick. The gating unit G (with adding the Gumbel noise) generates the soft gating decisions G^s. The soft decisions are converted into hard decisions during the forward pass. The soft decisions are retained for the backward pass, making the gating unit differentiable.

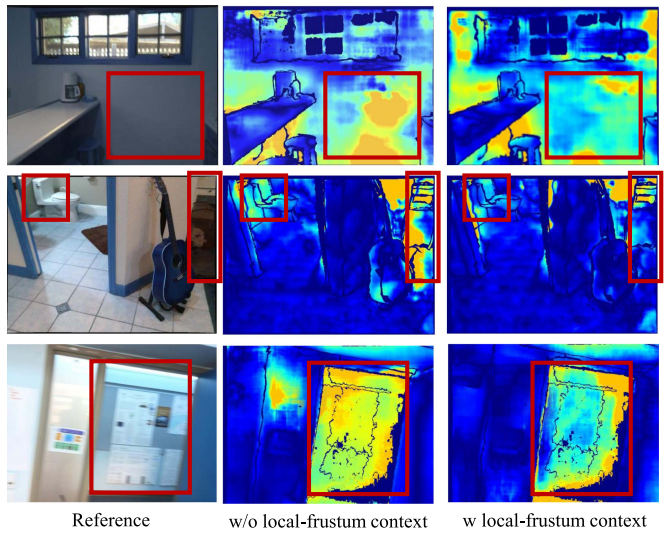

Fig. 8. Visual comparison of depth estimation with and without the local-frustum-based context aggregation. Please pay attention to the results of the challenging areas highlighted in the figure.

where $\mathbf{F}_k^a \in \mathbb{R}^\kappa$ is the aggregated feature of the k-th sampled point in the central ray, $P_k^s = \{F_k^\theta\}, \theta \in (1, \Theta)$ is the feature map of the neighbouring rays at the k-th layer of the frustum.

Last, we replace the $\mathbf{c}_K$ and $\mathbf{F}_k$ by $\mathbf{c}_K^a$ and $\mathbf{F}_k^a$, respectively, in formulation (5) and formulation (6). Therefore, the SDF prediction and zero-crossing location are turned to be:

$$\begin{aligned} \overline{s}_k &= \text{MLP}_s([\mathbf{c}_K^a, \mathbf{F}_k^a, \overline{d}_k]), \\ l &= \text{MLP}_l(\mathbf{c}_K^a). \end{aligned} \tag{11}$$

This local-frustum-based context aggregation improves the performance on datasets where challenging regions and low-quality images exist. The low-quality images are typically captured due to motion blur or bad lighting conditions, which cannot be well handled by existing methods. We found that the attentional gating unit, without any semantic supervision, tends to select the pixels that belong to the same object as the central pixel. That is why we claim that the proposed method is able to select semantically relevant neighboring rays for context aggregation. Please see a visual illustration of its effects in Fig. 8.

Loss Functions: We adopt a multi-task learning strategy to optimize the network. The two tasks, i.e., SDF estimation and zero-crossing position estimation, are inherently relevant and could reinforce each other by optimizing the following loss:

$$\mathcal{L} = w_s \mathcal{L}_s + w_l \mathcal{L}_l + w_{sl} \mathcal{L}_{sl}, \tag{12}$$

where $\mathcal{L}_s$ and $\mathcal{L}_l$ are the loss of the SDF estimation and the zero-crossing location estimation, respectively:

$$\begin{aligned} \mathcal{L}_s &= \sum_{k=1}^{K} L_1(s_k, \hat{s_k}), \\ \mathcal{L}_l &= L_1(l, \hat{l}), \end{aligned} \tag{13}$$

where $\hat{s_k}$ and $\hat{l}$ are the ground-truth, $L_1(\cdot)$ denotes the L1 loss function. $\mathcal{L}_{sl}$ is a relational loss that penalizes the inconsistency between the predicted SDFs and the predicted zero-crossing position:

$$\mathcal{L}_{sl} = \begin{cases} 1, & s_l^a \times s_l^b > 0 \\ 0, & s_l^a \times s_l^b \leq 0, \end{cases} \tag{14}$$

where s_l^a and s_l^b are the predicted SDF of the closest two sampled points around the predicted zero-crossing position on the ray. w_s, w_l, w_{sl} are the pre-defined weights.

D. Implementations

We provide implementation details of the training and inference. The input image size are 640×512, 1600×1200, and 640×480 for the DTU, the Tanks & Temples, and the ScanNet datasets, respectively. The 2D U-Net consists of 6 convolutional layers and 6 deconvolutional layers, each followed by a batch normalization layer and a ReLU layer, except for the last ones. The 3D cost volume is fed into a 3D U-Net which consists of three 3D convolutional layers and three 3D deconvolutional layers. On each ray, the number of hypothesized points K is 16. The feature fetching from images and volume are achieved by using bilinear interpolation and trilinear interpolation, respectively. The hidden dimension of $\mathbf{z}, \mathbf{z}^f, \mathbf{z}^u, \mathbf{z}^o, \mathbf{c}_k, \mathbf{h}_k$ are 50. MLP_l and MLP_s both contain 4 fully-convolutional layers. The weights w_s, w_l, w_{sl} of multi-task learning loss function are 0.1, 0.8, 0.1, respectively. Epipolar transformer and the LSTM are jointly trained. We use Adam optimizer with initial learning rate 0.0005 which is decreased by 0.9 for every 2 epochs. The training takes 48 hours. The inference time is about 2 seconds. We filter and fuse the depth maps to produce 3D point cloud like previous work [74]. The receptive field t of the local-frustum-based context aggregation is 9. During the training of the attentional gating unit, we use a similar strategy to [32] that penalizes the trivial solution, e.g., simply using all the neighboring pixels. We found this strategy would greatly facilitate the training. The RayMVSNet++ is trained and tested on an NVIDIA Tesla V100-SXM2.

IV. Results and Evaluation

A. Datasets and Evaluation Metrics

We performed a series of experiments on multiple datasets to evaluate how well our method performs on different scenarios. The experimental datasets are:

- *DTU* [1]: The DTU dataset contains 79 training scans and 22 testing scans, all captured under changing lighting conditions. Since DTU did not provide SDF annotations, we densely generate the point-wise SDFs from the reconstructed surfaces [45], [74]. Besides, three challenging test subsets focusing on regions with *Specular reflection*, *Shadow* and *Occlusion* are created from the DTU test set. These regions are manually annotated and are designed for evaluating the method's performance in challenging cases. Please refer to the supplemental material, available online for the subsets details.
- *Tanks & Temples* [31]: To evaluate the generality, we test our method on the Tanks & Temples dataset which contains large-scale complex scenes, using the trained model on *DTU* without any fine-tuning.
- *ScanNet* [11]: The ScanNet dataset is originally collected for the purpose of RGB-D reconstruction and scene understanding. Since the images are captured in various indoor scenes under ordinary conditions, we utilize the ScanNet dataset for examining the methods' ability on data with low-quality images. Specifically, we collect 31,051 image triples for training and 1,467 image triples for testing. The test set could be divided into two subsets: *Textureless* and *Large depth variation*. The point-wise SDFs are generated from the reconstructed surfaces [45], [74].

TABLE I
Statistics of the Experimental Datasets

Dataset	Subset	#View	#Object	#Scene
DTU [1]	Train	3871	79	-
	Test	1078	22	-
	Specular reflection	233	5	-
	Shadow	294	6	-
	Occlusion	254	5	-
ScanNet [11]	Train	31051	-	851
	Test	967	-	401
	Textureless	329	-	117
	Large depth variation	171	-	80
Tanks and Temples [31]	Test	2100	8	-

The statistics of the experimental datasets are reported in Table I.

We use the following metrics to evaluate the performances on different datasets, respectively:

- *Accuracy & Completeness* [49]: the metric that evaluates the accuracy of the reconstructed points (i.e., how close the reconstructed points are to the ground-truth surface) and the completeness of the reconstructed points (i.e., how much of the ground-truth surface is modeled by the reconstructed points). Besides, an overall score is computed as the mean of the accuracy and completeness to indicate the performance considering both the two factors. We use this metric to evaluate the performance on *DTU*.
- *F-score* [31]: the metric that evaluates the precision and recall of the reconstructed points with a specific distance threshold. We use this metric to evaluate the performance on *Tanks & Temples*. The distance thresholds are different for the tested scenes according to [31]. F-score is different to the overall score in Accuracy & Completeness, as it uses the harmonic mean, instead of the arithmetic mean, of precision and recall, resulting in a more balanced metric for measuring the two factors at the same time.
- *Depth accuracy* [34]: we use several metrics for evaluating the estimated depth map comprehensively. The metrics include: AbsRel, SqRel, Log10, RMSE, RMSELog, $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$, and Percentage @ x. Table V reports the details. We use this metric to evaluate the performance on *DTU* and *ScanNet*.

B. Performance on DTU

Evaluation on Reconstructed Point Cloud: To evaluate RayMVSNet on *DTU*. we compare *Accuracy & Completeness* of the reconstructed point cloud. The quantitative results are shown in Table II. It shows that our method not only produces competitive results in terms of *Accuracy* and *Completeness*, but also achieves the state-of-the-art *Overall* performance. This demonstrates the effectiveness of our method, especially on balancing the trade-off between *Accuracy* and *Completeness*. The qualitative comparisons are visualized in Fig. 9. It is shown

TABLE II
QUANTITATIVE RESULTS ON THE *DTU* DATASET

Method	Accuracy	Completeness	Overall
Gipuma [17]	0.283	0.873	0.578
MVSNet [74]	0.396	0.527	0.462
R-MVSNet [75]	0.383	0.452	0.417
CIDER [69]	0.417	0.437	0.427
P-MVSNet [37]	0.406	0.434	0.420
Point-MVSNet [7]	0.342	0.411	0.376
Fast-MVSNet [79]	0.336	0.403	0.370
Att-MVSNet [38]	0.383	0.329	0.356
CasMVSNet [19]	0.325	0.385	0.355
CVP-MVSNet [72]	0.296	0.406	0.351
PatchmatchNet [62]	0.427	0.277	0.352
UCS-Net [10]	0.338	0.349	0.344
AACVP-MVSNet [78]	0.357	0.326	0.341
U-MVS [68]	0.354	0.353	0.354
RayMVSNet	0.341	0.319	0.330
RayMVSNet++	0.344	0.312	**0.328**

We compare all methods using the distance metric [1]. The numbers are reported in *mm* (Lower is better).

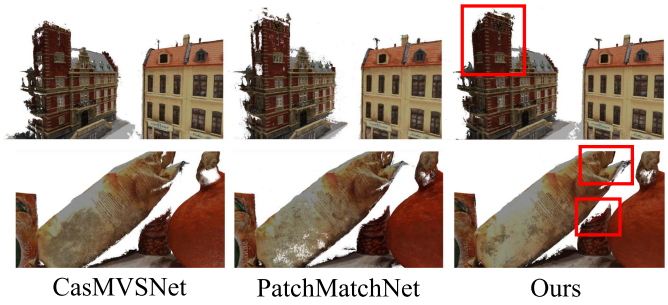

Fig. 9. Visual comparison of the reconstructed point cloud by RayMVSNet and the baselines. Please pay attention to the results of the challenging areas highlighted in the figure.

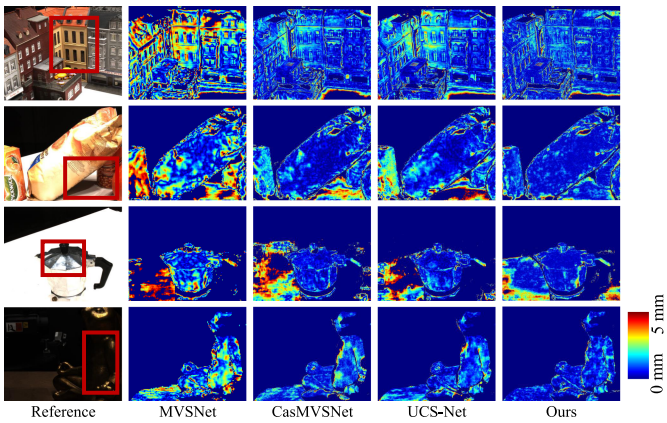

Fig. 10. Visual comparison of the estimated depth map by RayMVSNet and the baselines. Please pay attention to the results of the challenging areas highlighted in the figure.

that our method achieves high-quality reconstruction in various scenarios. In particular, our method outperforms the baselines in scenes with textureless regions, heavy occlusion, and complex geometry.

Evaluation on Challenging Regions: To further demonstrate our advantage, we compare RayMVSNet with existing works, in terms of the predicted depth map. The quantitative comparisons on the whole *DTU* test set (Fig. 11(a)) and the challenging subsets (Fig. 11(b), (c), and (d)) are reported. The Percentage @ x metric is used. The percentage (Y-axis) represents the ratio of the pixels whose depth prediction error is smaller than the specific error thresholds (X-axis). Higher percentages represent better performances. It is clear that our method outperforms all the baselines in all error thresholds. Crucially, our method is more general and robust in challenging cases as shown in Fig. 10, thanks to the prior learnt from the ray-based 1D implicit field.

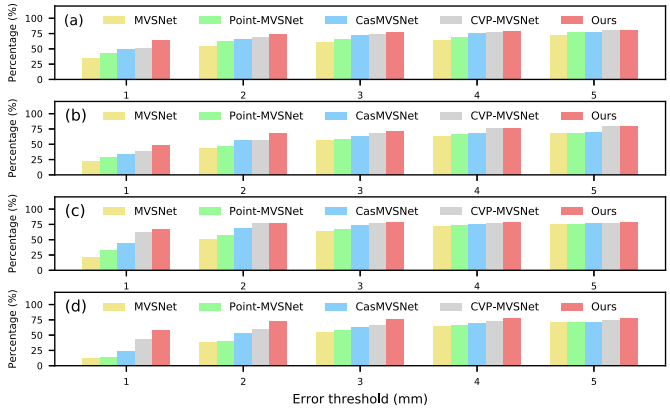

Fig. 11. Quantitative comparisons on the depth map prediction of the whole *DTU* test set (a) and the challenging test subsets: *Specular reflection* (b), *Shadow* (c) and *Occlusion* (d). Ours represents the proposed RayMVSNet. The percentage (Y-axis) represents the ratio of the pixels whose depth prediction error is smaller than the specific error thresholds (X-axis).

C. Performance on Tanks & Temples

We compare our method with the baselines on *Tanks & Temples*. Following the protocol of previous work [19], we use the network trained on *DTU*. *F-score* is the evaluation metric. The quantitative results are shown in Table III. RayMVSNet achieves the best performance, demonstrating the generality of epipolar transformer and ray-based 1D implicit field on large-scale scenes. RayMVSNet++ outperforms RayMVSNet on several test scenes while maintaining comparable mean performance to RayMVSNet. This is because most of the images in Tanks & Temples are captured under good conditions, e.g., in sufficient and stable lighting conditions without motion blur, which RayMVSNet is sufficient to handle. RayMVSNet++ is inferior to the baseline of D2HC-RMVSNet [71] in the scenes with large planar regions, such as Horse and Light house. The reason might be that D2HC-RMVSNet adopted a hybrid recurrent regularization module on the cost volume which provides a mechanism to implicitly involve the structural prior of planar regions for performance improvements.

D. Performance on ScanNet

Evaluation on Depth Estimation: We first evaluate our method in terms of depth estimation on *ScanNet*. The dataset contains low-quality images, so it is especially suitable for the evaluation of the proposed local-frustum-based context aggregation. The

TABLE III
QUANTITATIVE RESULTS ON THE *TANKS & TEMPLES* DATASET

Method	Family	Francis	Horse	Light house	M60	Panther	Playground	Train	Mean
MVSNet [74]	55.99	28.55	25.07	50.79	53.96	50.86	47.90	34.69	43.48
R-MVSNet [75]	69.96	46.65	32.59	42.95	51.88	48.80	52.00	42.38	48.40
PVA-MVSNet [77]	69.36	46.80	46.01	55.74	57.23	54.75	56.70	49.06	54.46
CVP-MVSNet [72]	76.50	47.74	36.34	55.12	57.28	54.28	57.43	47.54	54.03
CasMVSNet [19]	76.37	58.45	46.26	55.81	56.11	54.06	58.18	49.51	56.84
UCS-Net [10]	76.09	53.16	43.03	54.00	55.60	51.49	57.38	47.89	54.83
D2HC-RMVSNet [71]	74.69	56.04	**49.42**	**60.08**	59.81	59.61	**60.04**	**53.92**	59.20
U-MVS [68]	76.49	60.04	49.20	55.52	55.33	51.22	56.77	52.63	57.15
RayMVSNet	**78.55**	**61.93**	45.48	57.59	**61.00**	**59.78**	59.19	52.32	**59.48**
RayMVSNet++	77.82	60.10	44.51	58.21	58.32	57.23	59.20	52.34	58.47

We use the F-score as the evaluation metric (higher is better).

TABLE IV
QUANTITATIVE RESULTS ON THE *SCANNET* DATASET

Method	AbsRel↓	SqRel↓	Log10↓	RMSE↓	RMSELog↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Bts [34]	0.117	0.052	0.049	0.270	0.151	0.862	0.966	0.992
Bts* [34]	0.088	0.035	0.038	0.228	0.128	0.916	0.980	0.994
MVSNet [74]	0.098	0.046	0.042	0.256	0.143	0.893	0.968	0.989
CasMVSNet [19]	0.088	0.039	0.039	0.251	0.128	0.912	0.976	0.992
UCS-Net [10]	0.075	0.029	0.032	0.197	0.107	0.936	0.984	0.995
MVS2D [73]	0.069	0.022	0.030	0.188	0.097	0.951	0.990	**0.998**
RayMVSNet	0.074	0.029	0.037	0.195	0.107	0.947	0.986	0.996
RayMVSNet++	**0.058**	**0.016**	**0.025**	**0.157**	**0.085**	**0.963**	**0.993**	**0.998**

We use multiple metrics to comprehensively evaluate our method and several baselines on depth estimation.

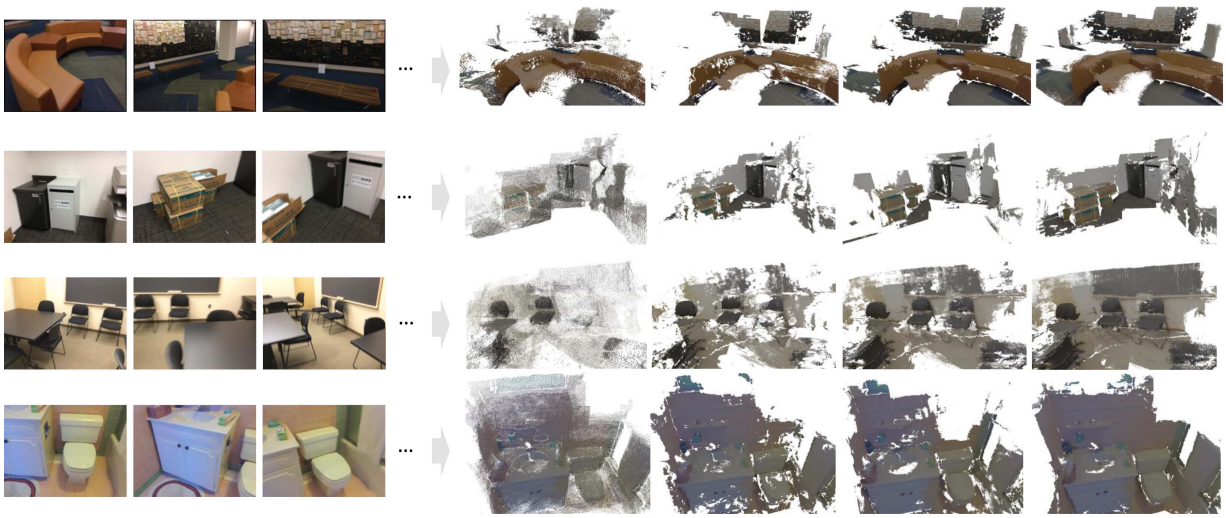

Fig. 12. Visual comparisons on the reconstructed point cloud on the *ScanNet* dataset [11]. RayMVSNet++ achieves better reconstruction results in terms of both accuracy and completeness, thanks to the local-frustum-based context aggregation which introduces more context to tolerate the imperfectness of the input images.

evaluation metrics for depth estimation were adopted. In this experiment, we set the receptive field t of the local-frustum-based context aggregation as 9 for RayMVSNet++. The results are reported in Table IV. We see RayMVSNet++ achieves the best performance in all metrics over existing methods. This demonstrates that RayMVSNet++ could tolerate the imperfection on the input images, i.e., motion blur or inferior lighting conditions. In particular, RayMVSNet++ outperforms RayMVSNet, confirming our motivation of developing RayMVSNet++, i.e., aggregating context on challenging regions and low-quality images exist. The visual comparisons on depth estimation are provided in Fig. 13.

Evaluation on Reconstructed Point Cloud: We also evaluate the quality of the reconstructed point cloud produced by our method. The *Accuracy & Completeness* metrics are adopted. As reported in Table VI, RayMVSNet++ achieves the best overall

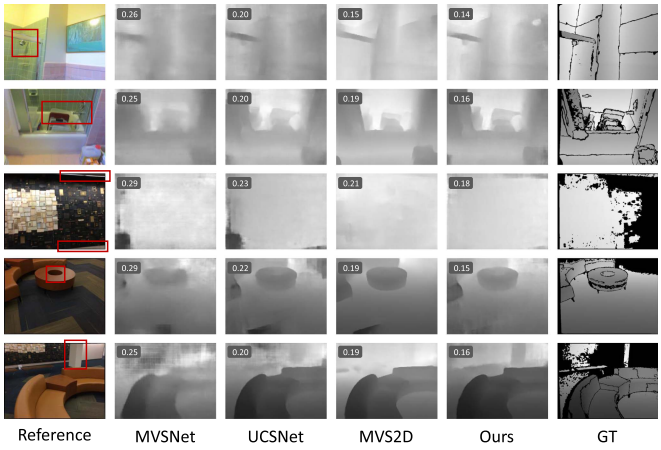

Fig. 13. Visual comparisons on the estimated depth on the ScanNet dataset. It shows that RayMVSNet++ achieves more accurate depth estimation in the challenging regions as highlighted. The RMSE (m) is reported on the upper-left of each example.

TABLE V
EVALUATION METRICS FOR DEPTH ESTIMATION

Metric	Formulation
AbsRel	$\frac{1}{N}\sum_i \frac{\lvert d_i - d_i^*\rvert}{d_i^*}$
SqRel	$\frac{1}{N}\sum_i \frac{(d_i - d_i^*)^2}{d_i^*}$
Log10	$\frac{1}{N}\sum_i \lvert\log_{10} d_i - \log_{10} d_i^*\rvert$
RMSE	$\sqrt{\frac{1}{N}\sum_i (d_i - d_i^*)^2}$
RMSELog	$\sqrt{\frac{1}{N}\sum_i (\log d_i - \log d_i^*)^2}$
$\delta < 1.25^k$	$\frac{1}{N}\sum_i (\max(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}) < 1.25^k)$
Percentage@x	$\frac{1}{N}\sum_i I(\lvert d_i - d_i^*\rvert < x)$

d_i and d_i^* denote the estimated depth and ground truth, respectively. N is the pixel bumber of the depth image

TABLE VI
QUANTITATIVE RESULTS ON THE *SCANNET* DATASET

Method	Accuracy	Completeness	Overall
Bts* [34]	0.048	0.213	0.130
MVSNet [74]	0.037	0.187	0.112
UCS-Net [10]	0.034	0.142	0.088
MVS2D [73]	**0.031**	0.147	0.089
RayMVSNet	0.034	0.153	0.094
RayMVSNet++	0.032	**0.138**	**0.083**

We evaluate the reconstructed point cloud using the distance metric [1]. The numbers are reported in *m* (lower is better).

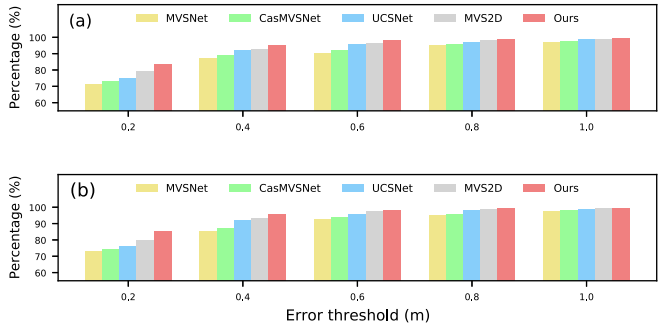

Fig. 14. Quantitative comparisons on the depth map prediction of the challenging test subsets in *ScanNet: Textureless* (a), *Large depth variation* (b). Ours represents the proposed RayMVSNet++. The percentage (Y-axis) represents the ratio of the pixels whose depth prediction error is smaller than the specific error thresholds (X-axis).

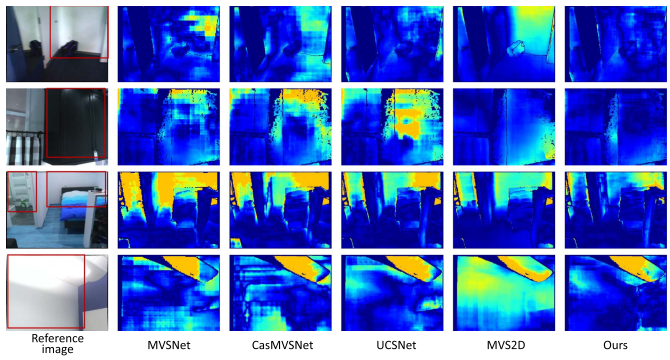

Fig. 15. Visual comparisons on the *Textureless* test set in the *ScanNet* dataset.

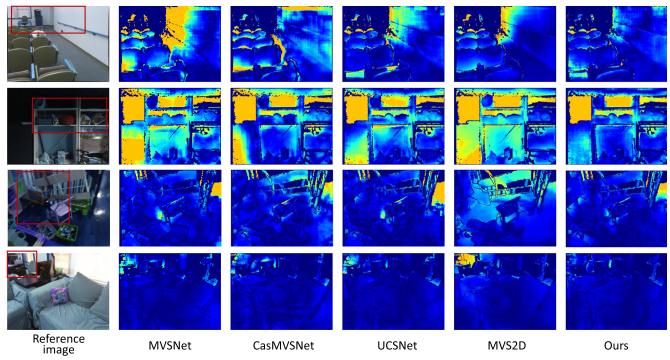

Fig. 16. Visual comparisons on the *Large depth variation* test set in the *ScanNet* dataset.

performance, which is consistent with the evaluation on depth estimation. We provide examples of visual comparisons on the reconstructed point cloud in Fig. 12.

Evaluation on Challenging Regions: We also study how our method performs in challenging regions to understand its effectiveness better. The experiment is conducted on the two subsets, i.e., *Textureless* and *Large depth variation*. Percentage @ x is the evaluation metric. Fig. 14 reports the results. We see RayMVSNet++ outperforms all the baselines in all test subsets with all error thresholds (X-axis). Notably, RayMVSNet++ outperforms the state-of-the-art methods by a large margin with an error threshold 0.2 mm. We visualize the examples of the challenging regions in Figs. 15 and 16. Note that the depth estimation on the highlighted regions is extremely difficult due to the textureless surfaces and the large depth variation.

E. Ablation Study

In Table VII, we conduct ablation studies to quantify the efficacy of several crucial components in RayMVSNet and RayMVSNet++. Unless specifically mentioned otherwise, the experiments are conducted on the *DTU* dataset.

TABLE VII
ABLATION STUDIES OF RAYMVSNET

Method	Accuracy	Completeness	Overall
w/o epipolar transformer	0.347	0.339	0.343
w/o 2D image feature	0.345	0.352	0.348
w/o 3D volume feature	0.434	0.322	0.378
vis-max feature aggregation	0.345	0.331	0.338
w/o ray-based inference	0.573	0.642	0.608
Ray with Transformer	**0.339**	0.343	0.341
Ray with average pooling	0.356	0.406	0.381
Ray with max pooling	0.466	0.383	0.424
w/o SDF prediction	0.354	0.330	0.342
Visibility-aware view aggregation	0.345	0.331	0.338
RayMVSNet	0.341	**0.319**	**0.330**

The performance under distance metric is reported (lower is better).

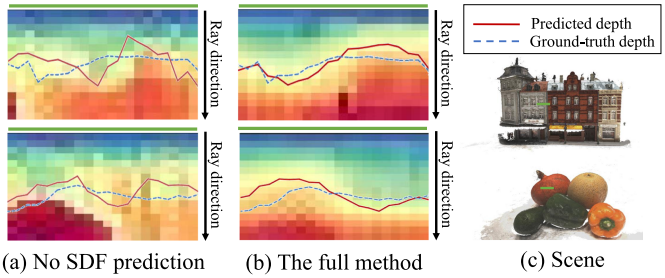

Fig. 17. Mid-layer feature map t-SNE visualization of the w/o SDF prediction baseline (a) and the full method (b) for the green segment marked in the scenes in (c).

Feature Aggregation: The cross-view feature aggregation is a key component of RayMVSNet. To evaluate the importance, we compare the full method to several baselines without some specific component: *w/o epipolar transformer*, *w/o 2D image feature* and *w/o 3D volume feature*. To be specific, *w/o epipolar transformer* denotes the baseline that discards the epipolar transformer and uses the fetched multi-view features $\mathbf{F}_p^I$ instead of the aggregated attention-aware feature of epipolar transformer $\mathbf{F}_p^A$ in equation (3). *w/o 2D image features* represents the baseline that discards the multi-view 2D image feature $\mathbf{F}_{\mu,p}^A$, $\mathbf{F}_{\sigma,p}^A$, and $\mathbf{F}_{1,p}^A$ in equation (3). *w/o 3D features* is the baseline that discards the 3D volume feature $\mathbf{F}_p^V$ in equation (3). It clearly shows that all these baselines make the performance decline. It is worth noting that *w/o epipolar transformer* achieves a lower completeness score, indicating epipolar transformer could make the reconstruction complete by providing more reliable cross-view correlations. We also compare our epipolar transformer to other multi-view feature aggregation methods. In the experiment of *vis-max feature aggregation*, we replace the epipolar transformer with the visibility-aware max-pooling feature aggregation [8]. The result indicates epipolar transformer is a better solution.

Ray-Based Inference: Our method learns the 1D implicit field by the ray-based inference. To show its necessity, a straightforward baseline is to learn the implicit field in the 3D space of the reference frustum, such that there is no ray-based inference. This baseline adopts the same cross-view feature aggregation as the full method, and predicts the SDF of sampled points in the reference frustum by using an MLP. The depth map is then generated by a ray-casting algorithm from the predicted SDFs. Unsurprisingly, experiments show this network is hard to converge and leads to low quantitative performance, which suggests that the ray-based 1D implicit field indeed simplifies the learning and is suitable to the MVS problem.

Other Ray-Based Implicit Field Models: In order to reveal the need of the proposed LSTM, we compare our method against several baselines with alternative models of processing sequential data. To be specific, we study the effects of replacing the LSTM with average pooling, max pooling, and Transformer [58], respectively. The *Ray with average pooling* and the *Ray with max pooling* baselines aggregate ray feature by average pooling and max pooling over all sampled points, respectively. The aggregated features are then used to predict the zero-crossing location. The point-wise SDF predictions are also performed as an auxiliary task. The result shows that our method outperforms all the baselines. In particular, the performance drops significantly with the *Ray with average pooling* and the *Ray with max pooling*, implying that the modeling of ray-based 1D implicit field is a non-trivial task. The *Ray with Transformer* is inferior to the full method, in terms of the *Overall score*, confirming that LSTM is more appropriate to our problem.

TABLE VIII
ABLATION STUDIES OF RAYMVSNET++

Method	RMSE(m)↓	p@0.2↑	p@0.4↑	p@0.6 ↑
w/o frustum	0.211	0.794	0.918	0.963
w/o gating unit	0.193	0.807	0.925	0.966
w/o Gumbel-Softmax	0.176	0.838	0.950	0.980
RayMVSNet++	**0.158**	**0.861**	**0.957**	**0.982**

P@x represents percentage@x.

No SDF Prediction: The SDF prediction is an auxiliary task in RayMVSNet. We demonstrate its influence by turning it off and comparing to the full method. The performance of *w/o SDF prediction* baseline is inferior to the full method, demonstrating the joint training of SDF prediction and zero-crossing position prediction is indeed helpful, due to the extra supervision of SDF. Examples are visualized in Fig. 17 which compares the mid-layer features of the full model and the baseline without SDF prediction. We can see that the mid-layer features of the full method, with SDF supervision, maintain a better monotonicity along the ray direction, resulting in more accurate predictions.

Alternative Multi-View Aggregation: We conduct an experiment on replacing our epipolar transformer with the visibility-aware multi-view feature aggregation method [8]. The results show that our method outperforms the alternative. This reveals the fact that attention mechanisms are indeed helpful to our multi-view feature aggregation task.

Local-Frustum-Based Context Aggregation: Frustum context aggregation is at the core of the proposed method. To reveal the effectiveness, we conduct several ablation studies by removing either the entire module or some key components. The experiments are conducted on the *ScanNet* dataset. The results are reported in Table VIII. We can make the following conclusions. First, the decline in performance on the baseline without the entire local-frustum-based context aggregation module (*w/o frustum*) indicates the proposed module is necessary. Second,

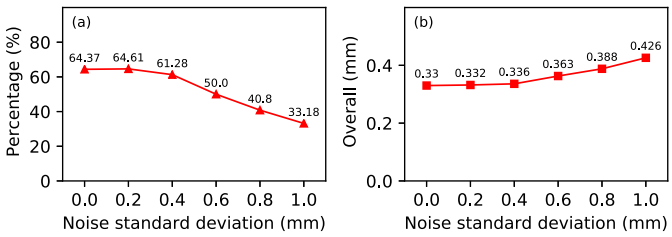

Fig. 18. Sensitivity to coarse depth quality. The percentage of pixel-wise depth predictions whose error is smaller than 1 mm (a) and the overall score of point cloud reconstruction (b) are reported.

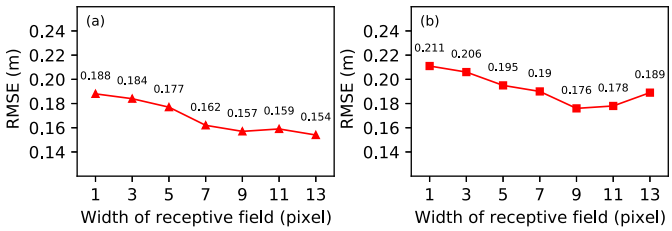

Fig. 19. Sensitivity to the width of the receptive field. The RMSE on the *Textureless* and *Large depth variation* test sets are reported. In general, we found our method is generally robust and not very sensitive to the width. It achieves the top performance when the width of the receptive field $t = 13$ and 9, respectively.

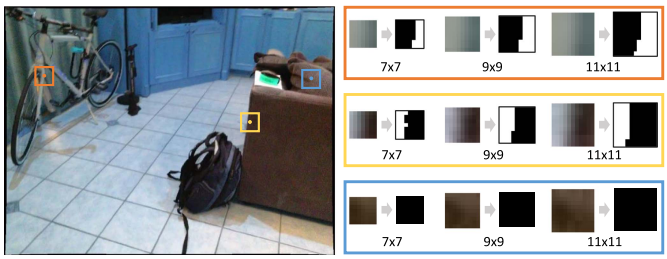

Fig. 20. The effectiveness of the gating unit in the local-frustum-based context aggregation. We see the gating unit is able to successfully select the semantically relevant pixels with different width of the receptive field.

the baseline that uses all context in the receptive field without a selection by the gating unit (*w/o gating unit*) leads to a relatively inferior performance, demonstrating the adaptive context selection is useful. Third, by using the soft decisions during both the forward and backward pass (*w/o Gumbel-Softmax*), the performance drops especially on the $RMSE$ and $p@0.2$ metrics. This is consistent with the conclusions of some recent methods that also unitized the Gumbel-Softmax trick [32], [59], [60].

F. Sensitivity to Coarse Depth Quality

We show our method is robust to the incorrectness of coarse depth prediction by conducting a pressure test. In the experiment, we add Gaussian noise to the predicted coarse depth maps, during both the training and testing phases. We report the performance of the depth map prediction and the point cloud reconstruction on *DTU*. Fig. 18 shows RayMVSNet is robust to moderate perturbation (noise standard deviation ≤ 0.4 mm). It is interesting to see that the quality of depth map prediction slightly

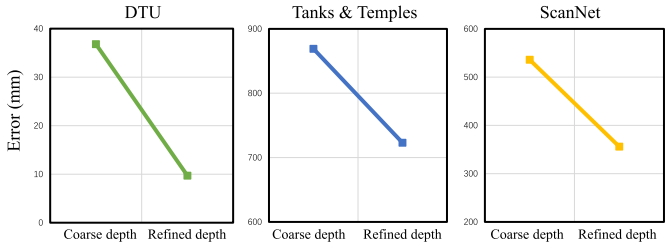

Fig. 21. The errors in depth estimation on the rays whose ground-truth depth is outside the enlarged searching region. We see that ray-based inference is able to improve the accuracy, demonstrating its ability to handle inaccurate coarse depth estimation.

increases when moderate noise is added. This demonstrates that data augmentation such as modest perturbation to coarse depth is helpful for training a more generalizable RayMVSNet. Moreover, we conduct experiments of replacing the MVSNet with other MVSNet variants, e.g., UCS-MVSNet, Fast-MVSNet, and CVP-MVSNet, for coarse depth estimation. We found consistent improvement of depth estimation for the alternative backbones. In particular, our method with a UCS-MVSNet backbone achieves a 0.326 overall score on the DTU dataset, which is slightly better compared to the original RayMVSNet.

G. Sensitivity to Width of Receptive Field

We also test RayMVSNet++ using different widths of the receptive field in the local-frustum-based context aggregation. We train our method on *ScanNet* with different width and test the trained models on the *Textureless* and *Large depth variation* test sets. The RMSE are showed in Fig. 19. When $t = 1$, the method essentially equals to the original RayMVSNet [67]. It shows that the local-frustum-based context aggregation is indeed helpful on *ScanNet* with more challenging examples. In particular, RayMVSNet++ is robust when $t \geq 7$, demonstrating our method is not sensitive to the parameter. It achieves the top performance when $t = 13$ and 9, respectively. It shows that the context across large neighboring pixels is more significant to the depth estimation in the textureless regions. Fig. 20 provides some examples of how the attentional gating unit performs with different widths of the receptive field. We have also tried increasing the width of the receptive field on *DTU*. However, we do not see significant performance improvements. This verifies the idea that the local-frustum-based context aggregation is only helpful to challenging datasets with low-quality images caused by poor lighting conditions or motion blur.

H. Handling Inaccurate Coarse Depth

Despite the conservative parameter settings, a small proportion of the true depth might fall outside of the search space induced by the estimated coarse depth. Although such cases are the minority ($< 3\%$), our method is able to alleviate this problem by estimating the relative location l on the ray. In such cases, during the ray-based inference, the estimated relative location l would be outside the enlarged searching region [0,1]. Since those cases exist in both the training and testing phases, our method is able

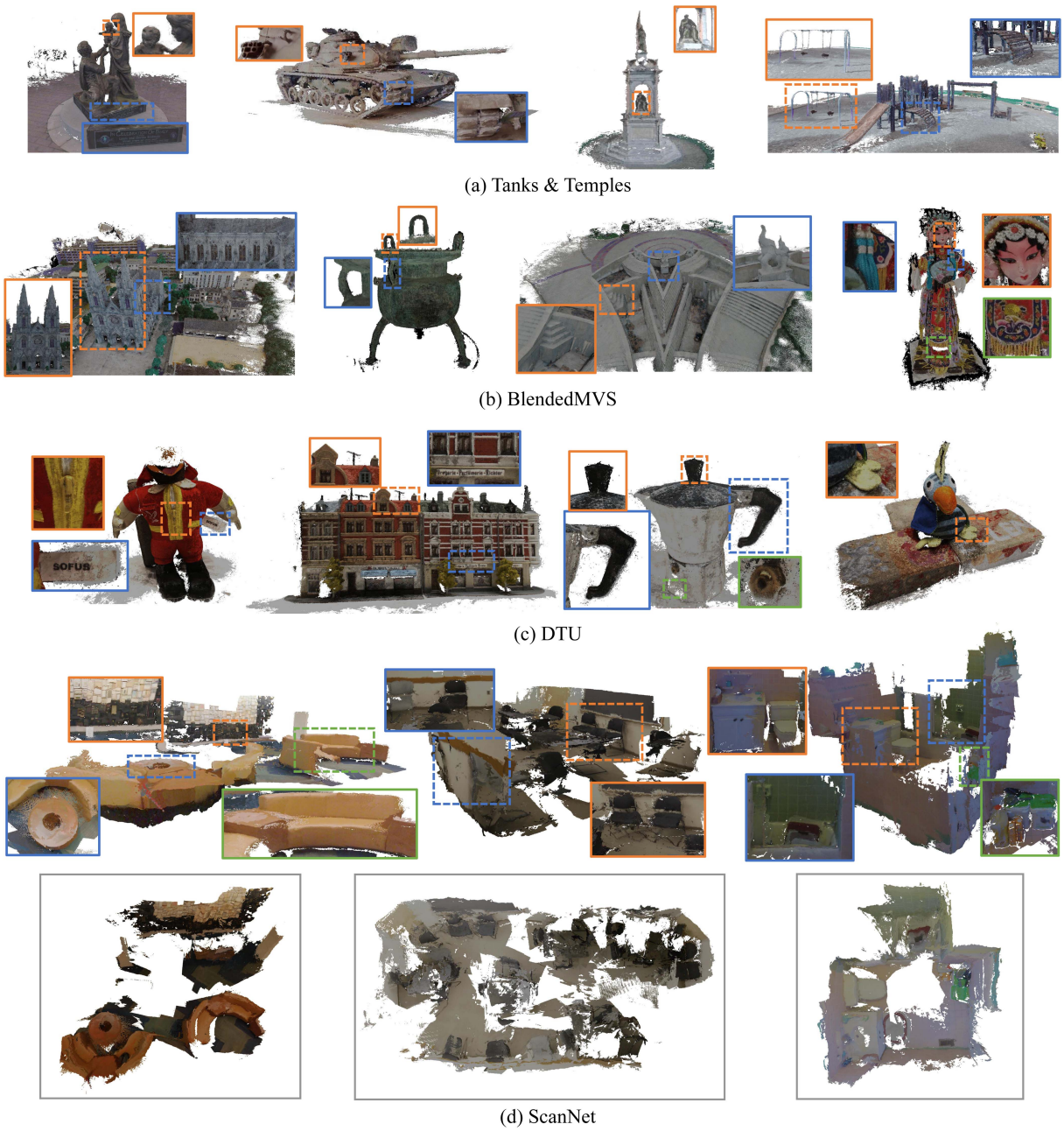

Fig. 22. Gallery of the reconstructed point cloud on (a) *Tanks & temples*, (b) *BlendedMVS*, (c) *DTU*, and (d) *ScanNet*. In (d), the scenes in the rectangles with solid line represents the reconstructed point cloud of the whole scene observed from the top view.

to learn to estimate those by the regression in equation 6. Fig. 21 provides visualizations of the depth estimation accuracy before and after the ray-based inference on the rays whose ground-truth depth is outside the enlarged searching region. We see that our method is able to improve the accuracy, demonstrating its ability to handle inaccurate coarse depth estimation. Note that the errors shown in the figure are determined by both the accuracy of depth estimation itself and the range of the enlarged searching region. We set the range of the enlarged searching region as 20 mm in DTU, 1000 mm in Tanks & Temples, and 600 mm in ScanNet.

I. *Qualitative Results*

We visualize the qualitative results of our method on several datasets in Fig. 22. Note that our method is able to reconstruct large-scale scenes with fine-grained geometry details, such as the highlighted regions.

V. Conclusion and Discussion

We have presented RayMVSNet++, which learns to directly optimize the depth value along each camera ray. An epipolar transformer is designed to enable sequential modeling of 1D

ray-based implicit fields, which essentially mimics the epipolar line search in traditional MVS. The ray-based approach demonstrates significant performance boost with only a low-res cost volume. In particular, a local-frustum-based context aggregation is proposed to extend the receptive field of the ray-based model, leading to more accurate and robust predictions. The method has been demonstrated to be effective on three public datasets, achieving state-of-the-art performance.

Our method has the following limitations. First, although we have demonstrated the method is robust to the coarse depth quality, there is still a small proportion of challenging regions whose depth cannot be accurately estimated due to the large error in the coarse depth prediction. Second, our method relies on accurate camera poses. For scenarios that do not meet this requirement, our method cannot produce accurate outputs, since it cannot optimize the camera pose and the 3D points simultaneously.

An interesting future direction is to further enhance the ray-based deep MVS approach so that cost volume convolution could be completely saved. In most deep MVS works, 3D point cloud is recovered from the estimated depth map as post-processing. Therefore, we would also like to study the end-to-end optimization of 3D point clouds [51]. Moreover, our method assumes the camera poses are given, it is interesting to explore estimating the camera pose [3] and reconstructing scene/object surfaces in a uniform framework, such that the two tasks would boost each other.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments.

References

[1] 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.
[2] A. M. Andrew, "Multiple view geometry in computer vision," *Kybernetes*, vol. 30, pp. 1333–1341, 2001.
[3] Q. Cai, L. Zhang, Y. Wu, W. Yu, and D. Hu, "A pose-only solution to visual reconstruction and navigation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 73–86, Jan. 2023.
[4] 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.*, Springer, 2020, pp. 213–229.
[5] R. Chabra et al., "Deep local shapes: Learning local SDF priors for detailed 3D reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 608–625.
[6] A. Chen et al., "MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14124–14133.
[7] R. Chen, S. Han, J. Xu, and H. Su, "Point-based multi-view stereo network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1538–1547.
[8] R. Chen, S. Han, J. Xu, and H. Su, "Visibility-aware point-based multi-view stereo network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3695–3708, Oct. 2021.
[9] Z. Chen and H. Zhang, "Learning implicit fields for generative shape modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5939–5948.
[10] S. Cheng et al., "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2524–2534.
[11] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5828–5839.
[12] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11065–11074.
[13] Y. Deng, J. Yang, and X. Tong, "Deformed implicit field: Modeling 3D shapes with learned dense correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10286–10296.
[14] Y. Ding et al., "Transmvsnet: Global context-aware multi-view stereo network with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8585–8594.
[15] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–22.
[16] Y. Duan, H. Zhu, H. Wang, L. Yi, R. Nevatia, and L. J. Guibas, "Curriculum DeepSDF," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 51–67.
[17] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 873–881.
[18] K. Genova, F. Cole, A. Sud, A. Sarna, and T. Funkhouser, "Local deep implicit functions for 3D shape," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4857–4866.
[19] 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.*, 2020, pp. 2495–2504.
[20] W. Hartmann, S. Galliani, M. Havlena, L. Van Gool, and K. Schindler, "Learned multi-patch similarity," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1586–1594.
[21] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu, "Epipolar transformer for multi-view human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 1036–1037.
[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[23] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "DeepMVS: Learning multi-view stereopsis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2821–2830.
[24] S. Im, H.-G. Jeon, S. Lin, and I. S. Kweon, "DPSNet: End-to-end deep plane sweep stereo," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–12.
[25] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-softmax," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–12.
[26] 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.*, 2017, pp. 2307–2315.
[27] C. Jiang et al., "Local implicit grid representations for 3D scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6001–6010.
[28] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 364–375.
[29] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2020, pp. 5156–5165.
[30] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–12.
[31] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–13, 2017.
[32] S. Kong and C. Fowlkes, "Pixel-wise attentional gating for scene parsing," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2019, pp. 1024–1033.
[33] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *Proc. Int. Conf. Mach. Learn.*, pp. PMLR, 2019, pp. 3744–3753.
[34] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," 2019, *arXiv: 1907.10326*.
[35] Y. Li et al., "Attention-guided unified network for panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7026–7035.
[36] Y. Liu et al., "Neural rays for occlusion-aware image-based rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7824–7833.
[37] 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.*, 2019, pp. 10452–10461.

[38] K. Luo, T. Guan, L. Ju, Y. Wang, Z. Chen, and Y. Luo, "Attention-aware multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1590–1599.

[39] 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.*, 2021, pp. 5732–5740.

[40] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3D reconstruction in function space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4460–4470.

[41] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 405–421.

[42] Z. Murez, T. Van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "Atlas: End-to-end 3D scene reconstruction from posed images," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K,. Springer, Aug. 23–28, 2020, pp. 414–431.

[43] H. J. Nelson and N. Papanikolopoulos, "Learning continuous object representations from point cloud data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 2446–2451.

[44] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3504–3515.

[45] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "DeepSDF: Learning continuous signed distance functions for shape representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 165–174, 2019.

[46] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, UK, Springer, Aug. 23–28, 2020, pp. 523–540.

[47] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, and K. Xu, "Geometric transformer for fast and robust point cloud registration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11143–11152.

[48] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2304–2314.

[49] 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 Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 519–528.

[50] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," 2018, *arXiv: 1803.02155*.

[51] Y. Shi, J. Huang, H. Zhang, X. Xu, S. Rusinkiewicz, and K. Xu, "SymmetryNet: Learning to predict reflectional and rotational symmetries of 3D shapes from single-view RGB-D images," *ACM Trans. Graph.*, vol. 39, no. 6, pp. 1–14, 2020.

[52] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2012, pp. 746–760.

[53] V. Sitzmann, E. R. Chan, R. Tucker, N. Snavely, and G. Wetzstein, "MetaSDF: Meta-learning signed distance functions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 10136–10147.

[54] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "LoFTR: Detector-free local feature matching with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8922–8931.

[55] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "NeuralRecon: Real-time coherent 3D reconstruction from monocular video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15598–15607.

[56] T. Takikawa et al., "Neural geometric level of detail: Real-time rendering with implicit 3D shapes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11358–11367.

[57] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Stoll, and C. Theobalt, "PatchNets: Patch-based generalizable deep implicit 3D shape representations," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 293–309.

[58] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[59] A. Veit and S. Belongie, "Convolutional networks with adaptive inference graphs," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–18.

[60] T. Verelst and T. Tuytelaars, "Dynamic convolutions: Exploiting spatial sparsity for faster inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2320–2329.

[61] D. Wang, X. Cui, S. Salcudean, and Z. J. Wang, "Generalizable neural radiance fields for novel view synthesis with transformer," 2022, *arXiv:2206.05375*.

[62] 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.*, 2021, pp. 14194–14203.

[63] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 27171–27183.

[64] Q. Wang et al., "IBRNet: Learning multi-view image-based rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4690–4699.

[65] Y. Wei, S. Liu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "NerfingMVS: Guided optimization of neural radiance fields for indoor multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5610–5619.

[66] 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.*, 2021, pp. 6187–6196.

[67] J. Xi, Y. Shi, Y. Wang, Y. Guo, and K. Xu, "RayMVSNet: Learning ray-based 1D implicit fields for accurate multi-view stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8595–8605.

[68] H. Xu et al., "Digging into uncertainty in self-supervised multi-view stereo," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6078–6087.

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

[70] Q. Xu, W. Wang, D. Ceylan, R. Mech, and U. Neumann, "DISN: Deep implicit surface network for high-quality single-view 3D reconstruction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 492–502.

[71] J. Yan et al., "Dense hybrid recurrent multi-view stereo net with dynamic consistency checking," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 674–689.

[72] 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.*, 2020, pp. 4877–4886.

[73] Z. Yang, Z. Ren, Q. Shan, and Q. Huang, "MVS2D: Efficient multi-view stereo via attention-driven 2D convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8574–8584.

[74] 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.*, 2018, pp. 767–783.

[75] 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.*, 2019, pp. 5525–5534.

[76] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 4805–4815.

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

[78] A. Yu et al., "Attention aware cost volume pyramid based multi-view stereo network for 3D reconstruction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 448–460, 2021.

[79] 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.*, 2020, pp. 1949–1958.

[80] J. Zhang, Y. Yao, S. Li, Z. Luo, and T. Fang, "Visibility-aware multi-view stereo network," in *Proc. Brit. Mach. Vis. Conf.*, 2020, pp. 1–12.

[81] J. Zhang, Y. Yao, and L. Quan, "Learning signed distance field for multi-view surface reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6525–6534.

[82] K. Zhang, M. Liu, J. Zhang, and Z. Dong, "PA-MVSNet: Sparse-to-dense multi-view stereo with pyramid attention," *IEEE Access*, vol. 9, pp. 27908–27915, 2021.

[83] K. Zhang, G. Riegler, N. Snavely, and V. Koltun, "NeRF++: Analyzing and improving neural radiance fields," 2020, *arXiv: 2010.07492*.

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

[85] Y. Zhao, K. Xu, E. Zhu, X. Liu, X. Zhu, and J. Yin, "Triangle lasso for simultaneous clustering and optimization in graph datasets," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 8, pp. 1610–1623, Aug. 2019.

[86] Z. Zheng, T. Yu, Q. Dai, and Y. Liu, "Deep implicit templates for 3D shape representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1429–1439.

[87] J. Zhu, B. Peng, W. Li, H. Shen, Z. Zhang, and J. Lei, "Multi-view stereo with transformer," 2021, *arXiv:2112.00336*.

Yifei Shi (Member, IEEE) received the PhD degree in computer science from NUDT in 2019. He is an associate professor with the College of Intelligence Science and Technology, National University of Defense Technology (NUDT). During 2017-2018, he was a visiting student research collaborator at Princeton University, advised by Thomas Funkhouser and Szymon Rusinkiewicz. His research interests mainly include computer vision, computer graphics, especially on object/scene analysis and manipulation by machine learning and geometric processing techniques. He has published 20+ papers in top-tier conferences and journals, including CVPR, ECCV, ICCV, SIGGRAPH Asia, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, and *ACM Transactions on Graphics*.

Junhua Xi received the master's degree from NUDT in 2013. She is currently working toward the PhD degree with the College of Computer Science and Technology, NUDT. Her research interests include 3D vision and robotics, especially on object analysis, multi-view stereo, and large-scale scene reconstruction.

Dewen Hu (Senior Member, IEEE) received the BSc and MSc degrees from Xi'an Jiaotong University, China, in 1983 and 1986, respectively, and the PhD degree from the National University of Defense Technology, in 1999. In 1986, he was with the National University of Defense Technology. From October 1995 to October 1996, he was a visiting scholar with The University of Sheffield, U.K. In 1996, he was promoted as a professor. He has authored more than 200 articles in journals, such as the Brain, *Proceedings of the National Academy of Sciences of the United States of America*, *NeuroImage*, *Human Brain Mapping*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Signal Processing*, *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Medical Imaging*, and *IEEE Transactions on Biomedical Engineering*. His research interests include pattern recognition and cognitive neuroscience. He is currently an associate editor of *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Zhiping Cai (Member, IEEE) received the BEng, MASc, and PhD degrees in computer science and technology from the National University of Defense Technology (NUDT), China, in 1996, 2002, and 2005, respectively. He is a full professor with the College of Computer, NUDT. His current research interests include artificial intelligence, network security, and Big Data. He is a senior member of the CCF.

Kai Xu (Senior Member, IEEE) received the PhD degree from the College of Computer, NUDT, in 2011. He is a professor with the College of Computer, NUDT. He conducted visiting research with Simon Fraser University and Princeton University. His research interests include geometric modeling and shape analysis, especially on data-driven approaches to the problems in those directions, as well as 3D vision and its robotic applications. He has published more than 80 research papers, including 20+ SIGGRAPH/TOG papers. He has co-organized several SIGGRAPH Asia courses and Eurographics STAR tutorials. He serves on the editorial board of *ACM Transactions on Graphics*, *Computer Graphics Forum*, *Computers & Graphics*, and *Visual Computer*. He also served as program co-chair of CAD/Graphics 2017, ICVRV 2017 and ISVC 2018, as well as PC member for several prestigious conferences including SIGGRAPH, SIGGRAPH Asia, Eurographics, SGP, PG, etc.