UFORecon: Generalizable Sparse-View Surface Reconstruction from Arbitrary and Unfavorable Sets

Youngju Na   Woo Jae Kim   Kyu Beom Han   Suhyeon Ha   Sung-Eui Yoon

KAIST

{yjna2907,wkim97,qbhan,suhyeon.ha,sungeui}@kaist.ac.kr

Abstract

*Generalizable neural implicit surface reconstruction aims to obtain an accurate underlying geometry given a limited number of multi-view images from unseen scenes. However, existing methods select only informative and relevant views using predefined scores for training and testing phases. This constraint makes the model impractical because we cannot always ensure the availability of favorable combinations in real-world scenarios. We observe that previous methods output degenerate solutions under arbitrary and unfavorable sets. Building upon this finding, we propose **UFORecon**, a robust view-combination generalizable surface reconstruction framework. To this end, we apply cross-view matching transformers to model interactions between source images and build correlation frustums to capture global correlations. In addition, we explicitly encode pairwise feature similarities as view-consistent priors. Our proposed framework largely outperforms previous methods not only in view-combination generalizability but also in the existing generalizable protocol trained with favorable view-combinations. The code is available at https://github.com/Youngju-Na/UFORecon.*

1. Introduction

Reconstructing 3D geometries from images of multiple camera viewpoints is a fundamental problem in the computer vision field, also applicable to robotics [6, 28, 49], autonomous driving [13, 22], and AR/VR applications [7, 26]. Multi-view stereo (MVS) is a widely used technique for reconstructing 3D geometry from multi-view images. Conventional MVS methods [2, 10, 19, 34] find correspondences across input images taken from different viewpoints. With the recent advances in deep learning, learning-based MVS methods [3, 8, 11, 44] leverage differentiable homography warping with 3D Convolutional Neural Network (CNN) to effectively search correspondences in feature space. Recent progress in 3D reconstruction [15, 30, 45]
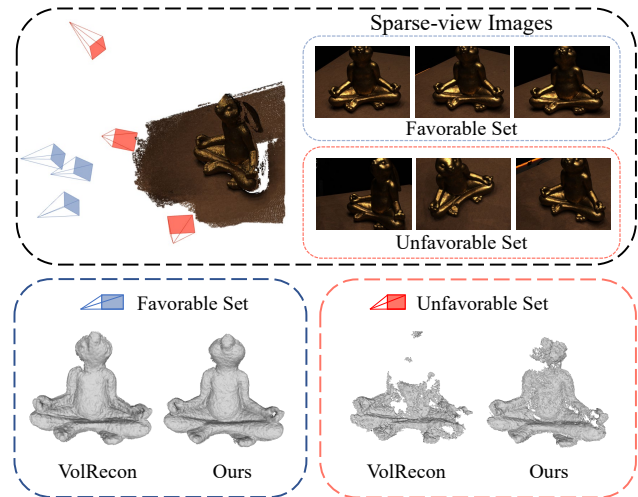

Figure 1. **Reconstruction results from different view combinations**. Both are trained only with the best-selected training protocol and tested with favorable (Blue) and unfavorable sets (Red), respectively. VolRecon [32] leads to a degenerate geometry in the unfavorable set while achieving accurate geometry in the favorable set. Our approach produces reasonable geometry on both sets.

increasingly emphasizes the use of implicit neural representation [27]. This trend is notable in the field of neural surface reconstruction [29, 38, 46, 48]. Unlike Neural Radiance Fields (NeRF) [27], which employs a volume density to infer 3D geometry implicitly, the use of Signed Distance Function (SDF) field has proven to be effective for learning surfaces because of their universal zero level-set definition of surfaces. However, as these approaches optimize on a single scene with dense camera viewpoints, there is an increasing effort to learn a generalizable representation that can adaptively reconstruct geometry from unseen scenes with sparse camera viewpoints [23, 25, 32, 43].

Despite the remarkable achievements, existing generalizable implicit surface reconstruction methods use the fixed, best combination for each camera based on the view-selection scores [44]. This score is predefined for each camera pair via a matching-based reconstruction, such as

COLMAP [34]. These *favorably* selected views usually share a significant amount of overlaps, which can help the model easily capture the 3D geometry of the target scene. We argue that this assumption limits the practicality of surface reconstruction tasks in two key scenarios: (1) where inferring from arbitrary combinations of camera views that were not available during training and (2) where the combination of views share a limited amount of overlapping areas with each other for reconstruction.

To verify this argument, we start by introducing the view-combination score (*VC* Score) that measures the informativeness of the input image set for reconstruction. We demonstrate that conventional research gradually faces higher difficulty as the *VC* score decreases. We then introduce our proposed framework, UFORecon, a robust view-combination generalizable surface reconstruction framework. More specifically, we learn the matching features between source view images using an attention-based cross-view transformer. Then, we build a cascaded correlation frustum based on each source view, effectively combining MVS with implicit surface reconstruction for robust reconstruction from varied view combinations. Furthermore, we encode the explicit similarity for all input pairs and intermediate depths into sample points to predict a 3D representation. As a result, we achieve effective view-combination generalizability without additional data or losses. Finally, we introduce a random set training strategy for robustness under arbitrary view combinations.

As shown in Figure 1, while existing methods suffer from degraded performance under the unfavorable set, our proposed method achieves a reasonable geometry. We attribute this to our method's capability to learn the correlation among source view images.

In summary, our contributions are as follows:

- We propose a new concept of *view-combination generalizability*, which represents an ability to reconstruct geometry under arbitrary and unfavorable views, in generalizable neural scene representation.

- We propose an effective model that integrates cross-view features with correlation volumes and explicit feature similarities to learn cross-view interaction and reconstruct under arbitrary and unfavorable sets.

- We validate the efficacy of random set training as a robust strategy to stably enhance view-combination generalizability.

2. Related Works

2.1. Depth Map-based Multi-view Stereo

Multi-view stereo (MVS) is a branch of the 3D reconstruction task that estimates the underlying geometry of a scene or an object given images captured from multiple views. The common method for classical MVS is finding correspondence or matching features of multi-view images and projecting them into a certain 3D representation [36]. Among various representations (*e.g.*, point clouds [9, 20], voxel grids [14, 16], depths map [2, 10, 33]), depths map-based MVS has shown to be effective in recent learning-based MVS thanks to their robustness and flexibility. MVS-Net [44] proposes to encode camera parameters and deep image features with differentiable homography to build 3D cost volume. CasMVSNet [11] further reduces the memory consumption of the cost volume by proposing a coarse-to-fine multi-stage depth estimation framework. Trans-MVSNet [8] proposes a feature-matching transformer to extract cross-view correlation and robust global context across multi-view images. Though these methods have shown promising results, they show limited results in non-Lambertian regions, and their estimation capability is limited to estimating the depths of known images, while volume rendering-based implicit methods could render color or depth images of novel viewpoints.

2.2. Implicit Neural Scene Reconstruction

Implicit neural representations for 3D scenes have gained significant attention since NeRF [27], which represents the 3D scene as an implicit continuous signal, approximated with neural networks, *e.g.*, multi-layer perceptrons (MLPs). Especially in surface reconstruction, NeuS [38] and VolSDF [46] exploit implicit neural representations for signed distance fields (SDFs) in 3D scenes. Utilizing geometric cues [48] (e.g., depth, normal) or adding constraints [35] has proven effective in sparse-view reconstruction. However, these approaches barely consider the relationship among source views and involve intensive test-time optimization. To address this, generalizable neural scene representation methods construct feature volumes [4, 47] or aggregate source views using transformer attention modules [31, 39]. Particularly for surface reconstruction, SparseNeuS [25] constructs hierarchical feature volumes using deep image features to enable geometry-aware reasoning. VolRecon [32] and ReTR [23] utilize transformers to effectively fuse multi-view image features. C2F2NeuS [43] effectively integrates MVS and implicit surface reconstruction by building cascaded frustum to encode global geometry features. However, these methods' performance is often bounded by camera sets with large overlaps and exhibits sensitivity to viewpoint variations. In our research, we model feature correlation among images and employ view-consistent matching prior to effectively handling arbitrary or unfavorable view sets.

2.3. Feature Matching Transformer

Transformers, which primarily utilize attention mechanisms [37], have been increasingly employed in diverse 3D understanding applications. Notable among these are
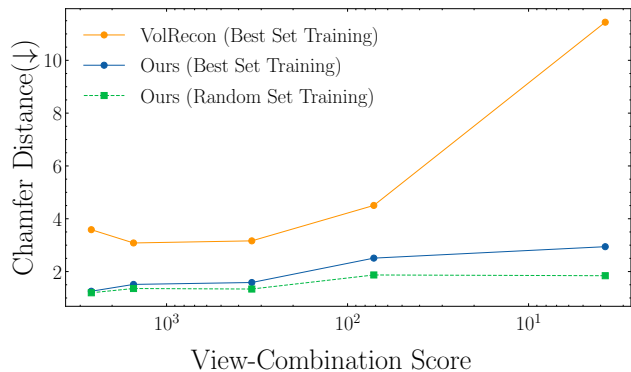
Figure 2. **Comparison of Chamfer Distance (CD) by view-combination (*VC*) scores for generalizable implicit surface reconstruction methods.** We define the VC score to represent the informativeness of view combinations in reconstruction. The higher VC score represents a more favorable combination. Our method shows better generalizability and accuracy over VolRecon [32] across all VC scores. Our random set training (Sec. 4.6) further improves the view-combination generalizability.

transformer-based MVS methods introduced by STTR [21], MVSTER [40], and TransMVSNet [8]. These methods significantly improve feature matching by employing both self- and cross-attention techniques to effectively train on features from multiple viewpoints. Other than the MVS methods, GMFlow [41] exploits a similar transformer-based architecture to estimate optical flow by reformulating the estimation as a global image feature-matching problem between frames. More recently, MatchNeRF [5] introduces cross-view feature matching from the pre-trained transformer in GMFlow [41] as an explicit geometry guidance for generalizable Novel View Synthesis (NVS). Despite these advancements, the exploration of cross-view feature matching in implicit 3D reconstruction remains under-developed. Our approach effectively combines transformer-based feature correlation with neural implicit surface methods, significantly enhancing view-combination generalizability and 3D reconstruction.

3. Motivation

The existing generalizable methods adhere to two critical assumptions that limit their ability to effectively reconstruct from varying view combinations. The assumptions are: 1) The network is trained on images with the best view-selection scores for the specific target view. 2) The evaluation similarly assumes the favorable combination for reconstruction by selecting the source cameras with high view selection scores or selecting the neighboring cameras to the target camera. We define a property of a view-combination that deviates from such assumptions as *unfavorable*. In contrast, favorable combinations tend to share a relatively large

overlap in their camera view frustum, and the respective camera positions are likely to be cluttered (*i.e.*, small baseline and rotation). We observe that a network trained only with such combinations easily overfits, restricting generalizability under arbitrary view-combination variations.

To measure the effectiveness of a view-combination for the reconstruction, we define a new metric, *view-combination score* (*VC* score), to evaluate the collective informativeness of a viewpoints combination for 3D reconstruction. We derive the score from view selection score [44], which represents the usefulness of an image pair for reconstruction. The view selection score [44] $s(i, j)$ between two cameras i and j are defined as follows:

$$s(i,j) = \sum_{\mathbf{p}} (\mathcal{G}(\theta_{ij}(\mathbf{p})), \tag{1}$$

where $\mathcal{G}$ is a piecewise Gaussian function that favors a certain baseline angle θ_{ij}, which denotes the angle between the two camera centers. $\mathbf{p}$ is a common track in both view i and j obtained with COLMAP [34]. The complete derivation process of the view-selection score and the hyperparameters is thoroughly detailed in the Appendix. We extend it to a set of multiple images by averaging all of the pairwise scores between images in the set as follows:

$$VC = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s(i,j). \tag{2}$$

To show the validity of the VC score and verify the conventional approach suffers from the degradation of performance under unfavorable combinations, we test the view-generalizability of the previous approach in various VC score levels. As depicted in Figure 2, Chamfer Distance (CD) incrementally rises as the VC score reduces in the previous state-of-the-art (SoTA) reconstruction method [32]. Such result reflects the necessity of providing a prior on the informativeness of image combination for robust reconstruction. Inspired by these findings, our objective centers on enhancing *view-combination generalizability* by providing cross-view correlation features as robust prior for the image combination, boosting the quality of reconstructions under arbitrary and unfavorable sets of images.

4. Methods

In this section, we first introduce the general formulations of volume rendering-based generalizable surface reconstruction (Sec. 4.1). Then, we introduce the overall structure of our proposed UFORecon (Figure 3) which consists of cross-view matching feature extraction (Sec. 4.2), cascaded correlation volume generation strategy (Sec. 4.3) that encodes global geometry across input views, explicit 2D geometry-aware matching priors (Sec. 4.4) that provides robust view-pair guidance, and view aggregation transformer
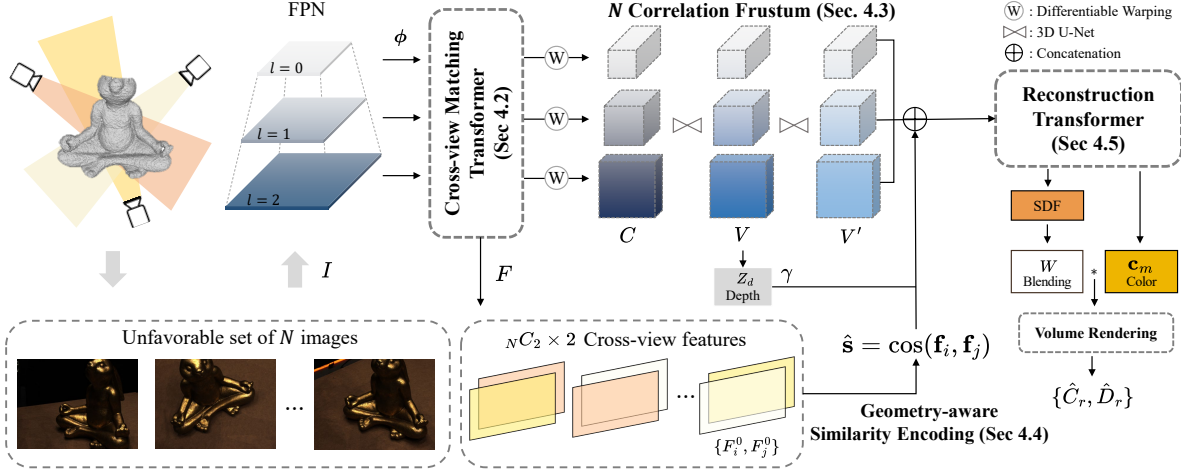
Figure 3. **Overall pipeline of UFORecon**. Our cross-view matching transformer extracts cross-view matching features from multi-level image features of the given image set (Sec. 4.2). Cross-view matching features are then represented as 3D volumes (*i.e.*, cascaded correlation frustums, (Sec. 4.3) and as the 2D features (*i.e.*, geometry aware similarity encoding, (Sec. 4.4). Our reconstruction transformer (Sec. 4.5) fuses various representations of matching features and geometry features (*i.e.*, depth) for volume rendering and color blending.

(Sec. 4.5) to aggregate global volumes and 2D priors with self-attention transformers. Finally, we introduce our training schemes (Sec. 4.6).

4.1. Preliminaries

The goal of generalizable neural surface reconstruction is to reconstruct the underlying geometry of an arbitrary scene with the posed images $\{I_i, P_i\}_{i=1}^N$ where $I_i \in \mathbb{R}^{H \times W \times 3}$ is the i-th view's image and $P_i \in \mathbb{R}^{3 \times 4}$ is the corresponding camera parameters. A general framework extracts features from 2D images and builds a global feature volume [23, 25, 32, 43]. Then the NeRF-like MLP decoder leverages these features with interpolation to guide 3D representations (*e.g.*, radiance, SDF). To generalize through scenes and to enable sparse-view inference, a limited number of source images (*e.g.*, $N = 3$) are utilized to synthesize a target image during training. Along a ray emitted from a given pixel coordinate in the target image, M 3D points are sampled as $\{\mathbf{p}_m = r(t_m) = \mathbf{o} + t_m \mathbf{v}, \ m = 1, ..., M\}$, where $\mathbf{o}$ and $\mathbf{v}$ denote the ray origin and direction respectively. To estimate the color and SDF value of 3D point $\mathbf{p}_m$, we acquire the corresponding image features and volume features. We project $\mathbf{p}_m$ onto the source image I_i as $\pi_i(\mathbf{p}_m)$ and obtain the corresponding image feature $\mathbf{f}_i^{\text{img}}$ with bilinear interpolation. In addition, the volume features $\mathbf{f}^{\text{vol}}$ are obtained with trilinear interpolation at the normalized position. The color of the 3D position $\mathbf{c}_m$ is determined as a weighted sum of image pixel colors $I_i(\mathbf{p}_m)$, where the weights are estimated by $W(\cdot)$.

$$\mathbf{c}_m = \sum_{i=1}^N W\left(\mathbf{f}^{\text{vol}}(\mathbf{p}_m), \{\mathbf{f}_i^{\text{img}}(\pi_i(\mathbf{p}_m))\}_{i=1}^N\right) I_i(\mathbf{p}_m). \tag{3}$$

The predicted 3D representations are aggregated into the 2D domain with volume rendering as follows:

$$\hat{C}(r) = \sum_{m=1}^M T_m \alpha_m \mathbf{c}_m, \tag{4}$$

where $T_m = \prod_{k=1}^{m-1}(1 - \alpha_k)$ denotes the accumulative transmittance, and α_m indicates discrete opacity values defined as follows:

$$\alpha_m = 1 - \exp\left(-\int_{t_m}^{t_{m+1}} \rho(t)dt\right), \tag{5}$$

where $t_m \leq t_{m+1}$. The derivation of $\rho(t)$ commonly follows the original definition in NeuS [38] or VolSDF [46].

4.2. Cross-View Matching Transformer

When input views become sparse, where the 3D space is under-constrained, it is important to identify correlations between the input views [35]. Furthermore, as we assume that a view combination can be different, uncovering the relationships between input images and using them as reconstruction priors becomes more crucial for view-combination generalizability. While most existing neural surface reconstruction methods extract image features independently, recent works [8, 42] have shown significant results by explicitly extracting cross-view correlation features of an input pair of images using transformer-based feature matching.

We initially extract multi-scale features ϕ_i^l independently for each image I_i using a Feature Pyramid Network (FPN) [24]. For each level l, the hierarchical features are then processed through a cross-view matching transformer $\mathcal{T}_M$ in a pairwise manner $F_i^l = \sum_{j \neq i} \mathcal{T}_M(\phi_i^l, \phi_j^l)$ for two

purposes: 1) to construct correlation frustums that learn global implicit correlations among multi-view source images (Sec. 4.3), and 2) to extract explicit feature similarity scores (Sec. 4.4) for all source image pairs. Since both branches learn the correlation between source images, the framework seamlessly integrates the two parts with interdependent objectives, effectively serving as a global matching encoder.

4.3. Cascaded Cross-View Correlation Frustum

Motivated by recent MVS methods that have shown the effectiveness of cost volume integration in generalizing sparse-view inference from arbitrary unseen scenes, we build a cascaded global volume with correlation features [8] to provide cross-view information. Our model constructs perspective frustums instead of regular Euclidean volumes for each source view by setting each view as a reference following C2F2NeUS [43]. For each reference image I_i, we align rest of the source images using differentiable warping to obtain total N pairwise correlation frustums $C_i^l \in \mathbb{R}^{1 \times d \times h \times w}$ for all levels L as follows:

$$C_i^l = \sum_{\substack{j=1 \\ j \neq i}}^{N} \langle F_i^l, \hat{F}_{j \to i}^l \rangle, \quad l = 1, \dots, L \qquad (6)$$

where F_i^l denotes reference-view feature and $\hat{F}_{j \to i}^l$ denotes cross-view feature F_j^l warped to i-th view from our cross-view matching transformer, and d, h, and w denote the resolution of the volume. We deliver the detailed aggregation of correlation volumes in the Appendix. Multi-level correlation frustums are processed with 3D CNNs and output intermediate volumes $V_i^l \in \mathbb{R}^{c \times d \times h \times w}$ and depths $D_i^l \in \mathbb{R}^{1 \times h \times w}$. The intermediate volumes V_i^l, initially tailored for depth estimation, are transformed with additional 3D CNNs to generate V'^l_i. We output our global feature volumes by adding V'^l_i from all source views as $V'^l = \sum_{i=1}^{N} V'^l_i$, serving as a global correlation. Our correlation frustum is different from C2F2NeUS [43] in that we additionally utilize cross-view matching features from transformers and single-channel correlation frustums with much lower memory cost. Furthermore, we believe that correlation volumes guide our model to leverage the inherent geometrical relationships between different views extensively.

4.4. Geometry-aware Similarity Encoding

Recently, correspondence matching information on 2D has shown to be useful for 3D understanding [5, 41]. To further guide our network for view-combination generalizability, we explicitly encode cosine similarity for all possible source pairs. We use the extracted cross-view features for all $\binom{N}{2}$ pairs from the cross-view matching transformer. We only use the coarsest feature for similarity encoding.

To obtain the 2D cross-view features for the sampled 3D points $\mathbf{p}_m$, we project the 3D points onto the 2D cross-view matching features F_i^0 and F_j^0 and acquire the corresponding features $\mathbf{f}_i = F_i^0(\pi_i(\mathbf{p}_m))$ and $\mathbf{f}_j = F_j^0(\pi_j(\mathbf{p}_m))$. After that, we calculate the cosine similarity of the two vectors $\hat{\mathbf{s}} = \cos(\mathbf{f}_i, \mathbf{f}_j)$ in a group-wise manner [12]. We average the cosine similarities for all pairs and obtain encoded similarity vector $\mathbf{f}_s$, allowing the dimension invariant to input source images. Given an arbitrary view combination for inference, the explicit similarity scores provide a robust geometry prior, regularizing the feature space for matching.

4.5. Reconstruction Transformers

View-aggregation Transformer. We combine global correlation from our correlation frustums, cross-view matching features, and similarity scores using our view-aggregation transformer $\mathcal{T}_A$. We acquire 2D cross-view matching features and similarity scores for all sampled points in the 3D space by projecting the points onto each image view and applying bilinear interpolation. Additionally, we employ trilinear interpolation on the points to feature volumes V'^l to obtain global correlation features. Finally, we add a learnable token feature $\mathbf{f}_0$ to capture global consistent rendering features as in [23, 32]. All features above are aggregated via linear self-attention transformers [17, 37]:

$$\mathbf{f}^p, \{\mathbf{f}'_i\}_{i=1}^N = \mathcal{T}_A(\mathbf{f}_0, \{\mathbf{f}_i\}_{i=1}^N, \mathbf{f}_v, \mathbf{f}_s), \qquad (7)$$

where $\mathbf{f}^p$ is a projection vector and $\{\mathbf{f}'_i\}_{i=1}^N$ are aggregated features further used for rendering and $\mathbf{f}_v$ denotes the concatenated volume features sampled from V'^l for all levels.

Geometry-aware Ray Transformer. To estimate implicit geometry representation (*e.g.*,SDF) along the rays, we first apply positional encoding to the points along the ray. This is achieved by calculating the difference between the camera coordinates of sampled points and the intermediate depth obtained from the cascaded correlation frustum as in [44]. We then apply positional encoding [17] $\gamma(\cdot)$ to embed the order of the ray and to extract geometry-aware features $\mathbf{f}'_i$ with the self-attention linear transformer $\mathcal{T}_R$ as follows:

$$\{\mathbf{f}'_i\}_{i=1}^M = \mathcal{T}_R(\mathbf{f}_p, \text{cat}(\gamma(z_i - z_d))_{i=1}^M). \qquad (8)$$

where z_i denotes depth values of a i-th sampled point in the camera coordinate and z_d denotes the intermediate depth value of the ray. We utilize color blending to obtain color value for each point following [32] and aggregate all points along the ray with volume rendering as in Eq. 4. We adopt the method of NeuS [38] to volume render SDF.

4.6. Training of UFORecon

Random Set Training. First, we follow the common protocol of training with a pre-determined best view combination

based on the view-selection score [44] and expect generalization to arbitrary or unfavorable camera configurations not seen during training. Beyond the existing training strategy, we suggest a random set training, which randomly selects N source views in the training phase, expecting the network to handle arbitrary view combinations. A detailed explanation and evidence are presented in the Appendix.

Loss function. Our loss function is defined as a weighted sum of two loss functions as:

$$\mathcal{L} = \mathcal{L}_{color} + \alpha \mathcal{L}_{depth}, \tag{9}$$

where $\mathcal{L}_{color}$ is formulated to minimize the ground truth colors and rendered color. Similarly, $\mathcal{L}_{depth}$ is formulated to minimize the ground truth depths and rendered depth, following VolRecon [32].

5. Experiments

In this section, we first explain our experimental settings, including datasets, baseline, and implementation details (Sec. 5.1). Second, we qualitatively and quantitatively evaluate our method on generalizable surface reconstruction mainly focusing on view-combination generalizability (Sec. 5.2). Last, we analyze the effectiveness of our individual elements (Sec. 5.3).

5.1. Experimental Settings

Datasets. We report our results on the DTU datasets [1], for the challenging scenario of 3 input views in unseen scenes following [23, 25, 32, 43]. The DTU dataset comprises scenes captured from 49 distinct frontal viewpoints, each accompanied by its respective camera matrix. We use 3 source views from each test scene to build a combination and evaluate the generalization ability of our method. For all possible view combinations, we calculate the VC scores and rank them based on the scores in descending order. Subsequently, we evenly divide these combinations into three groups: *favorable*, *normal*, and *unfavorable*.

Baseline. To demonstrate the effectiveness of our method, we mainly compare it with (1) SparseNeus [25], VolRecon [32], ReTR [23], and C2F2NeuS [43] the state-of-the-art generalizable neural implicit reconstruction method. For evaluating scene and viewpoints generalizability, we additionally compare with (2) Generalizable neural rendering methods [4, 39, 47], and (3) Neural implicit reconstruction [38, 46] which require per-scene training. Additionally, we include (4) traditional MVS methods [8, 34].

Implementation Details. In training, we take $N = 4$ source images and one target image with the resolution of 640×512. We train the network in an end-to-end manner with Adam optimizer [18] on a single RTX 4090 GPU. The learning rate is set to 1e-4. We use a hierarchical sampling strategy in both training and testing with 64 points for both coarse and fine sampling. For global feature volume, we use $L = 3$ for cascaded strategy and follow the resolution of TransMVSNet [8] and set the number of depth hypotheses for each level to 48, 32, and 8 respectively. More implementation details can be found in the Appendix.

5.2. Evaluation on Surface Reconstruction

Chamfer Distance (CD) is used as an evaluation metric in all experiments. Note that for fair comparisons, our training adheres to the pre-determined best view combinations, aligning with the existing protocols [23, 25, 32, 43] unless otherwise specified.

Reconstruction on Favorable Sets. Following the existing best-selected test protocol [23, 25, 32, 43], we evaluate with the fixed favorable views for all test scenes. As shown in Table 1, our approach outperforms the state-of-the-art models even in the conventional setting. It implies that the superiority of our model is also consistent in all *VC* score levels.

Reconstruction on Unfavorable Sets. To analyze the proposed view-combination generalizability, we test with *unfavorable* sets. To this end, we sample a view set consisting of 3 images from the unfavorable view group and use the same set for all 15 test scenes. As shown in Table 2, our method consistently achieves significantly better performance in all scenes. Our method achieves better view-combination generalizability by capturing robust prior for image combination. In addition, we qualitatively compare our method in various *VC* score levels. As shown in Figure 4, ours shows much better reconstruction quality compared to existing works, particularly showing huge advantages in unfavorable sets. We further tested with unfavorable sets of BlendedMVS dataset without fine-tuning as shown in Figure 5. It shows that our method is generalizable to unfavorable view-combinations of other datasets.

5.3. Analysis on UFORecon

Ablation Study. We first perform ablation experiments to assess the impact of each component within our approach. Table 3 shows the performance in an unfavorable scenario as specific components are removed from the framework. Each component contributes to the overall performance. Each level of correlation frustum contributes to better performance. Encoding similarity further improves the performance even with a full level of correlation frustums. Also, note that depth supervision plays an important role in our framework similar to [23, 32].

Analysis on Correlation Frustums. A higher level of frustum encodes relatively finer and local details while a lower level encodes global geometry features. As shown in Table 3, the performance degradation is more severe when excluding lower levels.

Analysis on Explicit Similarity Encoding. We conduct a focused analysis on the impact of similarity encoding by

Scan	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean (CD)↓
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
TransMVSNet [8]	1.07	3.14	2.39	1.30	1.35	1.61	0.73	1.60	1.15	0.94	1.34	0.46	0.60	1.20	1.46	1.35
VolSDF [46]	4.03	4.21	6.12	1.63	3.24	2.73	2.84	1.63	5.14	3.09	2.08	4.81	0.60	3.51	2.18	3.41
NeuS [38]	4.57	4.49	3.97	4.32	4.63	1.95	4.64	3.83	5.40	5.60	6.47	6.68	2.96	5.57	6.11	4.90
SparseNeuS-ft [25]	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 [47]	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 [39]	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 [25]	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 [23]	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
C2F2NeuS [43]	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
Ours	**0.76**	**2.05**	**1.31**	_0.82_	**1.12**	**1.18**	_0.74_	_1.17_	**1.11**	**0.71**	**0.88**	_0.58_	0.54	**0.86**	0.99	**0.99**
Ours*	_0.77_	_2.10_	_1.34_	0.87	_1.15_	_1.16_	**0.71**	1.25	_1.17_	0.81	_0.90_	**0.57**	_0.51_	**0.86**	_0.97_	1.01

Table 1. **A quantitative results on favorable sets.** Our method outperforms traditional 3D reconstruction methods (first section), per-scene surface reconstruction methods (second section), generalizable novel-view synthesis methods (third section), and generalizable surface reconstruction methods (fourth section). The **Bold** numbers mean the best, and the underlined values indicate the second-best scores. (*) denotes the use of a random set training strategy.

Scan	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean (CD)↓
TransMVSNet [8]	8.32	8.68	7.46	8.16	8.22	6.74	7.74	8.36	10.67	9.70	8.86	7.99	8.32	8.90	7.13	8.35
SparseNeuS [25]	5.24	5.00	5.76	4.95	3.80	4.12	3.87	3.44	3.48	3.23	4.61	4.24	2.13	4.04	4.46	4.16
VolRecon [32]	3.43	3.64	4.26	4.63	2.43	3.40	2.81	2.41	2.36	2.49	3.79	3.55	1.44	3.60	3.38	3.18
ReTR [23]	3.00	3.98	3.78	4.22	2.22	2.93	3.00	2.51	2.24	2.36	2.36	3.92	1.63	2.83	3.07	2.94
Ours	_1.39_	_2.25_	_1.65_	_1.96_	_1.53_	_1.61_	_1.22_	_1.92_	_1.36_	_1.66_	_1.75_	_1.29_	_0.73_	_1.70_	_1.39_	_1.56_
Ours*	**1.31**	**2.00**	**1.41**	**1.36**	**1.24**	**1.58**	**1.06**	**1.44**	**1.37**	**0.99**	**1.45**	**0.96**	**0.58**	**1.34**	**1.09**	**1.28**

Table 2. **A quantitative results on unfavorable sets.** (*) denotes the use of a random set training strategy.

Methods	Unfavorable
w/o CF	3.19
CF ($L = 1$)	1.91
CF ($L = 2$)	1.71
w/o similarity encoding	1.62
w/o L_{depth}	2.26
UFORecon	1.56

Table 3. **Ablation study on each element in UFORecon.** CF denotes cross-view correlation frustum. All experiments are tested with unfavorable sets.

Method	Favorable	Unfavorable	#params
VolRecon [32]	1.38	3.18	0.88 M
VolRecon [32] + MatchNeRF [5]	1.26	2.47	4.97 M
Ours	0.99	1.56	1.5 M

Table 4. **Effect of explicit similarity score.** We combine two representative baselines in generalizable surface reconstruction [32] and correspondence matching-based NVS method [5]. This hybrid model shows 10% improvements in favorable sets and 22% in unfavorable sets. Our method outperforms this hybrid model with better efficiency.

contrasting our approach with a hybrid model. Specifically, we integrate MatchNeRF [5], known for its utilization of explicit similarity scores for generalizable rendering, with VolRecon [32]. This combined model is then tested on both favorable and unfavorable datasets. As indicated in Table 4, incorporating explicit feature similarity notably enhances the quality of generalizable surface reconstructions espe-

cially in unfavorable conditions. This reflects that explicitly feature similarity provides robust view-combination prior. Nevertheless, our proposed approach outperforms in both favorable and unfavorable conditions, effectively capturing correlation features with much fewer parameters.

Analysis on Random Set Training. We analyze the effect of the random set training. We expect the network to generalize better to various combinations without losing accuracy. As shown in Figure 2, it consistently improves the reconstruction quality at all *VC* score levels. In unfavorable sets, random set training shows 17% of improvements as shown in Table 2. Qualitatively, random set training consistently achieves more complete reconstructions as shown in Figure 6. However, we find that directly applying random set training on the previous method [32] destabilizes training dynamics and leads to a suboptimal result. We conjecture that utilizing cross-view correlation features guides the network to stably improve the reconstruction performance.

6. Conclusion

In this paper, we have introduced a novel concept of *view-combination generalizability* in the field of generalizable implicit surface reconstruction. Our analysis showed that existing methods tend to overfit to specific learned view combinations, resulting in suboptimal performance with unfavorable combinations. To address this challenge, we have proposed a novel framework, UFORecon, that leverages cross-view matching features to learn correlation among in-
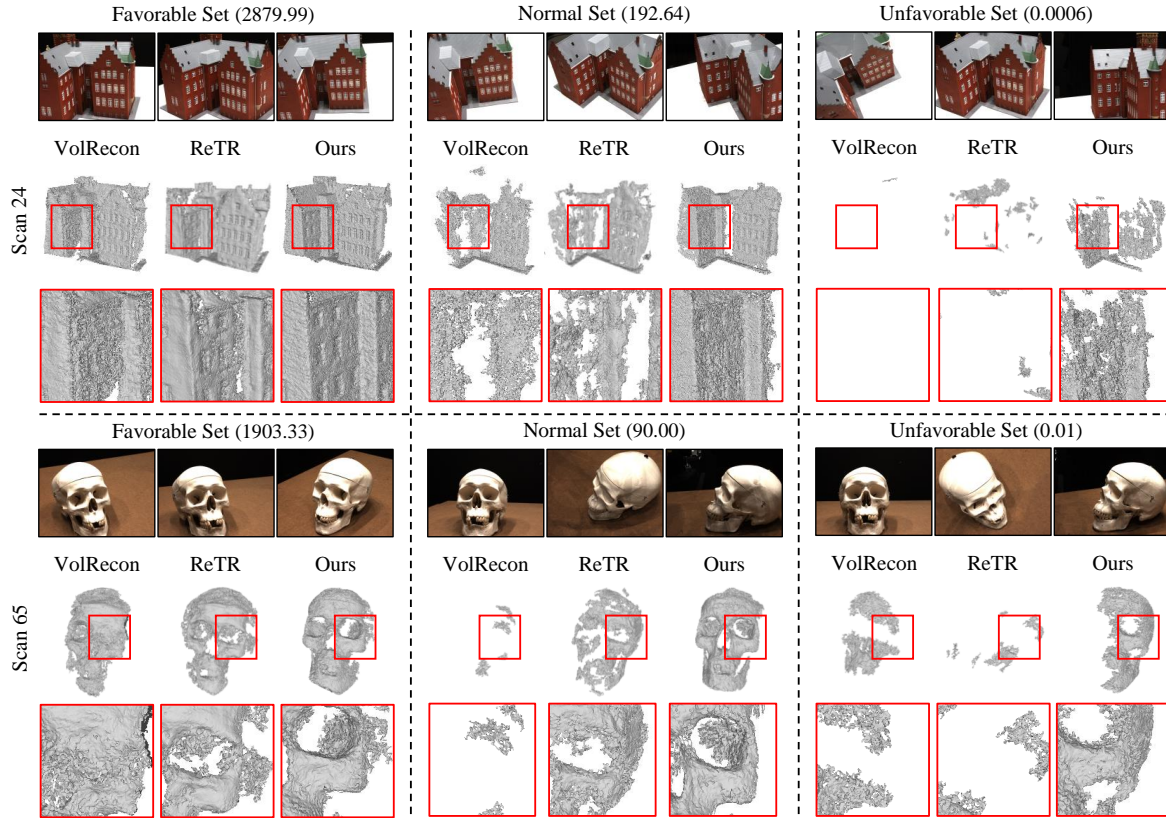
Figure 4. **A qualitative results of Surface Reconstruction across various VC Levels.** The numbers in parentheses denote the view-combination score. Our method consistently outperforms previous methods at all levels and in different scenes. More qualitative results can be found in the Appendix.
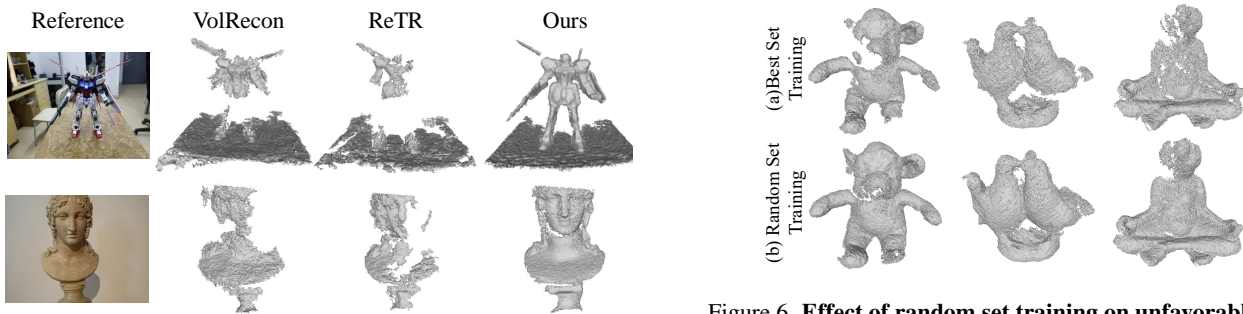

Figure 5. **Unfavorable 4-views tests on BlendedMVS dataset.**

Figure 6. **Effect of random set training on unfavorable sets.**

put images. This is achieved by building correlation frustums with cross-view features and encoding explicit 2D matching similarities across all source image pairs. Notably, our method demonstrates superior performance over previous surface reconstruction methods, not only in unfavorable view combinations but also in favorable ones.

Limitations and future work. Although our method achieves advancements in view-combination generalizability for surface reconstruction, it is still challenging to recon-struct scenes with much higher complexity. In addition, we assume that the ground-truth camera poses are given, We include a detailed discussion on this in the Appendix. We hope that our contributions will inspire further research to extend this to such directions.

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*, 120:153–168, 2016. 6

[2] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, pages 766–779. Springer, 2008. 1, 2

[3] Di Chang, Aljaž Božič, 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*, pages 665–680. Springer, 2022. 1

[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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2, 6, 7

[5] Yuedong Chen, Haofei Xu, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Explicit correspondence matching for generalizable neural radiance fields. *arXiv preprint arXiv:2304.12294*, 2023. 3, 5, 7

[6] Qiyu Dai, Yan Zhu, Yiran Geng, Ciyu Ruan, Jiazhao Zhang, and He Wang. Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1757–1763. IEEE, 2023. 1

[7] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 1

[8] 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. 1, 2, 3, 4, 5, 6, 7

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

[10] 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

[11] 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. 1, 2

[12] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3273–3282, 2019. 5

[13] Christian Häne, Torsten Sattler, and Marc Pollefeys. Obstacle detection for self-driving cars using only monocular cameras and wheel odometry. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5101–5108. IEEE, 2015. 1

[14] 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

[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. 1

[16] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30, 2017. 2

[17] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020. 5

[18] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015. 6

[19] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International journal of computer vision*, 38:199–218, 2000. 1

[20] 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

[21] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021. 3

[22] Zhuopeng Li, Lu Li, and Jianke Zhu. Read: Large-scale neural scene rendering for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1522–1529, 2023. 1

[23] 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. 1, 2, 4, 5, 6, 7

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[25] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, pages 210–227. Springer, 2022. 1, 2, 4, 6, 7

[26] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt. Scalable 6-dof localization on mobile devices. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 268–283. Springer, 2014. 1

[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. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[28] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*, pages 1347–1356. PMLR, 2022. 1

[29] 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. 1

[30] 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. 1

[31] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 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. 1, 2, 3, 4, 5, 6, 7

[33] 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. 2

[34] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 1, 2, 3, 6, 7

[35] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4190–4200, 2023. 2, 4

[36] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 5

[38] 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. *arXiv preprint arXiv:2106.10689*, 2021. 1, 2, 4, 5, 6, 7

[39] 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, 6, 7

[40] 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

[41] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022. 3, 5

[42] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4

[43] Luoyuan Xu, Tao Guan, Yuesong Wang, Wenkai Liu, Zhaojie Zeng, Junle Wang, and Wei Yang. C2f2neus: Cascade cost frustum fusion for high fidelity and generalizable neural surface reconstruction. *arXiv preprint arXiv:2306.10003*, 2023. 1, 2, 4, 5, 6, 7

[44] 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, 3, 5, 6

[45] 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. 1

[46] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 1, 2, 4, 6, 7

[47] 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, 7

[48] 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. 1, 2

[49] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 1