C2F2NeUS: Cascade Cost Frustum Fusion for High Fidelity and Generalizable Neural Surface Reconstruction

Luoyuan Xu[1], Tao Guan[1], Yuesong Wang[1]*, Wenkai Liu[1], Zhaojie Zeng[1], Junle Wang[2], Wei Yang[1]

[1] School of Computer Science and Technology, Huazhong University of Science and Technology

[2] Tencent

{xu_luoyuan, qd_gt, yuesongwang, wenkai_liu, zhaojiezeng, weiyangcs}@hust.edu.cn, wangjunle@gmail.com

Abstract

There is an emerging effort to combine the two popular 3D frameworks using Multi-View Stereo (MVS) and Neural Implicit Surfaces (NIS) with a specific focus on the few-shot / sparse view setting. In this paper, we introduce a novel integration scheme that combines the multi-view stereo with neural signed distance function representations, which potentially overcomes the limitations of both methods. MVS uses per-view depth estimation and cross-view fusion to generate accurate surfaces, while NIS relies on a common coordinate volume. Based on this strategy, we propose to construct per-view cost frustum for finer geometry estimation, and then fuse cross-view frustums and estimate the implicit signed distance functions to tackle artifacts that are due to noise and holes in the produced surface reconstruction. We further apply a cascade frustum fusion strategy to effectively captures global-local information and structural consistency. Finally, we apply cascade sampling and a pseudo-geometric loss to foster stronger integration between the two architectures. Extensive experiments demonstrate that our method reconstructs robust surfaces and outperforms existing state-of-the-art methods.

1. Introduction

Reconstructing 3D structures from a set of images is a fundamental task in computer vision, with widespread applications in fields such as architectural preservation, virtual/augmented reality, and digital twins. Multi-view stereo (MVS) is a widely-used technique for addressing this task, exemplified by MVSNet [52] and its successors [38, 42, 43, 49, 51]. These methods construct 3D cost volumes based on the camera frustum, rather than regular euclidean space, to achieve precise depth map estimation. However, these methods typically require post-processing steps, such as depth map filtering, fusion, and mesh recon-
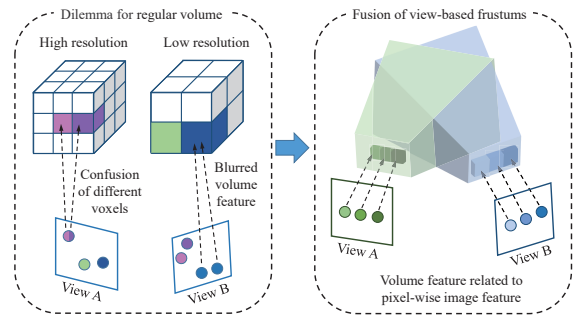
*Corresponding author. Contact him at yuesongwang@hust.edu.cn

Figure 1. A regular volume doesn't simultaneously fit all cameras well, and can easily run into a dilemma when choosing the resolution. In a low-resolution volume, a single voxel may cover multiple pixels of an image, resulting in blurred volume features, such as the blue pixel in view B. In a high-resolution volume, multiple voxels may cover only a single pixel of an image, causing confusion of different voxels, such as the pink-purple pixels in view A. Instead, frustum volume is view-dependent and extracts pixel-level image features. Hence, we build the cost frustum for each view and adopt the fusion of per-view frustum so as to fit each view.

struction, to reconstruct the 3D surface of the scene, and can not well handle noises, textureless regions, and holes.

The implicit scene representation approaches, e.g., Neural Radiance Fields (NeRF) [27] and its peer Neural Signed Distance Function [35, 36, 39], achieves remarkable results in view synthesis and scene reconstruction. The implicit surface reconstruction approaches typically employ Multilayer Perceptrons (MLPs) to implicitly fit a volume field. We then can extract scene geometry and render views from the implicit volume field. These approaches usually require a large number of images from different viewpoints and adopt a per-scene optimization strategy, which means they are not generalizable to unknown scenes.

There is an emerging effort [3, 4, 16, 59] to merge the two technical paths. MVSNeRF [4] combines NeRF [27] with MVSNet [52] for generalizable view synthesis. RC-MVSNet [3] utilizes NeRF's neural volume rendering to

handle view-dependent effects and occlusions. The most related to our approach is the SparseNeUS [23] for generalizable surface reconstruction method for sparse views. It builds a regular euclidean volume (i.e., cube) to encode geometric information by aggregating 2D feature maps of multiple images. The features sampled from it and the corresponding positions are used to estimate the signed distance function (SDF). However, a regular volume doesn't fit a camera's view naturally, which can be better modeled as view frustum. More specifically, as illustrated in Fig. 1, a volume with a higher resolution costs more memory and collect redundant image features, while a coarser volume causes quality degradation. Instead, we propose to build the cost frustum for each view and this strategy has been proven to be effective on MVSNet [52] and its successors.

In this paper, we propose a novel integration scheme that combines MVS with neural implicit surface reconstruction. To encode the global and local geometric information of the scene, we adopt the cascade architecture of CasMVS-Net [13], which is a volume pyramid. Specifically, we first construct a volume on the camera frustum and then convert it into a cascade geometric frustum. As shown in Fig. 1, to fit each camera's view well, we build a cascade frustum for every view and then fuse them using a proposed cross-view and cross-level fusion strategy that effectively captures global-local information and structural consistency. By combining the 3D position, fused feature, and view direction, we estimate the SDF and render colors using volume rendering [39]. Moreover, we utilize the intermediate information output by MVS part to apply cascade sampling and a pseudo-geometric loss, which further improves the quality of the reconstructed surface. Our experiments on the DTU [1] and BlendedMVS [53] datasets demonstrate the effectiveness and generalization ability of our proposed method, surpassing existing state-of-the-art generalization surface reconstruction techniques.

Our approach makes the following contributions:

- We introduce a novel exploration approach that integrates MVS and implicit surface reconstruction architectures for end-to-end generalizable surface reconstruction from sparse views.

- We propose a cross-view and cross-level fusion strategy to effectively fuse features from multiple views and levels.

- We further utilize information from the MVS part to apply cascade sampling and a pseudo-geometric loss to the neural surface part, promoting better integration between the two architectures.

2. Related Work

Neural Surface Reconstruction Neural implicit representations enable the representation of 3D geometries as continuous functions that are computable at arbitrary spatial locations. Due to the ability to represent complex and detailed shapes in a compact and efficient manner, these representations show significant potential in tasks such as 3D reconstruction [7,15,17,29,30,39,54,55,57], shape representation [2,12,26,31], and novel view synthesis [20,27,33,37].

To avoid relying on ground-truth 3D geometric information, many of these methods employ 2D images as supervision through classical rendering techniques, such as surface rendering and volume rendering. While some methods [17, 21, 29, 55] reconstruct the surface and render 2D images using surface rendering, they often require accurate object masks, which can be challenging to obtain in practical scenarios. As NeRF [27] successfully integrates implicit neural functions and volume rendering and generates photo-realistic novel views, some methods [7, 30, 39, 54] incorporate SDF into neural volume rendering to achieve surface reconstruction without additional masks. Despite these advancements, further improvement in surface quality is achieved by introducing additional geometric priors [11,59]. However, these methods require a large number of dense images, and it is difficult to generalize to unknown scenes, which restricts the deployment at the level of these methods.

Some methods [4, 6, 22, 40, 56] generate novel views in unknown scenarios in a generalization manner. These methods construct radiative neural fields on sparse views, and can inference on unknown scenarios without any fine-tuning after training in multiple known scenarios. Moreover, some other methods [8, 14, 28] synthesize novel views on a single scene with sparse views. However, these methods are difficult to generate high-quality geometries.

To overcome these deficiencies, SparseNeUS [23] provides a preliminary solution by encoding geometric information using a regular euclidean volume, VolRecon [32] introduces multi-view image features through the view transformer to advance this scheme, ReTR [18] uses hybrid extractor to obtain multi-level euclidean volume and then utilize reconstruction transformer to improve the performance. However, these methods are challenging to achieve high-quality reconstructions due to the regular volume doesn't fit a camera's view naturally. Moreover, VolRecon [32] and ReTR [18] additionally introduce ground truth depth labels, which are usually expensive to obtain.

Multi-view Stereo With the rapid advancements in deep learning techniques, MVS methods [5,24,25,38,42,44,45, 52] based on depth map fusion have shown remarkable performance on various benchmarks [1, 53]. The pioneering MVSNet [52] architecture constructs a 3D cost volume by leveraging differentiable homography warping operations, and generates the depth map through cost volume regularization. The key to success lies in its utilization of camera

frustums instead of regular euclidean spaces for constructing 3D cost volumes. Some attempts [13, 51] progressively optimize the depth map by refining the camera frustum in a coarse-to-fine manner. Some attempts [9, 19, 41, 60] introduce transformers to improve reconstruction performance. Some other attempts [3, 10, 46–50] train networks in an unsupervised manner. However, these methods require a series of post-processing operations, such as depth map filtering, depth map fusion, and mesh reconstruction, to reconstruct the 3D structure of the scene, and can not well handle noises, textureless regions, and holes.

The Integration of MVS and Neural Implicit Scene Representation The integration of MVS and neural implicit scene representation generates significant interest among researchers, leading to several recent explorations [3, 4, 16, 59]. MVSNeRF [4] constructs a cost volume to enable geometry-aware scene reasoning. It then uses volume rendering [27] in combination with position, view direction, and volume features to perform neural radiation field reconstruction and achieve generalizable view synthesis. RC-MVSNet [3] adds an independent cost volume for volume rendering, allowing the network to learn how to handle view-dependent effects and occlusions and improve the quality of depth maps. MVSDF [59] leverages the geometry and feature consistency of Vis-MVSNet [58] to optimize the SDF, resulting in more robust geometry estimation.

However, MVSNeRF [4] is difficult to generate high-quality surfaces, RC-MVSNet [3] requires cumbersome post-processing steps to obtain surfaces, and MVSDF [59] cannot generalize to unknown scenes and requires dense images. Our method, on the other hand, differs significantly as it focuses on achieving high fidelity and generalizable surface reconstruction for sparse views.

3. Method

In this section, we explain the detailed structure of our proposed C2F2NeUS, which is a novel integration scheme that better combines MVS with neural implicit surface representation. With this integration, C2F2NeUS achieves high fidelity and generalizable surface reconstruction for sparse views in an end-to-end manner. As illustrated in Fig. 2, by fusing the view-dependent frustums in the MVS part, we obtain more accurate geometric features which are sent to the neural implicit surface part to predict SDF and extract surfaces.

Specifically, we first construct a view-dependent cascade geometric frustum for each view to encode geometric information of the scene and fully exploit the advantage of MVS(Sec. 3.1). For a given set of 3D coordinates, we then sample and fuse the feature from these frustums by using the proposed cross-view and cross-level fusion strategy (Sec. 3.2). This strategy can effectively capture global

local information and structural consistency. Next, we introduce how to predict SDF and render color from the fused feature (Sec. 3.3). And the SDF prediction network generates an SDF field which is used for surface reconstruction, this representation leverages the smooth and complete geometry of SDF. To train the SDF prediction network in an unsupervised manner, we render color via volume rendering. Finally, we introduce the training loss of our end-to-end framework (Sec. 3.4).

3.1. Cascade Geometric Frustum Generation

To make the implicit neural surface reconstruction generalizable and capture the scene information more accurately, we follow the volume pyramid of CasMVSNet [13] and encode the global and local geometric information of the scene by building a cascade geometric frustum. Unlike SparseNeUS [23], which utilizes a regular euclidean volume, we construct the volume from the perspective MVS. In MVS, the reference view is most important, and other source views contribute to the depth estimation for the reference view. Since a single view-dependent frustum cannot describe the complete scene, we create a frustum for each image and treat the image as the reference view and other images as source views. Besides, we estimate the corresponding depth maps from each cost frustum to construct the cascade geometric frustums.

To accomplish this, we first extract feature maps $\{F_i\}_{i=0}^{N-1}$ using a 2D feature extraction network for N images $\{I_i\}_{i=0}^{N-1}$ of the scene. With the corresponding camera parameters $\{K_i, R_i, T_i\}_{i=0}^{N-1}$ of each image, we then build a 3D cost frustum $C \in \mathbb{R}^{c \times d \times h \times w}$ for the reference camera via differential homography warping operations, where c, d, h, w are dimensions of feature, number of depth samples, height, width respectively. In our implementation, we construct a 3D cost frustum $\{C_i\}_{i=0}^{N-1}$ for each image as the reference view and use the remaining images as source views. The 3D cost frustums C_i are then regularized by 3D CNN Ψ_1 to obtain the intermediate volumes $V_i \in \mathbb{R}^{c \times d \times h \times w}$. On the one hand, the intermediate volumes V_i are used to estimate the probability volumes $P_i \in \mathbb{R}^{1 \times d \times h \times w}$ which are used to regress the depth maps $D_i \in \mathbb{R}^{1 \times 1 \times h \times w}$ of the current reference views I_i. On the other hand, the intermediate volumes V_i are further regularized by a new 3D CNN Ψ_2 to obtain the geometric frustums $G_i \in \mathbb{R}^{c \times d \times h \times w}$.

$$V_i, P_i, D_i = \Psi_1(C_i), \qquad G_i = \Psi_2(V_i). \qquad (1)$$

The depth maps D_i are used to redefine the depth hypothesis and construct the cascade 3D cost frustums, ultimately constructing cascade geometric frustums $\{G_i^j\}_{i=0,...,N-1}^{j=0,...,L-1}$, where L is the cascade level number.
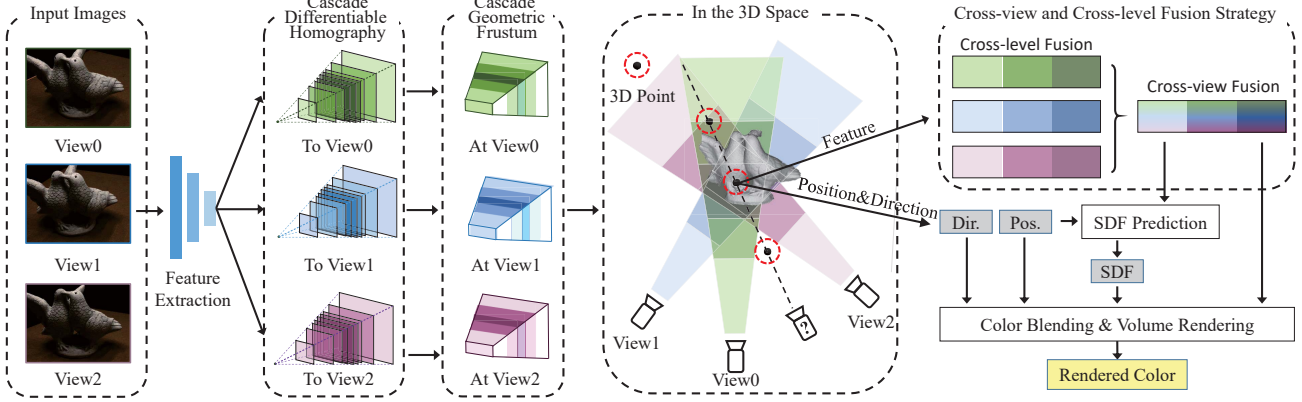
Figure 2. The overview of C2F2NeUS. We first construct the cascade geometric frustum of each view to capture global and local geometric information about the scene. Then we apply a cross-view and cross-level fusion strategy to effectively fuse features from multiple views and levels. Finally, the positions of 3D points and their corresponding fused features are fed to an SDF prediction network which is trained by rendered colors from volume rendering.

3.2. Cross-view and Cross-level Fusion Strategy

The cost frustum of each view and level has different importance. Intuitively, regions with relatively smaller angles w.r.t. the viewpoint in finer frustums are more crucial. Therefore, we estimate the weight for each frustum to represent importance, similar to Vis-MVSNet [58]. Another problem in the fusion process is that points near the surface can sample features from all three pyramid levels, but points far away from the surface can only extract features from the coarse level. Straightforward fusion strategies, such as adding features from all views and levels, will confuse features of different levels, while directly concatenating all features together make it difficult to deal with an arbitrary number of input views. Consequentially, we propose a cross-view and cross-level fusion strategy that treats each view and level differently. This fusion strategy effectively captures the spatial and structural information of the scene and produces a more precise surface.

We introduce an adaptive weight $A_i^j \in \mathbb{R}^{1 \times d \times h \times w}$ for each geometric frustum G_i^j, which is normalized using the sigmoid function. Therefore, we can rewrite Equ. 1 as

$$V_i, P_i, D_i = \Psi_1(C_i), \qquad G_i, A_i = \Psi_2(V_i). \quad (2)$$

To integrate both the global information from coarser frustums and the local information of finer frustums, we concatenate features at different levels and sum the features at different viewpoints according to their weights. Specifically, we sample the corresponding features $g_i^j = G_i^j(p) \in \mathbb{R}^{1 \times c}$ and weights $a_i^j = A_i^j(p) \in \mathbb{R}^{1 \times 1}$ of a given 3D position $p \in \mathbb{R}^{1 \times 3}$ from all frustums using bilinear interpolation. Then, we concatenate features and sum the weights of different levels $j = 0, ...L - 1$ for each viewpoint I_i, and obtain new features $g_i^L = \text{cat}(\{g_i^j\})$, and new

weights $a_i^L(p) = \text{sum}(\{a_i^j\})$, respectively, where $g_i^L(p) \in \mathbb{R}^{1 \times Lc}$ and $a_i^L(p) \in \mathbb{R}^{1 \times 1}$. Finally, we fuse the concatenated features g_i^L from different viewpoints $\{I_i\}_{i=0}^{N-1}$ based on their respective weights a_i^L. The final geometric feature of the given 3D position p is defined as $f_{geo} = \Sigma_{i=0}^{N-1} a_i^L \cdot g_i^L / \Sigma_{i=0}^{N-1} a_i^L$, where $f_{geo} \in \mathbb{R}^{1 \times Lc}$.

3.3. SDF Prediction and Volume Rendering

We would like to exploit the advantage of neural implicit surface reconstruction, i.e., the surface extracted from a neural SDF network is usually very smooth and consistent.

SDF Prediction. Given an SDF prediction network Φ consisting of MLP and an arbitrary 3D position p with its corresponding geometric feature f_{geo}, we first encode the position p using position encoding $\gamma(\cdot)$. We then use the encoded position and geometric feature f_{geo} as input to the SDF prediction network Φ to predict the SDF $s(p)$ of 3D position p. Our SDF prediction operation is defined as:

$$s(p) = \Phi(\gamma(p), f_{geo}). \quad (3)$$

Blending Weights. Similar to IBRNet [40], we use blending weights to estimate color of a 3D position p and view direction r. We extract 2D color feature maps from N input images via a new feature extract network. For a given 3D position p with its corresponding geometric feature f_{geo} and view direction r, we project p onto N input views and extract corresponding color features f_i^{col} from color feature maps using bilinear interpolation. We then compute the mean u and variance v of the sampled color features f_i^{col} for different views to capture cross-image information and concatenate each feature f_i^{col} with u and v. A small shared

4

MLP Γ is used to process the concatenated features and generate new features f_i^{col2} that contain color information. We also compute the direction difference $\Delta r = r - r_i$ between the view direction r and each input image's viewpoint r_i. The color features f_i^{col2}, direction differences Δr, and geometric features f_{geo} are fed into a new MLP network Γ_{col} for generating blending weights $w_i(p)$.

$$w_i(p) = \Gamma_{col}(\Gamma(f_i^{col}, u, v), \Delta r, f_{geo}). \quad (4)$$

Finally, we use the softmax operator to normalize blending weights $\{w_i(p)\}_{i=0}^{N-1}$.

Volume Rendering. As there is no ground-truth 3D geometry, to supervise the SDF prediction network, we render the color of the query ray and calculate its consistency with the ground-truth color. Specifically, we perform the ray point sampling, where each sampled position p and viewpoint d are used to predict the corresponding SDF $s(p)$ and blending weights $w_i(p)$. We then project the position p onto N input images to extract their respective colors $c_i(p)$ and compute the color $c(p)$ of position p as a weighted sum of the sampled color $c_i(p)$ and blending weights $w_i(p)$. Next, we apply volume rendering as in NeUS [39] to render the color of the ray by aggregating the SDF and color of each position p along the ray. The rendered color is compared to the ground-truth color for calculating the consistency loss.

Cascade Sampling and Pseudo-depth Generation. To further leverage the benefits of MVS and enhance the quality of the extracted surfaces, we incorporate cascade sampling on the frustum and a pseudo-geometric loss, which enforce a stronger integration between MVS and neural implicit surface. In this work, we apply an adaptive sampling strategy using the intermediate probability volume P_i of the cascade frustum generation network. Specifically, we take the query image as the reference view, and other input images as the source views, and send them to the cascade frustum generation network to obtain the depth maps D_{que}^j and probability volumes P_{que}^j at different levels. We use the probability volume $P_{que}^{j=0}$ of the coarsest layer for cascade sampling. We then compute the mean α and standard β deviation of the probability volume $P_{que}^{j=0}$ along the depth channel. The adaptive sample ranges $[t_n, t_f]$ are defined as follows:

$$[t_n, t_f] = [\alpha - \beta, \alpha + \beta]. \quad (5)$$

With the high-resolution depth maps $D_{que}^{j=L-1}, D_i^{j=L-1}$ of the query image and other input images, we compute the geometric consistency to obtain the respective effective masks. The masked depth maps can be considered pseudo-depth labels, which use to supervise the SDF prediction network. Specifically, the masked depth map of the query image is used to compute a pseudo-depth consistency loss.

Further, we fuse the masked depth maps of different images into point clouds, which are used to directly supervise the SDF prediction network.

3.4. Loss Function

Ground-truth 3D geometric labels are difficult to obtain, to address this issue, our framework employs an unsupervised learning approach. Specifically, we introduce a training loss $\mathcal{L}_{total}$ as a combination of two unsupervised losses for training the depth map and SDF, respectively.

$$\mathcal{L}_{total} = \mathcal{L}_{dep} + \mathcal{L}_{sdf}. \quad (6)$$

For the loss $\mathcal{L}_{dep}$, we supervise the intermediate depth map D_i^j of cascade geometric frustum network by utilizing several losses of SMU-MVSNet [49].

For the loss $\mathcal{L}_{sdf}$, we adopt the same loss item of SparseNeUS [23], i.e. the color consistency loss $\mathcal{L}_{cc}$, the Eikonal term $\mathcal{L}_{eik}$, the sparseness regularization term $\mathcal{L}_{spa}$. Moreover, we introduce a geometry-based loss derived from pseudo-depth, which can provide reliable guidance without relying on expensive ground-truth geometry labels. The overall loss $\mathcal{L}_{sdf}$ is defined as follows:

$$\mathcal{L}_{sdf} = \mathcal{L}_{cc} + \mathcal{L}_{eik} + \mathcal{L}_{spa} + \mathcal{L}_{pdc} + \mathcal{L}_{pgs}. \quad (7)$$

The pseudo-depth consistency loss $\mathcal{L}_{pdc}$ is an L1 distance between the rendered depth and the pseudo-depth label. $\mathcal{L}_{pdc}$ is defined as:

$$\mathcal{L}_{pdc} = \frac{1}{X} \sum_{x=0}^{X-1} |d - \hat{d}|_1, \quad (8)$$

where d and $\hat{d}$ are the rendered depth and ground-truth depth respectively. $\mathcal{L}_{pgs}$ is the pseudo-geometry SDF loss [11]. The SDF values of the pseudo point cloud are zeroes. $\mathcal{L}_{pgs}$ is defined as:

$$\mathcal{L}_{pgs} = \frac{1}{||Q_2||} \sum_{q_2 \in Q_2} |s(q_2)|, \quad (9)$$

where Q_2 is a set of 3D points randomly selected from the pseudo point clouds.

4. Experiments

In this section, we demonstrate the effectiveness of our proposed method. Firstly, we provide a detailed account of our experimental settings, which includes implementation details, datasets, and baselines. Secondly, we present quantitative and qualitative comparisons on two widely used datasets, namely DTU [1] and BlendedMVS [53]. Finally, we conduct detailed ablation studies to analyze the contribution of different components of our proposed method.
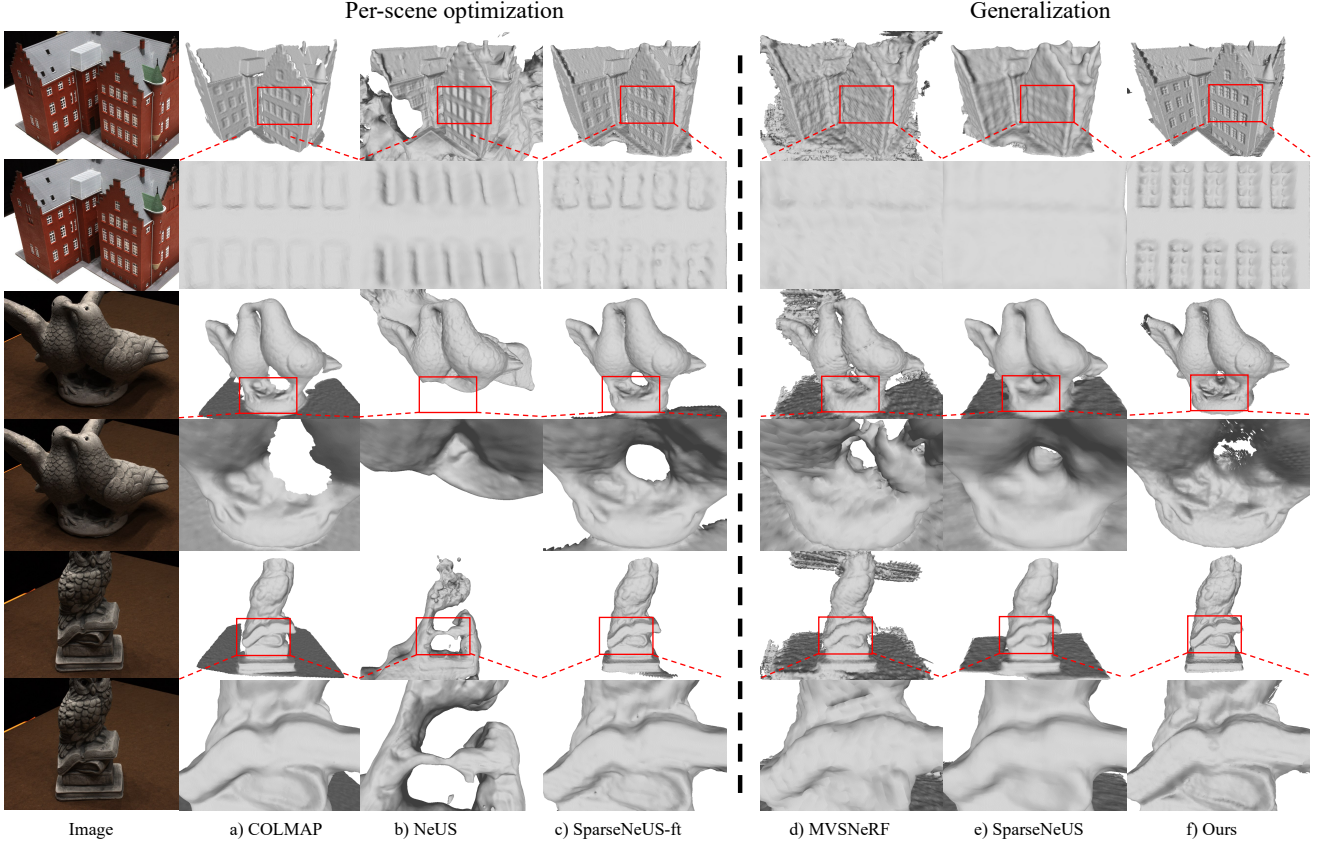
Per-scene optimization | Generalization

Image | a) COLMAP | b) NeUS | c) SparseNeUS-ft | d) MVSNeRF | e) SparseNeUS | f) Ours

Figure 3. The qualitative comparison of our method with other state-of-the-art methods on DTU.

Scan	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	**Mean**
COLMAP [34]	**0.90**	2.89	1.63	1.08	2.18	1.94	1.61	1.30	2.34	1.28	1.10	1.42	0.76	1.17	1.14	1.52
IDR [55]	4.01	6.40	3.52	1.91	3.96	2.36	4.85	1.62	6.37	5.97	1.23	4.73	0.91	1.72	1.26	3.39
VolSDF [54]	4.03	4.21	6.12	0.91	8.24	1.73	2.74	1.82	5.14	3.09	2.08	4.81	0.60	3.51	2.18	3.41
UNISURF [30]	5.08	7.18	3.96	5.30	4.61	2.24	3.94	3.14	5.63	3.40	5.09	6.38	2.98	4.05	2.81	4.39
NeUS [39]	4.57	4.49	3.97	4.32	4.63	1.95	4.68	3.83	4.15	2.50	1.52	6.47	1.26	5.57	6.11	4.00
IBRNet-ft [40]	1.67	2.97	2.26	1.56	2.52	2.30	1.50	2.05	2.02	1.73	1.66	1.63	1.17	1.84	1.61	1.90
SparseNeUS-ft [23]	1.29	**2.27**	1.57	0.88	1.61	1.86	1.06	1.27	1.42	1.07	0.99	0.87	0.54	1.15	1.18	1.27
PixelNerf [56]	5.13	8.07	5.85	4.40	7.11	4.64	5.68	6.76	9.05	6.11	3.95	5.92	6.26	6.89	6.93	6.28
IBRNet [40]	2.29	3.70	2.66	1.83	3.02	2.83	1.77	2.28	2.73	1.96	1.87	2.13	1.58	2.05	2.09	2.32
MVSNeRF [4]	1.96	3.27	2.54	1.93	2.57	2.71	1.82	1.72	2.29	1.75	1.72	1.47	1.29	2.09	2.26	2.09
SparseNeUS [23]	1.68	3.06	2.25	1.10	2.37	2.18	1.28	1.47	1.80	1.23	1.19	1.17	0.75	1.56	1.55	1.64
VolRecon† [32]	1.20	2.59	1.56	1.08	1.43	1.92	1.11	1.48	1.42	1.05	1.19	1.38	0.74	1.23	1.27	1.38
ReTR† [18]	1.05	2.31	1.44	0.98	**1.18**	**1.52**	0.88	1.35	1.30	0.87	1.07	0.77	0.59	1.05	1.12	1.17
Ours	1.12	2.42	**1.40**	**0.75**	1.41	1.77	**0.85**	**1.16**	**1.26**	**0.76**	**0.91**	**0.60**	**0.46**	**0.88**	**0.92**	**1.11**

Table 1. The quantitative results of different methods on DTU. † indicates supervised by ground truth depth labels

4.1. Experimental Settings

Implementation Details. We implement our method in PyTorch. In the cascade geometric frustum generation network, we adopt the same cascade scheme with CasMVS-Net [13], but make the following changes: we use the 2D feature extraction network consisting of 9 convolutional layers, share the 3D CNN Ψ across levels, and set the dimension of image and volume features to $c = 8$. During training, we take $N = 5$ images with a resolution of 640×512 as input and use an additional image as the query image to supervise the SDF prediction network Φ. The cascade stage number is set to $L = 3$. We train our end-to-end framework on one A100 GPU with a batch size of 2 for 300k iterations. We set the same learning rate and cosine decay schedule as NeUS [39]. The ray number is set to $X = 512$, and the sample number on each ray is $Y = N_{coarse} + N_{fine}$, where $N_{coarse} = 64$ and $N_{fine} = 64$. The weights of $\mathcal{L}_{pdc}$ and $\mathcal{L}_{pgs}$ in Equ. 7 are $0, 0$ before 10k iterations and $0.05, 1$ after that.
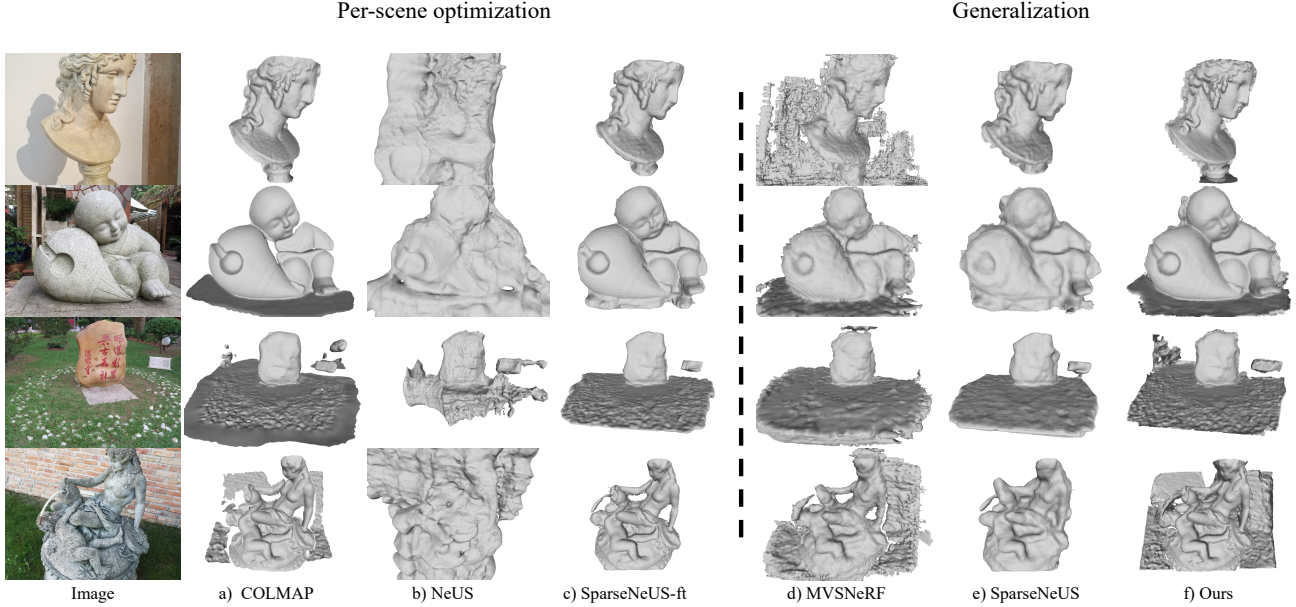
Figure 4. The qualitative comparison of our method with other state-of-the-art methods on BlendedMVS. We reconstruct the surface with our method without any fine-tuning.

Datasets. The DTU dataset [1] is a well-known indoor multi-view stereo dataset, consisting of 124 scenes captured under 7 distinct lighting conditions. Consistent with prior research [11, 23, 39, 59], we employ 75 scenes for training and 15 non-overlapping scenes for testing. Each test scene contains two sets of three images offered by SparseNeUS [23]. We evaluate our method using three views with a resolution of 1600×1152. To ensure fairness in evaluation, we adopt the foreground masks provided by IDR [55] to assess the performance of our approach on the test set, as in previous studies [11, 23, 39, 59]. To examine the generalization ability of our proposed framework, we conduct a qualitative comparison of our method on the BlendedMVS dataset [53] without any fine-tuning.

4.2. Comparisons on DTU

We perform surface reconstruction for sparse views (only 3 views) on the DTU dataset [1] and evaluate the predicted surface against the ground-truth point clouds using the chamfer distance metric. Tab. 1 and Fig. 3 present a summary of the comparison between our method and other existing methods, which demonstrate that our method achieves better performance. It is important to note that our method is solely trained on the training set without any fine-tuning on the test set to assess its generalization capability. Our method surpasses the generalizable version of SparseNeUS [23] by 32% and significantly outperforms its fine-tuning variant. Furthermore, our method exhibits superior performance compared to VolRecon [32] and ReTR [18] , which are the state-of-the-art generalizable neural implicit reconstruction methods and are supervised
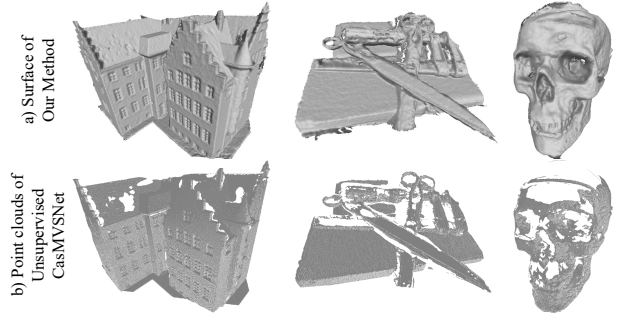

Figure 5. The qualitative comparison of our method with MVS. We present the surface reconstructed by our method, and point clouds of unsupervised CasMVSNet.

with ground-truth deep labels

4.3. Generalization on BlendedMVS

To showcase the generalization capabilities of our proposed method, we conduct additional tests on the Blended-MVS dataset [53] without any fine-tuning. The qualitative comparison between our method and other methods is presented in Fig. 4. The results indicate that our method exhibits a robust generalization ability and produces a more refined surface when compared to other generalizable neural implicit reconstruction methods.

4.4. Comparison with Unsupervised MVS

To compare our method with MVS, we retrain CasMVS-Net [13] with the unsupervised loss in Equ. 6. We estimate the depth maps of three views, filter the depth maps using geometric consistency and the masks provided by IDR [55],

Method	Inp. Size	Vol. Size	Cham. Dis.	GPU
Euclidean	800*600	96*96*96	1.77	**2231** M
Euclidean	800*600	192*192*192	1.62	7073 M
Frustum	200*144	200*144*48	1.50	2291 M
Frustum	400*288	400*288*48	**1.42**	4621 M

Table 2. Effect of camera frustum and regular euclidean space. We evaluate the two methods without cascade under different image sizes and volume sizes. Our method achieves a better performance.

Stage1	Stage2	Stage3	$\mathcal{L}_{pdc}$	$\mathcal{L}_{pgs}$	Cham. Dis
√					1.42
	√				1.28
	√		√		1.24
	√		√	√	1.18
		√			1.19
		√	√	√	**1.11**

Table 3. Effect of different components. The performance continues to improve as components increase, which demonstrates the effectiveness of each component.

and fuse them into a point cloud. The qualitative comparison is shown in Fig. 5, which demonstrates that our method is more robust and reconstructs a more complete surface.

4.5. Ablation Studies

Effect of Camera Frustum and Regular Euclidean Space. Regular volume doesn't simultaneously fit all cameras well, which leads to blurred features, particularly in sparser scenes. Instead, camera frustum volume can better model. We present a performance comparison between the regular euclidean volume and the camera frustum volume without cascade in Table 2. The comparison results reveal that our method achieves better performance with similar GPU memory.

Effect of Different Components. In this experiment, we present the results of different components to demonstrate their effectiveness. We fellow CasMVSNet [13] and adopt three level pyramid structure. **Stage1** indicates that only the first level of the pyramid is used for the SDF network. **Stage2** means we sample features from the first and second-level volumes and fuse them together. Similarly, **Stage3** fuses the features from the first, second, and third-level volumes. As shown in Tab. 3, the reconstructed surface quality significantly improves as we increase the cascade stage numbers. Moreover, The introduction of $\mathcal{L}_{pdc}$ and $\mathcal{L}_{pgs}$ further improves the surface quality.

Effect of Cross-view and Cross-level Fusion Strategy. To demonstrate the effectiveness of the proposed cross-view and cross-level fusion strategy, we remove this strategy on *Stage3* and adopt simple addition. As shown in Tab. 4

Method	Cham. Dis
Stage3 w/o fusion strategy	2.71
Stage3 w/ fusion strategy	**1.19**

Table 4. Effect of cross-view and cross-level fusion strategy. The performance severely degrades without the fusion strategy.
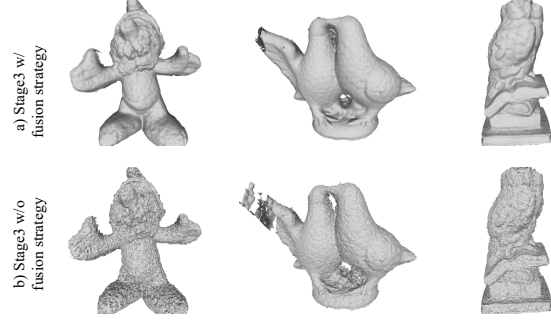

Figure 6. The qualitative comparison between with and without fusion strategy. Without our fusion strategy, the reconstructed surface becomes noisy.

and Fig. 6, using only simple addition will lead to severe performance degradation and extract noisy surfaces. On the other hand, with our fusion strategy, the performance is significantly improved, and a finer surface is extracted. This demonstrates the effectiveness of our proposed fusion strategy in capturing global-local information and structural consistency.

5. CONCLUSIONS

We propose a novel integration scheme, C2F2NeUS, for exploiting both the strengths of MVS and neural implicit surface reconstruction. Previous methods rely on regular euclidean volume for cross-view fusion, which doesn't simultaneously fit all cameras well and may lead to blurred features. We instead present a cascade geometric frustum for each view and conduct an effective fusion of all the views. Our method achieves state-of-the-art reconstruction quality for sparse inputs, which demonstrates its effectiveness. However, our method still suffers from several limitations, one is that the frustums can overlap with each other in 3D space resulting in redundant computations, and our approach constructs a cost frustum for each view, making it infeasible for dense views. In the future, we plan to optimize the frustum space and reduce computation in overlapping areas.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 2, 5, 7

[2] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2565–2574, 2020. 2

[3] Di Chang, Aljaz Bozic, Tong Zhang, Qingsong Yan, Yingcong Chen, Sabine Süsstrunk, and Matthias Nießner. Rc-mvsnet: Unsupervised multi-view stereo with neural rendering. In *European Conference on Computer Vision*, 2022. 1, 3

[4] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14104–14113, 2021. 1, 2, 3, 6

[5] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020. 2

[6] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7911–7920, 2021. 2

[7] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6260–6269, 2022. 2

[8] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 2

[9] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8594, 2022. 3

[10] Yikang Ding, Qingtian Zhu, Xiangyue Liu, Wentao Yuan, Haotian Zhang, and Chi Zhang. Kd-mvs: Knowledge distillation based self-supervised learning for multi-view stereo. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 630–646. Springer, 2022. 3

[11] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-neus: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022. 2, 5, 7

[12] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems 2020*, pages 3569–3579. 2020. 2

[13] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2020. 2, 3, 6, 7, 8

[14] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021. 2

[15] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1261, 2020. 2

[16] Mohammad Mahdi Johari, Yann Lepoittevin, and Franccois Fleuret. Geonerf: Generalizing nerf with geometry priors. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18344–18347, 2021. 1, 3

[17] Petr Kellnhofer, Lars C Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4287–4297, 2021. 2

[18] Yixun Liang, Hao He, and Ying-cong Chen. Rethinking rendering in generalizable neural surface reconstruction: A learning-based solution. *arXiv preprint arXiv:2305.18832*, 2023. 2, 6, 7

[19] Jinli Liao, Yikang Ding, Yoli Shavit, Dihe Huang, Shihao Ren, Jia Guo, Wensen Feng, and Kai Zhang. Wt-mvsnet: window-based transformers for multi-view stereo. *Advances in Neural Information Processing Systems*, 35:8564–8576, 2022. 3

[20] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 2

[21] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020. 2

[22] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7824–7833, 2022. 2

[23] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 210–227. Springer, 2022. 2, 3, 5, 6, 7

[24] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10452–10461, 2019. 2

[25] Keyang Luo, Tao Guan, Lili Ju, Yuesong Wang, Zhuo Chen, and Yawei Luo. Attention-aware multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1590–1599, 2020. 2

[26] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 2

[27] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020. 1, 2, 3

[28] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 2

[29] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 2

[30] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 2, 6

[31] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[32] Yufan Ren, Tong Zhang, Marc Pollefeys, Sabine Süsstrunk, and Fangjinhua Wang. Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16685–16695, 2023. 2, 6, 7

[33] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 2

[34] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 6

[35] Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems*, 33:10136–10147, 2020. 1

[36] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. 2021. 1

[37] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 2

[38] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14203, 2021. 1, 2

[39] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Neural Information Processing Systems*, 2021. 1, 2, 5, 6, 7

[40] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2, 4, 6

[41] Xiaofeng Wang, Zheng Zhu, Guan Huang, Fangbo Qin, Yun Ye, Yijia He, Xu Chi, and Xingang Wang. Mvster: Epipolar transformer for efficient multi-view stereo. In *European Conference on Computer Vision*, pages 573–591. Springer, 2022. 3

[42] Yuesong Wang, Tao Guan, Zhuo Chen, Yawei Luo, Keyang Luo, and Lili Ju. Mesh-guided multi-view stereo with pyramid architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2039–2048, 2020. 1, 2

[43] Yuesong Wang, Keyang Luo, Zhuo Chen, Lili Ju, and Tao Guan. Deepfusion: A simple way to improve traditional multi-view stereo methods using deep learning. *Knowledge-Based Systems*, 221:106968, 2021. 1

[44] Yuesong Wang, Zhaojie Zeng, Tao Guan, Wei Yang, Zhuo Chen, Wenkai Liu, Luoyuan Xu, and Yawei Luo. Adaptive patch deformation for textureless-resilient multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1621–1630, 2023. 2

[45] Zizhuang Wei, Qingtian Zhu, Chen Min, Yisong Chen, and Guoping Wang. Aa-rmvsnet: Adaptive aggregation recurrent multi-view stereo network. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6167–6176, 2021. 2

[46] Hongbin Xu, Zhipeng Zhou, Yu Qiao, Wenxiong Kang, and Qiuxia Wu. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 2, page 6, 2021. 3

[47] Hongbin Xu, Zhipeng Zhou, Yali Wang, Wenxiong Kang, Baigui Sun, Hao Li, and Yu Qiao. Digging into uncertainty

in self-supervised multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6078–6087, 2021. 3

[48] Luoyuan Xu, Tao Guan, Yuesong Wang, Yawei Luo, Zhuo Chen, Wenkai Liu, and Wei Yang. Self-supervised multi-view stereo via adjacent geometry guided volume completion. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2202–2210, 2022. 3

[49] Luoyuan Xu, Yawei Luo, Keyang Luo, Yuesong Wang, Tao Guan, Zhuo Chen, and Wenkai Liu. Exploiting the structure information of suppositional mesh for unsupervised multi-view stereo. *IEEE MultiMedia*, 29(1):94–103, 2021. 1, 3, 5

[50] Jiayu Yang, Jose M Alvarez, and Miaomiao Liu. Self-supervised learning of depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7526–7534, 2021. 3

[51] Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4877–4886, 2020. 1, 3

[52] 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

[53] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020. 2, 5, 7

[54] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Neural Information Processing Systems*, 2021. 2, 6

[55] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 2, 6, 7

[56] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2, 6

[57] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 2

[58] Jingyang Zhang, Yao Yao, Shiwei Li, Zixin Luo, and Tian Fang. Visibility-aware multi-view stereo network. *British Machine Vision Conference (BMVC)*, 2020. 3, 4

[59] Jingyang Zhang, Yao Yao, and Long Quan. Learning signed distance field for multi-view surface reconstruction. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6505–6514, 2021. 1, 2, 3, 7

[60] Jie Zhu, Bo Peng, Wanqing Li, Haifeng Shen, Zhe Zhang, and Jianjun Lei. Multi-view stereo with transformer. *arXiv preprint arXiv:2112.00336*, 2021. 3