This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

VolumeFusion: Deep Depth Fusion for 3D Scene Reconstruction

Jaesung Choe
KAIST

Sunghoon Im
DGIST

Francois Rameau
KAIST

Minjun Kang
KAIST

In So Kweon
KAIST

Abstract

To reconstruct a 3D scene from a set of calibrated views, traditional multi-view stereo techniques rely on two distinct stages: local depth maps computation and global depth maps fusion. Recent studies concentrate on deep neural architectures for depth estimation by using conventional depth fusion method or direct 3D reconstruction network by regressing Truncated Signed Distance Function (TSDF). In this paper, we advocate that replicating the traditional two stages framework with deep neural networks improves both the interpretability and the accuracy of the results. As mentioned, our network operates in two steps: 1) the local computation of the local depth maps with a deep MVS technique, and, 2) the depth maps and images' features fusion to build a single TSDF volume. In order to improve the matching performance between images acquired from very different viewpoints (e.g.*, large-baseline and rotations), we introduce a rotation-invariant 3D convolution kernel called PosedConv. The effectiveness of the proposed architecture is underlined via a large series of experiments conducted on the ScanNet dataset where our approach compares favorably against both traditional and deep learning techniques.*

1. Introduction

Multi-view stereo (MVS) is a fundamental research topic that has been extensively investigated over the past decades [18]. The main goal of MVS is to reconstruct a 3D scene from a set of images acquired from different viewpoints. This problem is commonly framed as a correspondence search problem by optimizing photometric or geometric consistency among groups of pixels in different images. Conventional non-learning-based MVS frameworks [15, 2, 13] generally achieve the reconstruction using various 3D representations [12]: depth maps [15, 2], point cloud [13], voxels [29, 19], and meshes [11].

The recent use of deep neural networks for MVS [39, 20, 22, 32, 43, 4, 16] has proven effective in addressing the limitations of traditional techniques like repetitive patterns, low-texture regions, and reflections. Usually, deep-based MVS

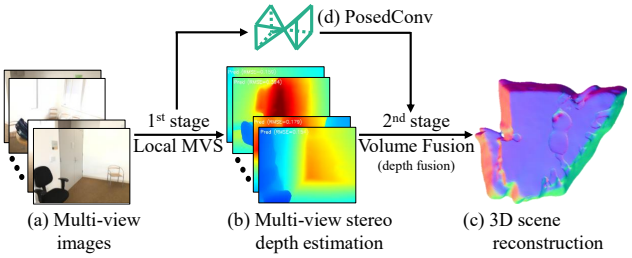


Figure 1. **Volume fusion.** Given (a) multi-view images and their camera parameters, our network aims at 3D scene reconstruction. (b) First, we estimate local multi-view depth maps. (b) Second, we introduce differentiable depth fusion with the guidance of pose-invariant features from (d) our PosedConv.

methods [43, 4, 16] center on the estimation of the pixel-wise correspondence between a reference image and its surrounding views. While this strategy can be elegantly integrated into deep learning frameworks, they can only work locally where frames largely overlap. For the full 3D reconstruction of the entire scene, these methods [43, 4, 16] require to perform a depth map fusion [31, 14] to merge local reconstructions as post-processing.

More recently, Murez *et al*. [32] suggest that a direct regression of the Truncated Signed Distance Function (TSDF) volume of the scene is more effective than using intermediate 3D representations, *i.e*. depth maps. The overall concept of their technique consists in the back-projection of all the extracted image features into a global scene volume from which the network directly regresses the TSDF volume. This end-to-end approach [32] has the advantage of being straightforward and scalable to a large number of images. However, this pioneering work [32] has difficulty in shaping the global structure of complex scenes, such as the corners of rooms or long hallways.

To address this issue, we propose to closely mimic the traditional 3D reconstruction pipeline with two distinct stages: local reconstruction and global fusion. However, unlike the previous studies [43, 10, 32] and concurrent papers [38, 1], we integrate these two stages in an end-to-end manner. First, our network computes the local geometry, *i.e*., dense depth maps from neighboring frames. Then,

we begin the depth fusion process by merging local depth maps as well as image features in a single volume representation where our network regresses TSDF for the final 3D scene reconstruction. This enables our end-to-end framework to learn a globally consistent volumetric representation in a single forward computation without the need for manually engineered fusion algorithms [37, 14]. To further enhance the robustness of our depth fusion mechanism, we propose the *Posed Convolution Layer* (Posed-Conv). Compared to the traditional 3D convolution layer that is solely invariant to translation, we propose a more versatile convolution layer that is invariant to both translation and rotation. In short, our Posed Convolution Layer helps to extract pose-invariant feature representation regardless of the orientation of the input image. As a result, our method demonstrates globally consistent shape reconstruction even under wide baselines or large rotations between the views. Our contributions are summarized as follows:

- A novel network taking advantage of local MVS and global depth fusion for 3D scene reconstruction.
- A new rotation and translation invariant convolution layer, called *PosedConv*.

2. Related Works

2.1. Multi-view Stereo

Multi-View Stereo (MVS) consists in the pixel-wise 3D reconstruction of a scene given a set of unstructured images along with their respective intrinsic and extrinsic parameters. In [12], the scene representation is used as an axis of taxonomy to categorize MVS into four sub-fields of research: depth maps [15, 2], point clouds [13, 28], voxels [29, 19], and meshes [11]. In particular, The depth map estimation approaches [15, 2, 21, 17] have been widely researched since these strategies are easily scalable to number of multi-view images. These methods usually rely on a small baseline assumption to ensure large overlap with a reference frame. Thus, the photometric matching is achieved via a plane-sweeping algorithm [7, 42, 21, 17] where the most probable depth is estimated for each pixel in the reference frame. Then, a depth map fusion algorithm [19, 14] is required to build a global 3D model from a set of depths.

With the rise of deep neural networks, learning-based MVS methods have achieved promising results. Inspired by stereo matching networks [26, 3], MVS studies [43, 4, 16, 22, 20, 39] have developed cost volume for unstructured multi-view matching. Relying on basic frameworks, such as DPSNet [22] or MVSNet [43], follow-up research proposes point-based depth refinement [4], cascaded depth refinement [16], and temporal fusion network [20, 10]. After exhaustively estimating a collection of depth maps, depth fusion [14, 8] starts to reconstruct the global 3D scene.

2.2. Depth Map Fusion

In their seminal work, Curless and Levoy [8] propose a volumetric depth map fusion approach able to deal with noisy depth maps through cumulative weighted signed distance function. Follow-up research, such as KinectFusion [23] or voxel hashing [35, 25], concentrate on the problem of volumetric representation via depth maps fusion. Recently, with the help of deep learning networks, learning-based volumetric approaches [24, 36, 40] have been proposed. For instance, SurfaceNet [24] and RayNet [36] infer the depth maps from multi-view images and their known camera poses. These methods are close to our strategy, but their networks are trained only by a per-view depth map which is not directly related to depth maps fusion. More recently, RoutedFusion [40] and Neural Fusion [41] introduce a new learning-based depth map fusion using RGB-D sensors. However, these papers [40, 41] concentrate on depth fusion algorithm using noisy and uncertain depth maps from multi-view images, not RGB-D sensors. To the best of our knowledge, our method is the first learning-based depth maps fusion from multi-view images for 3D reconstruction.

3. Volume Fusion Network

In this work, we present a novel strategy to effectively reconstruct 3D scenes from an arbitrary set of images $\mathcal{I}$ where the cameras' parameters for each frame are assumed to be known. The cameras' parameters include the intrinsic matrix $\mathbf{K}$ and the extrinsic parameters (rotation matrix $\mathbf{R}\in\mathbb{R}^{3\times3}$ and translation vector $\mathbf{t}\in\mathbb{R}^{3\times1}$). To achieve this reconstruction, we design a novel volume fusion network that consists of two stages: First, each image in the set is processed through our depth network (Sec. 3.1). Based on a pixel-wise photometric matching between adjacent frames, this network regresses both a depth map $\tilde{\mathcal{Z}}$ and an overlapping mask $\tilde{\mathcal{M}}$ for every frame. Second, using the resulting per-frame depth maps and overlapping masks, we formulate a depth fusion process as a volume fusion (Sec. 3.2). We further enhance the features extracted from each posed image through our PosedConv (Sec. 3.3). The overall architecture is presented as in Fig. 2.

3.1. Multi-view Stereo

In this section, we describe the first stage of our approach: the local MVS network using three neighbor frames $\{\mathcal{I}_{n-1}, \mathcal{I}_n, \mathcal{I}_{n+1}\}$. Following previous studies [22, 27], we construct an initial cost volume $\mathcal{V}_n^{\mathcal{I}}$ using the neighbor image feature maps $\{\mathcal{F}_{n-1}, \mathcal{F}_n, \mathcal{F}_{n+1}\}$ to infer a depth map in the reference camera view ($\mathcal{I}_n$). In contrast to the previous studies [43, 4, 16] that only concentrate on accurate depth estimation, our network additionally infers an overlapping mask. The overlapping mask $\tilde{\mathcal{M}}_n$[1] is computed

[1] We visualize an overlapping mask in the supplementary material.

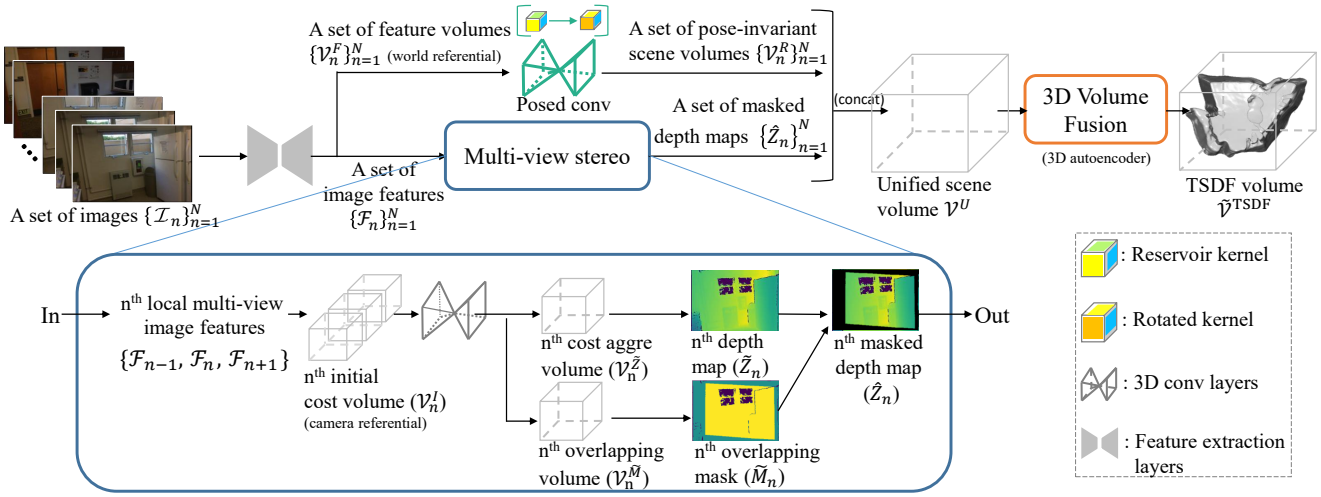


Figure 2. **Overall architecture.** Our volume fusion network consists in two stages: multi-view stereo and Volumetric Depth Fusion. Given a set of N images $\{\mathcal{I}_n\}_{n=1}^N$, we extract image features $\{\mathcal{F}_n\}_{n=1}^N$. These image features are used (1) to infer a masked depth map $\hat{\mathcal{Z}}_n$ in the multi-view stereo stage, and (2) to extract pose-invariant feature volume $\{\mathcal{V}_n^R\}_{n=1}^N$ in *PosedConv*. Then, we construct a unified scene volume $\mathcal{V}^U$ from $\{\mathcal{V}_n^R\}_{n=1}^N$ and $\{\hat{\mathcal{Z}}_n\}_{n=1}^N$. Finally, in volume fusion stage, we obtain a TSDF volume $\tilde{\mathcal{V}}^{\text{TSDF}}$ which is the full 3D scene reconstruction results.

to quantify the probability of per-pixel overlap among three adjacent frames in a referential camera view. The purpose of the overlapping mask is to filter out the uncertain depth values that we cannot geometrically deduce, *i.e.*, without correspondence across neighbor frames.

In details, we compute matching cost by computing the initial cost volume $\mathcal{V}_{\text{n}}^{\mathcal{I}}$ through stacked hourglass networks [33, 34, 3]. Then, we obtain a cost aggregated volume[2] $\mathcal{V}_{\text{n}}^{\tilde{\mathcal{Z}}} \in \mathbb{R}^{1\times \text{D}\times \text{H}\times \text{W}}$ and an overlapping volume $\mathcal{V}_{\text{n}}^{\tilde{\mathcal{M}}} \in \mathbb{R}^{2\times \text{D}\times \text{H}\times \text{W}}$. These volumes are used for the inference of a depth map $\tilde{\mathcal{Z}} \in \mathbb{R}^{\text{H}\times \text{W}}$ and an overlapping mask $\tilde{\mathcal{M}} \in \mathbb{R}^{\text{H}\times \text{W}}$. Regarding the estimation of the overlapping mask, we formulate its estimation as a binary classification problem (*i.e.*, overlap/non-overlap). First, the overlapping probability $\mathcal{P}_{\text{n}}^{\tilde{\mathcal{M}}} \in \mathbb{R}^{1\times \text{D}\times \text{H}\times \text{W}}$ of the overlapping volume $\mathcal{V}_{\text{n}}^{\tilde{\mathcal{M}}}$ is obtained by a softmax operation. Then, the overlapping mask $\tilde{M}_{\text{n}}$ can be directly estimated by a max-pooling operation along the depth axis of the probability volume. Specifically, the overlapping probability $\tilde{\mathcal{M}}_{u,v}$ at the pixel location (u, v) can be estimated as follow:

$$\tilde{\mathcal{M}}_{u,v} = \text{MaxPool}_{d\in D}\,(\mathcal{P}_{\text{n}}^{\tilde{\mathcal{M}}}[d, v, u]), \tag{1}$$

where $\mathcal{P}_{\text{n}}^{\tilde{\mathcal{M}}}[d, v, u]$ is the probability of the overlapping volume at the voxel $[d, v, u]$. To learn the overlapping mask, a per-pixel L1-Loss between the ground-truth and the estimated mask is employed:

$$\mathcal{L}_M = \sum_{(u,v)\in\tilde{\mathcal{M}}} \|\tilde{\mathcal{M}}_{u,v} - \mathcal{M}_{u,v}\|_1, \tag{2}$$

[2]We represent a 3D dimensional volume with its channels as four axes: channel (C), depth (D), height (H), and width (W).

where $\mathcal{L}_M$ is an overlapping loss and $\mathcal{M}_{u,v}$ is a true overlapping mask $\mathcal{M}$ at pixel (u, v). Note that the overlapping mask is an essential element for our depth map fusion stage (see Sec. 3.2).

Apart from the overlapping mask, the local depth map $\tilde{z}_{\text{n}}$ is also compute from the cost aggregated volume $\mathcal{V}_{\text{n}}^{\tilde{\mathcal{Z}}}$. The depth estimation $\tilde{z}_{u,v}$ at pixel (u, v) is computed as follows:

$$\tilde{z}_{u,v} = \sum_{d=1}^{\text{D}} \frac{z_{\text{min}} \times D}{d} \cdot \sigma(\mathbf{a}_{u,v}^d), \tag{3}$$

where z_{min} is a hyper-parameter defining the minimum range of depth estimation, $\sigma(\cdot)$ represents the softmax operation, $\mathbf{a}_{u,v}^d$ is the d-th plane value of the vector at the pixel (u, v) in the cost aggregated volume $\mathcal{V}_{\text{n}}^{\tilde{\mathcal{Z}}}$. We set the hyper-parameters as D=48 and z_{min}=0.5 meter in our experiments. We compute the depth loss $\mathcal{L}_{\mathcal{Z}}$ from the estimated depth map $\tilde{\mathcal{Z}}$ and a true depth map $\mathcal{Z}$ as follows:

$$\mathcal{L}_{\mathcal{Z}} = \sum_u \sum_v \mathcal{M}_{u,v} \times \text{smooth}_{\text{L}_1}(z_{u,v} - \tilde{z}_{u,v}), \tag{4}$$

where $\tilde{z}_{u,v}$ is the value of the predicted depth map $\tilde{\mathcal{Z}}$ at pixel (u, v), and $\text{smooth}_{\text{L}_1}(\cdot)$ is the smooth L1-loss function. Note that we mask out the depth map with the ground-truth mask $\mathcal{M}$ to impose the loss only on the pixels where the correspondence among the neighbor views exists.

3.2. Volumetric Depth Fusion

In the previous section, we describe the first stage of our method specialized in regressing a per-view depth map $\tilde{\mathcal{Z}}_{\text{n}}$ and an overlapping mask $\tilde{\mathcal{M}}_{\text{n}}$ to obtain a

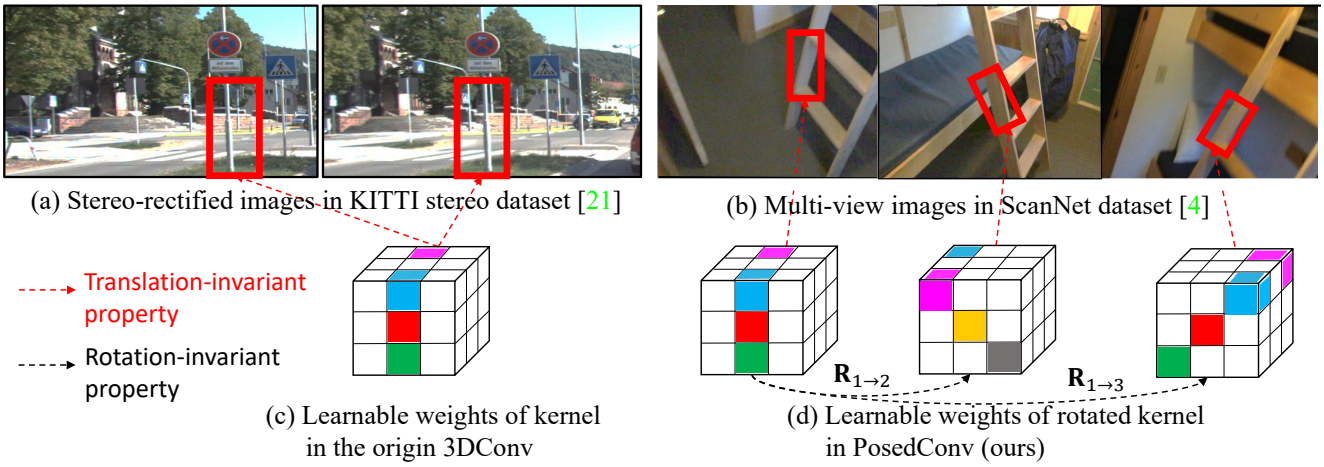


Figure 3. **Illustration of origin 3DConv and our PosedConv layer.** Compared to (a) the stereo-rectified images [30], (b) multi-view images [9] describe the identical objects (red boxes) in different viewpoints. Accordingly, (c) the original 3DConv can be used in cost volumes [3, 44] to compute the matching cost among the stereo-rectified images. (d) Our PosedConv has both rotation-invariant and translation-invariant property that properly extracts features from differently posed images for the robust matching in the world-referential coordinate (*i.e.*, unified scene volume $\mathcal{V}^U$).

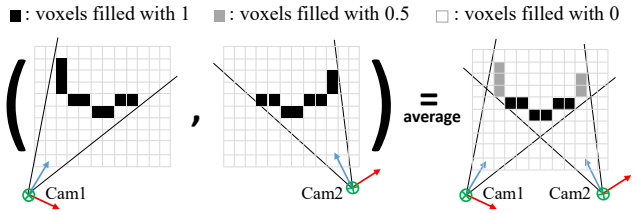


Figure 4. **Top-view illustration of the volume fusion using two depth maps.**

masked depth map $\hat{\mathcal{Z}}_n$. To integrate these local estimations $\{\hat{\mathcal{Z}}_n\}_{n=1}^N$, we design a volume fusion that embeds a set of masked depth maps $\{\hat{\mathcal{Z}}_n\}_{n=1}^N$ into a unified scene volume $\mathcal{V}^U \in \mathbb{R}^{(C+1)\times V_x \times V_y \times V_z}$. This stage aims at regressing the TSDF of a global scene by fusing local geometric information, *i.e.*, $\{\hat{\mathcal{Z}}_n\}_{n=1}^N$.

The traditional fusion process [37, 14] attempts to achieve this reconstruction by using the photometric consistency check of the pixel values that are back-projected using the inferred depth maps [14, 31]. However, these approaches suffer from the following drawbacks: 1) the quality of the final reconstruction depends on the accuracy of the initial depth maps, and 2) the brightness consistency assumption is violated under challenging conditions such as changing lighting conditions and homogeneously textured regions. Alternatively, recent learning-based approaches have been developed, such as RoutedFusion [40]. This approach produces high-quality 3D reconstruction but also requires reliable depth maps from RGB-D sensors. However, it is hardly suited for depth maps obtained via multi-view images which tend to be significantly more noisy.

To overcome these issues, our depth fusion strategy propagates the masked depth maps $\{\hat{\mathcal{Z}}_n\}_{n=1}^N$ as well as the feature volumes $\mathcal{V}^{\mathcal{F}}$, which are computed from the image features $\mathcal{F}_n$ by back-projecting the image features into the world-referential coordinate system. This strategy allows the network to re-profile the surface of the 3D scene by computing the matching cost of the image features guided by the masked depth maps. To fuse the per-view masked depth maps $\hat{\mathcal{Z}}$, we iteratively compute the per-view masked depth map and image feature. First, we declare a 3D volume following [32] and initialize all voxels with 0. Then, we back-project a masked depth map and compute their voxel location $[i, j, k]^\top$. The value of each voxel occupied by a back-projected depth map is incremented by 1. We repeat this process for all views, and then average the volume as shown in Fig. 4. This strategy allows to compute the 3D surface probability using all the previously computed depths.

The embedded voxels in the unified scene volume $\mathcal{V}^U$ are the initial guidance for shaping the global structure of the target scene through volume fusion. To further enhance the geometric property of each embedded image feature, we introduce the concept of the Posed Convolution Layer (PosedConv) for an accurate reconstruction of the target scene.

3.3. Posed Convolution Layer

Renowned stereo matching approaches [26, 3, 6, 5] employ a series of 3D convolution layers (3DConv) to find dense correspondences between the left-right image features. Since this strategy proved to be effective, it became the most commonly used technique to build a cost volume. As a result, it has also been widely applied in un-rectified multi-view stereo pipelines [20, 22, 10]. Nonetheless, it ap-

pears that this representation is not appropriate for multi-view stereo. To understand this problem, we need to analyze both configurations: calibrated stereo pair and un-calibrated multi-view stereo. For calibrated stereo images, corresponding patches in both views are acquired from the same orientation (aligned optical axis); hence, the translation-invariant convolution operation is suitable to find consistent matches (see Fig. 3-(a)). However, MVS scenarios are more complex since images can be acquired from various orientations; therefore, the 3DConv is not appropriate since it is not rotation invariant (see Fig. 3-(b)). This phenomenon leads to poor matching quality when the orientations of the different viewpoints vary too greatly.

To cope with this limitation, we design rotation-dependent 3D convolution kernels, called PosedConv, using the known rotation matrix as shown in Fig. 3. First, we declare that the variable for our posed convolution layer which is named reservoir kernel $\mathbf{W} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times w \times w \times w}$ is similar to that of naive 3DConv. The parameters C_{out} and C_{in} denote the number of output and input channels, respectively, and the odd number w represents the size of the 3D spatial window. Assuming that the reservoir kernel is aligned with the unified scene volume $\mathcal{V}^U$, we compute the rotated kernel $\mathbf{W}_{\text{n}}^R$ to be aligned to each corresponding camera viewpoint using the rotation matrix $\mathbf{R}_{1\to\text{n}}$[3]. The rotated kernel extracts more consistent features than the naive 3DConv since the proposed kernel is more robust to maintain the same receptive field even under rotation changes. Consequently, a pose-invariant feature volume $\mathcal{V}_{\text{n}}^R$ are computed by convolving the rotated kernels $\mathbf{W}_{\text{n}}^R$ with the feature volume $\mathcal{V}_{\text{n}}^{\mathcal{F}}$ as:

$$\begin{aligned}\mathcal{V}_{\text{n}}^R &= (\mathcal{V}_{\text{n}}^{\mathcal{F}} * \mathbf{W}_{\text{n}}^R)(\mathbf{v}) \\ &= \sum_{\mathbf{v}'\in\Omega} \mathcal{V}_{\text{n}}^{\mathcal{F}}(\mathbf{v}+\mathbf{v}') \cdot \mathbf{W}_{\text{n}}^R(\mathbf{v}') \\ &= \sum_{\mathbf{v}'\in\Omega} \mathcal{V}_{\text{n}}^{\mathcal{F}}(\mathbf{v}+\mathbf{v}') \cdot \mathbf{W}(D_R \cdot \mathbf{v}'),\end{aligned} \tag{5}$$

where $\mathbf{v}=[i,j,k]^\intercal$ is the voxel coordinates in world referential, and $\Omega=\{\mathbf{v}'\in\mathbb{Z}^3 | [\text{-}w', \text{-}w', \text{-}w']^\intercal, ..., [+w', +w', +w']^\intercal\}$ is a set of signed distances from the center of the reservoir kernel to each voxel $\mathbf{v}$ in the kernel where $w'=(w\text{-}1)/2$. The dot operator $\cdot$ indicates dot product, and D_R represents the modified rotation matrix through our *Discrete Kernel Rotation*[4]. Then, we calculate the part of the unified scene volume $\mathcal{V}^U$ by averaging a set of the pose-invariant feature volume $\{\mathcal{V}_{\text{n}}^R\}_{\text{n=1}}^N$. Our PosedConv, elaborately designed for varying camera orientation, plays an important role in building the unified scene volume $\mathcal{V}^U$ and robust 3D scene reconstruction.

[3]We set the first camera coordinate as the world coordinate.

[4]The details of D_R is available in the supplementary material.

3.4. Volumetric 3D Reconstruction

The unified scene volume $\mathcal{V}^U$ is aggregated through stacked hourglass 3DConv layers [33, 34, 3] to compute the TSDF of the entire scene. After the aggregation process, we obtain a TSDF volume $\tilde{\mathcal{V}}^{\text{TSDF}} \in \mathbb{R}^{V_{\text{x}} \times V_{\text{y}} \times V_{\text{z}}}$, as shown in Fig. 2. The estimated TSDF volume $\tilde{\mathcal{V}}^{\text{TSDF}}$ involves the value of truncated signed distance from the surface, and it is trained in a supervised manner as follows:

$$\mathcal{L}_{\text{TSDF}} = \sum_{(x,y,z)} \left| \tilde{\mathcal{V}}_{x,y,z}^{\text{TSDF}} - \mathcal{V}_{x,y,z}^{\text{TSDF}} \right|_1, \tag{6}$$

where $\mathcal{V}_{x,y,z}^{\text{TSDF}}$ is the ground-truth TSDF volume and $|\cdot|_1$ is the absolute distance measurements (*i.e.* L1-loss), and (x, y, z) is the voxel location within the TSDF volume.

Finally, our network is trained in an end-to-end manner with the following three different losses:

$$\mathcal{L}_{tot} = \alpha\mathcal{L}_{\mathcal{Z}} + \beta\mathcal{L}_{\mathcal{M}} + \gamma\mathcal{L}_{\text{TSDF}} \tag{7}$$

where α, β, and γ are 1.0, 0.5, and 2.0, respectively. The depth loss $\mathcal{L}_{\mathcal{Z}}$ and the overlapping loss $\mathcal{L}_{\mathcal{M}}$ guide the network to perform local multi-view stereo matching for explicit depth estimation. The TSDF loss $\mathcal{L}_{\text{TSDF}}$ is intended to transform the explicit geometry depth maps and the per-view image features into an implicit representation through our volume fusion.

4. Experiment

4.1. Implementation Details and Dataset

Following Murez *et al.* [32], we conduct the assessment of our method on the ScanNet dataset [9]. The ScanNet dataset consists of 800 indoor scenes. Each scene contains a sequence of RGB images, the corresponding ground-truth depth maps, and the cameras' parameters. Among them, 700 scenes are used for training while the remaining 100 scenes constitute our testing set. To obtain the ground-truth TSDF volume $\mathcal{V}^{\text{TSDF}}$, we follow the original scheme proposed by a previous study [32].

In the first stage of our network, we use three local views per frame to operate the local multi-view stereo matching. The input size of the image is 480(H)$\times$$640$(W) and the resolution of the estimated depth is 120(H)$\times$$160$(W). The size of the overlapping mask is identical to that of the depth map. After PosedConv extracts pose-invariant scene volume $\mathcal{V}_{\text{n}}^R$, we integrate the per-view information ($\{\hat{\mathcal{Z}}_{\text{n}}\}_{\text{n=1}}^N$ and $\{\mathcal{V}_{\text{n}}^R\}_{\text{n=1}}^N$) into the unified scene volume $\mathcal{V}^U$. Finally, the second stage of our network is trained in a supervised manner via the TSDF loss $\mathcal{L}_{\text{TSDF}}$. The unified scene volume $\mathcal{V}^U$ and the TSDF volume $\mathcal{V}^{\text{TSDF}}$ covers $0.04m$ per voxel. The resolution of the unified scene volume $\mathcal{V}^U \in \mathbb{R}^{(C+1)\times V_{\text{x}} \times V_{\text{y}} \times V_{\text{z}}}$ is $160(V_{\text{x}})\times64(V_{\text{y}})\times160(V_{\text{z}})$

Method	2D Depth Evaluation				3D Geometry Evaluation			
	AbsRel	AbsDiff	SqRel	RMSE	$\mathcal{L}_1$	Acc	Comp	F-score
COLMAP [37]	.137	.264	.138	.502	.599	.069	.135	**.558**
MVDepthNet [39]	.098	.191	.061	.293	.518	.040	.240	.329
GPMVS [20]	.130	.239	.339	.472	.475	**.031**	.879	.304
DPSNet [22]	.087	.158	.035	.232	.421	.045	.284	.344
Murez *et al*. [32]	.061	.120	.042	.248	.162	.065	.130	.499
VolumeFusion (ours)	**.049**	**.084**	**.021**	**.164**	**.141**	.038	**.125**	.508

Table 1. **Quantitative results on ScanNet dataset [9].** We provide two metrics: depth evaluation and 3D geometry evaluation.

for training and $416(V_x)\times128(V_y)\times416(V_z)$. We set the initial learning rate as 0.0001 and drop the learning rate by half after 50 epochs. We train our network for 100 epochs with 8 NVIDIA RTX 3090 GPUs, which takes about two days.

4.2. Comparison with State-of-the-art Methods

To validate the effectiveness of the proposed approach, we compare the reconstruction performance to a variety of traditional geometry-based and deep learning-based methods [37, 20, 22, 39, 32] in both 3D space and the 2.5D depth map domain. Specifically, we use the four common quantitative measures (AbsRel, AbsDiff, SqRel, and RMSE)[5] of depth map quality and three common criteria ($\mathcal{L}_1$, Accuracy(Acc), Completeness(Comp), and F-score)[5] on 3D reconstruction quality. The quantitative results are reported in Table 1. The evaluation is conducted with the 100 scenes from the ScanNet [9] test set following the evaluation pipeline described in Murez *et al*. [32]. As our experimental results show, the proposed method outperforms all competitors for all evaluation metrics of depth map by a large margin. We conjecture that the significant performance gap results from our depth map fusion method, which effectively matches multiple images even at a large viewpoint difference and increases the number of observations for matching.

In the evaluation of 3D reconstruction in Table 1, our approach outperforms the others especially on the $\mathcal{L}_1$ and Comp. It suggests that our method is outstanding for shaping the global structure of the scene. For the Acc metric, our method demonstrates the second-best results after GPMVS [20], which shows that temporal fusion [20] is a potential method to improve the quality of the multi-view depth map. However, for the 3D reconstruction of the overall scene ($\mathcal{L}_1$, Comp), our method largely outperforms the temporal fusion method [20]. The best F-score is obtained by COLMAP [37]. As described in DPSNet [22], COLMAP [37] has the strength to accurately reconstruct the edges or corners where distinctive features are extracted.

Additionally, we show qualitative results of the depth fusion results in Fig. 6, and the 3D scene reconstruction in Fig. 5. Compared to the most recent competitive technique [32], our method better preserves the global structure of the scene, especially for the shapes of complex rooms – *e.g*. with corridors. Moreover, concerning the quality of depth estimation, our method shows improved depth accuracy after fusion through volume fusion. We attribute the performance improvement to our two-stage approach that exploits the pose-invariant features that re-profile the surface of the scene with the robust matching in a world-referential coordinate, *i.e*., the unified scene volume $\mathcal{V}^U$).

[5] We provide a detailed description of these metrics in our supplementary material

Method	Evaluation			
	AbsRel	RMSE	$\mathcal{L}_1$	F-score
3D Conv	.058	.231	.166	.460
PosedConv	**.049**	**.164**	**.141**	**.508**

Table 2. **Ablation study of PosedConv.** We compare original convolution layer (3D Conv), and our PosedConv.

Method	Preserve		Evaluation	
	Posed Conv	Depth Fusion	RMSE	$\mathcal{L}_1$
Single-stage [32]			.248	.162
w/o depth	✓		.236	.159
Two-stage (ours)	✓	✓	**.164**	**.141**

Table 3. **Ablation study for depth fusion.** Note that 3 means preserve the depth fusion process as shown in Fig. 2.

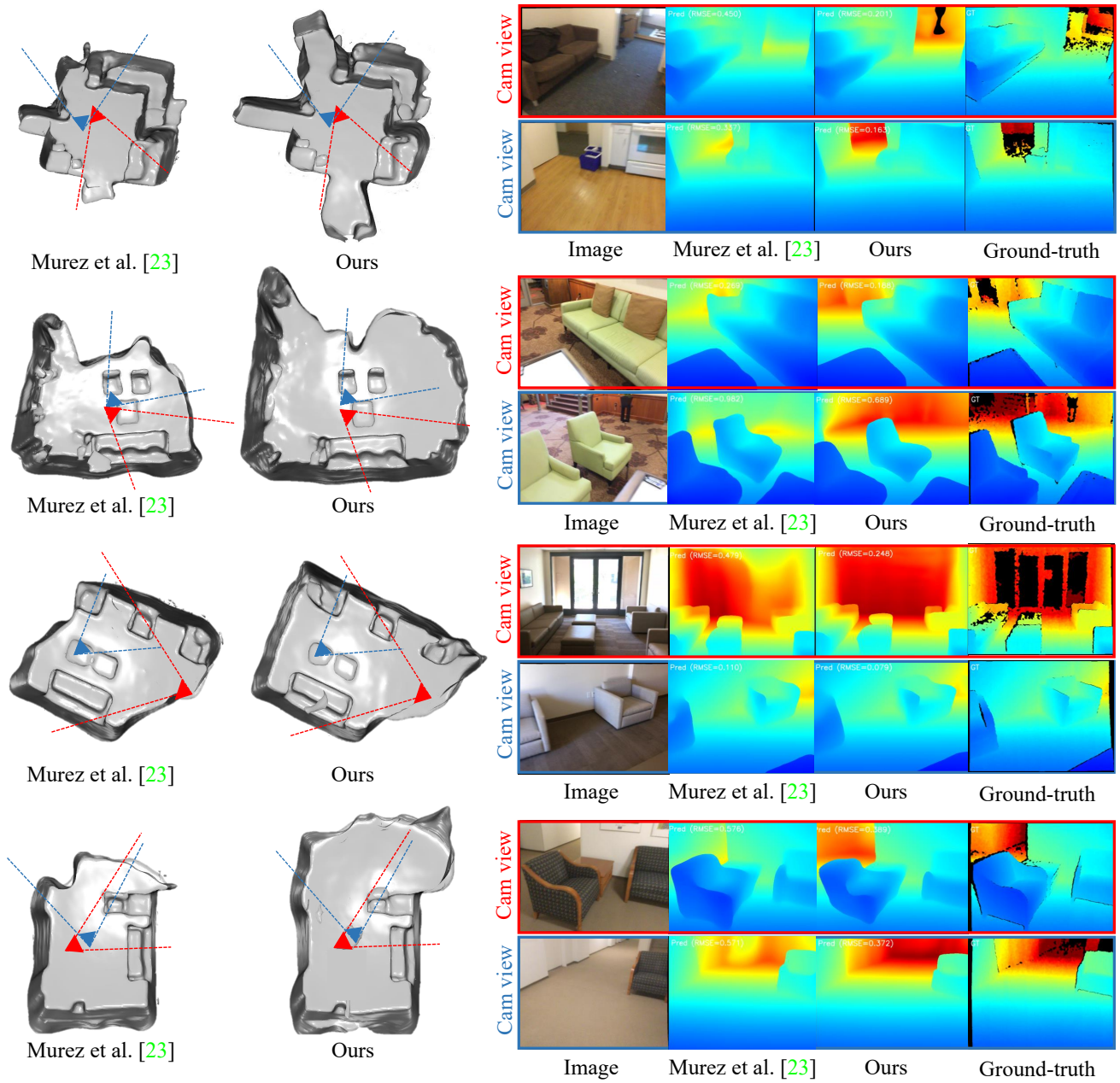
Method	Evaluation			
	AbsRel	RMSE	$\mathcal{L}_1$	F-score
Ours w/o $\mathcal{M}$	.060	.238	.162	.475
Ours w/ $\mathcal{M}$	**.049**	**.164**	**.141**	**.508**

Table 4. **Ablation study of an overlapping mask $\mathcal{M}$.**

4.3. Ablation Study

In this section, we propose to evaluate the contribution of each proposed component (PosedConv, depth map fusion, and overlapping mask), through an extensive ablation study. The obtained results are shown in Tables 2, 3, and 4.

In Table 2, we validate the PosedConv by comparing with the origin 3D Conv as in Fig. 3-(a)). This result highlights the relevance of PosedConv since it consistently improves the depth (AbsRel and RMSE) and reconstruction ($\mathcal{L}_1$ and F-score) accuracy. We attribute these results to



(a) 3D reconstruction results in mesh representation

(b) Depth maps from the reconstructed scene.

Figure 5. **Qualitative reconstruction results in ScanNet dataset [9].** (a) 3D scene reconstruction results from the recent work (Murez *et al*. [32]) and ours. (b) Depth results from the two selected camera viewpoints. Overall, our method shows better results, especially for the global structure of the target scenes.

the rotation-invariant and translation-invariant features extracted from our PosedConv, which helps to re-profile the 3D surface in the depth map fusion stage (Fig. 2).

In Table 3, we conduct an ablation study regarding the two-stage strategy of our method. Single-stage represents the direct TSDF regression as proposed in Murez *et al*. [32], and two-stage means our volume fusion network equipped with the PosedConv and depth map fusion. To verify the necessity of the depth fusion (*i.e*., depth maps embedding), we also include one additional method (w/o depth embedding) whose unified scene volume $\mathcal{V}^U$ only contains pose-invariant feature volume $\mathcal{V}^R$. The results demonstrate that our two-stage approach outperforms the single-stage strategy [32]. Moreover, when we embed masked depth maps

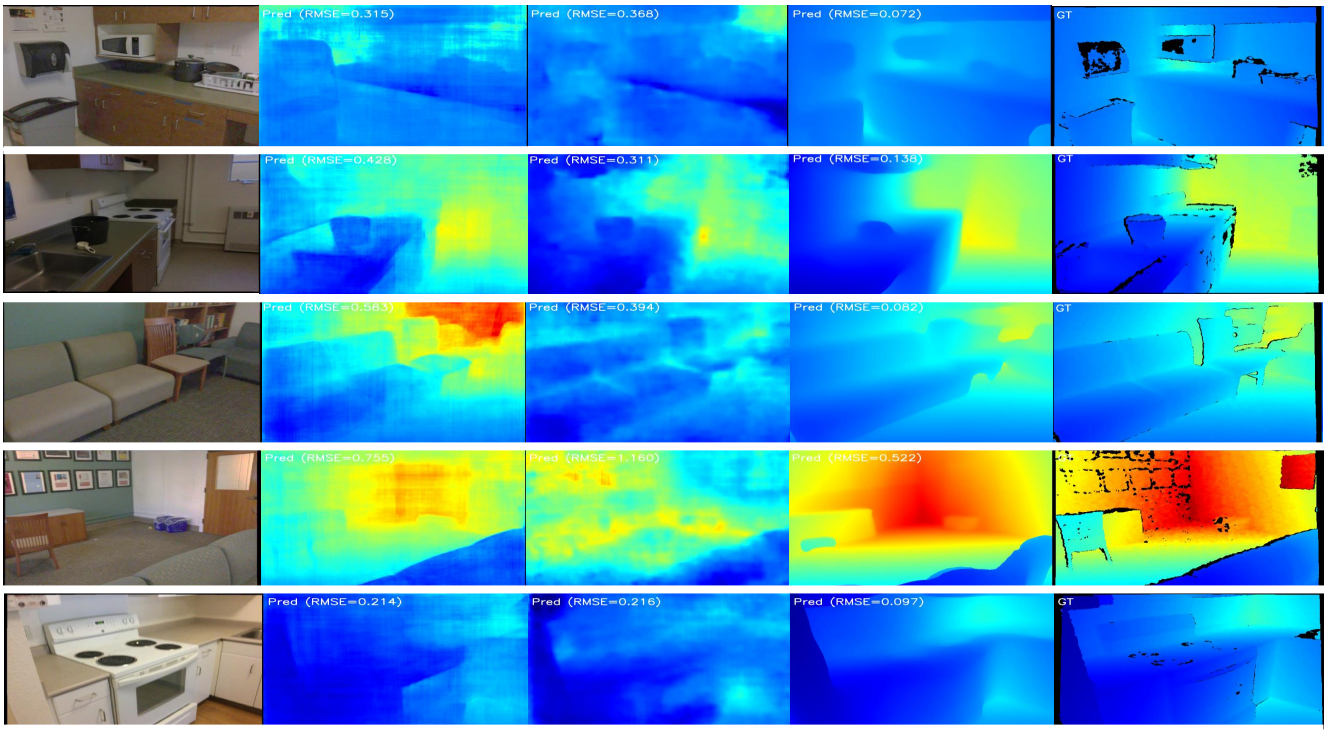


Image DPSNet [14] Before fusion ($\tilde{Z}$) After fusion (Ours) Ground-truth

Figure 6. **Qualitative results of our depth map fusion and recent depth-based MVS method [22].** Compared to our depth estimation from the local multi-view stereo (Before fusion $\tilde{Z}$) and from the final estimation (After fusion), it demonstrates that our depth fusion scheme is trained to re-profile the 3D surface so that the quality of depth maps after fusion outperforms depth maps before fusion $\tilde{Z}$.

$\hat{\mathcal{M}}$, the performance gap to the previous work [32] increases. Thanks to our volume fusion network with PosedConv and differentiable depth map fusion scheme, we obtain the best 3D reconstruction results.

Lastly, we validate the overlapping mask $\mathcal{M}$ in Table 4. This mask is used to filter out the depth values at the non-overlapping region between local multi-view images. It shows that applying overlapping masks consistently improves the quality of the depth maps and 3D reconstruction. Based on these results, we confirm that using overlapping masks under varying camera motion is effective for depth map fusion.

5. Conclusion

In this work, we presented an end-to-end volume fusion network for 3D scene reconstruction using a set of images with known camera poses. The specificity of our strategy is its two-stage structure that mimics traditional techniques: local multi-view stereo and Volumetric Depth Fusion. For the local depth maps estimation, we designed a novel multi-view stereo method that also estimates an overlapping mask. This additional output enables filtering out of the depth measurements that do not have the pixel correspondence among the neighbor views, which in turn, improves the overall reconstruction of the scene. In our depth fusion, we propagate the masked depth maps as well as the image features to re-profile the 3D surface by computing the matching cost. To improve the robustness of matching between images with different orientations – in the world-referential coordinate (*i.e.*, unified scene volume), we introduce the Posed Convolution Layer that extracts pose-invariant features. Finally, our network infers a TSDF volume that describes the global structure of the target scene. Despite the consistency and accuracy of our 3D reconstructions, volumetric representation requires huge computation power and memory that restricts the resolution of the resulting TSDF. To cope with this problem, future works for volume-free fusion via local multi-view information should be explored. In this context, our method represents a solid base for the development of future research in multi-view stereo and depth map fusion for 3D scene reconstruction.

ACKNOWLEDGMENT

This work was supported by NAVER LABS Corporation [SSIM: Semantic and scalable indoor mapping]

References

[1] Aljaž Božič, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *arXiv preprint arXiv:2107.02191*, 2021. 1

[2] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *European Conference on Computer Vision*, pages 766–779. Springer, 2008. 1, 2

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 2, 3, 4, 5

[4] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1538–1547, 2019. 1, 2

[5] Jaesung Choe, Kyungdon Joo, Tooba Imtiaz, and In So Kweon. Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation. *IEEE Robotics and Automation Letters*, 6(3):4672–4679, 2021. 4

[6] Jaesung Choe, Kyungdon Joo, Francois Rameau, and In So Kweon. Stereo object matching network. *arXiv preprint arXiv:2103.12498*, 2021. 4

[7] Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 358–363. IEEE, 1996. 2

[8] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996. 2

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 4, 5, 6, 7

[10] Arda Düzçeker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2, 4

[11] Carlos Hernández Esteban and Francis Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392, 2004. 1, 2

[12] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1, 2

[13] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 1, 2

[14] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881, 2015. 1, 2, 4

[15] David Gallup, Jan-Michael Frahm, Philippos Mordohai, Qingxiong Yang, and Marc Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1, 2

[16] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 2

[17] Hyowon Ha, Sunghoon Im, Jaesik Park, Hae-Gon Jeon, and In So Kweon. High-quality depth from uncalibrated small motion clip. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5413–5421, 2016. 2

[18] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

[19] Carlos Hernández, George Vogiatzis, and Roberto Cipolla. Probabilistic visibility for multi-view stereo. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 1, 2

[20] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2651–2660, 2019. 1, 2, 4, 6

[21] Sunghoon Im, Hyowon Ha, Gyeongmin Choe, Hae-Gon Jeon, Kyungdon Joo, and In So Kweon. Accurate 3d reconstruction from small motion clip for rolling shutter cameras. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):775–787, 2018. 2

[22] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: end-to-end deep plane sweep stereo. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 4, 6, 8

[23] Shahram Izadi, Richard A Newcombe, David Kim, Otmar Hilliges, David Molyneaux, Steve Hodges, Pushmeet Kohli, Jamie Shotton, Andrew J Davison, and Andrew Fitzgibbon. Kinectfusion: real-time dynamic 3d surface reconstruction and interaction. In *ACM SIGGRAPH 2011 Talks*, 2011. 2

[24] Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2307–2315, 2017. 2

[25] Olaf Kähler, Victor Prisacariu, Julien Valentin, and David Murray. Hierarchical voxel block hashing for efficient integration of depth images. *IEEE Robotics and Automation Letters*, 1(1):192–197, 2015. 2

[26] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, and Peter Henry. End-to-end learning of geometry and context for deep stereo regression. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 66–75. IEEE, 2017. 2, 4

[27] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[28] Maxime Lhuillier and Long Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE*

transactions on pattern analysis and machine intelligence, 27(3):418–433, 2005. 2

[29] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 1, 2

[30] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. 4

[31] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. Openmvg: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016. 1, 4

[32] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *Eur. Conf. Comput. Vis.*, 2020. 1, 4, 5, 6, 7, 8

[33] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, pages 2277–2287, 2017. 3, 5

[34] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 3, 5

[35] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):1–11, 2013. 2

[36] Despoina Paschalidou, Osman Ulusoy, Carolin Schmitt, Luc Van Gool, and Andreas Geiger. Raynet: Learning volumetric 3d reconstruction with ray potentials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2018. 2

[37] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 2, 4, 6

[38] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1

[39] Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *2018 International Conference on 3D Vision (3DV)*, pages 248–257. IEEE, 2018. 1, 2, 6

[40] Silvan Weder, Johannes Schonberger, Marc Pollefeys, and Martin R Oswald. Routedfusion: Learning real-time depth map fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 4

[41] Silvan Weder, Johannes L Schonberger, Marc Pollefeys, and Martin R Oswald. Neuralfusion: Online depth fusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3162–3172, 2021. 2

[42] Ruigang Yang and Marc Pollefeys. Multi-resolution real-time stereo on commodity graphics hardware. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003. 2

[43] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 1, 2

[44] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 4