# Multi-View Reconstruction using Signed Ray Distance Functions (SRDF)

Pierre Zins[1,2] Yuanlu Xu[2] Edmond Boyer[1,3] Stefanie Wuhrer[1] Tony Tung[2]
[1]Inria centre at the University Grenoble Alpes
[2]Meta Reality Labs, Sausalito, USA
[3]Meta Reality Labs, Zurich, Switzerland
name.surname@inria.fr, merayxu@gmail.com, tony.tung@fb.com

## Abstract

*In this paper, we investigate a new optimization framework for multi-view 3D shape reconstructions. Recent differentiable rendering approaches have provided breakthrough performances with implicit shape representations though they can still lack precision in the estimated geometries. On the other hand multi-view stereo methods can yield pixel wise geometric accuracy with local depth predictions along viewing rays. Our approach bridges the gap between the two strategies with a novel volumetric shape representation that is implicit but parameterized with pixel depths to better materialize the shape surface with consistent signed distances along viewing rays. The approach retains pixel-accuracy while benefiting from volumetric integration in the optimization. To this aim, depths are optimized by evaluating, at each 3D location within the volumetric discretization, the agreement between the depth prediction consistency and the photometric consistency for the corresponding pixels. The optimization is agnostic to the associated photo-consistency term which can vary from a median-based baseline to more elaborate criteria,* e.g. *learned functions. Our experiments demonstrate the benefit of the volumetric integration with depth predictions. They also show that our approach outperforms existing approaches over standard 3D benchmarks with better geometry estimations.*

## 1. Introduction

Reconstructing 3D shape geometries from 2D image observations has been a core issue in computer vision for decades. Applications are numerous and range from robotics to augmented reality and human digitization, among others. When images are available in sufficient numbers, multi-view stereo (MVS) is a powerful strategy that has emerged in the late 90s (see [58]). In this strategy, 3D geometric models are built by searching for surface locations in 3D where 2D image observations concur, a property called photo-consistency. This observation consistency strategy has been later challenged by approaches in the field that seek instead for observation fidelity using differentiable rendering. Given a shape model that includes appearance information, rendered images can be compared to observed images and the model can thus be optimized. Differentiable rendering adapts to several shape representations including point clouds, meshes and, more recently, implicit shape representations. The latter can account for occupancy, distance functions or densities, which are estimated either directly over discrete grids or through continuous MLP network functions. Associated to differentiable rendering these implicit representations have provided state-of-the-art approaches to recover both the geometry and the appearance of 3D shapes from 2D images.

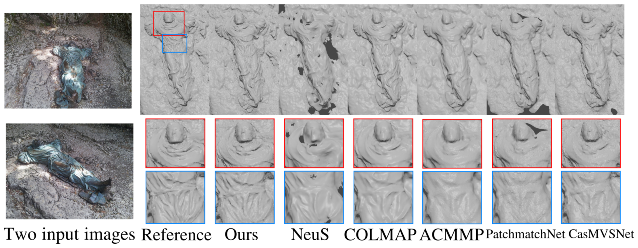

Figure 1. Reconstructions with various methods using 14 images of a model from BlendedMVS [70].

With the objective to improve the precision of the reconstructed geometric models and their computational costs, we investigate an approach that takes inspiration from differentiable rendering methods while retaining beneficial aspects of MVS strategies. Following volumetric methods we use a volumetric signed ray distance representation which we parameterize with depths along viewing rays, a representation we call the Signed Ray Distance Function or SRDF. This representation makes the shape surface explicit with depths while keeping the benefit of better distributed gradients with a volumetric discretization. To optimize this shape representation we introduce an unsupervised differentiable volumetric criterion that, in contrast to differentiable rendering approaches, does not require color estimation. Instead, the criterion considers volumetric 3D samples and evaluates whether the signed distances along rays agree at a sample when it is photo-consistent and disagree otherwise. While being volumetric our proposed approach shares the following MVS benefits:

i) No expensive ray tracing in addition to color decisions is required;
ii) The proposed approach is pixel-wise accurate by construction;
iii) The optimization can be performed over groups of cameras defined with visibility considerations. The latter enables parallelism between groups while still enforcing consistency over depth maps.

In addition, the volumetric scheme provides a testbed to compare different photo-consistency priors in a consistent way with space discretizations that do not depend on the estimated surface.

To evaluate the approach, we conducted experiments on real data from DTU Robot Image Data Sets [23], BlendedMVS [70] and on synthetic data from Renderpeople [3] as well as on real human capture data. Ablation tests demonstrate the respective contributions of the SRDF parametrization and the volumetric integration in the shape reconstruction process. Comparisons with both MVS and differential rendering methods also show that our method consistently outperforms state-of-the-art both quantitatively and qualitatively with better geometric details.

## 2. Related Work

### 2.1. Multi-view Stereo

Reconstructing 3D shapes from multiple images is a long-standing problem in computer vision. Traditional MVS approaches can be split into two categories. Seminal methods [6, 10, 30, 59] use a voxel grid representation and try to estimate occupancies and colors. While efficient, their reconstruction precision is inherently limited by the memory requirement of the 3D grid when increasing resolution. On the other hand depth-based methods [4,7,13,15,57,66] have been proposed that usually try to match image features from several views to estimate depths. Additional post-processing fusion and meshing steps [8, 27, 28, 36] are required to recover a surface from the multi-view depth maps. Despite the usually complex pipeline, multi-view depth map estimation offers the advantage to give access to pixel-accuracy, a strong feature for the reconstruction quality which has made this strategy dominant in MVS approaches. We also consider multi-view depth maps that are however included into a volumetric framework, through signed distances, with the aim to improve global consistency. With the advances in deep learning, several methods propose to learn parts of the MVS pipeline, such as the image feature matching [18,31,76] or the depth map fusion [12,54]. Others even propose to learn the full pipeline in an end-to-end manner [17, 20, 41, 53, 61, 64, 68, 69, 76]. These learning based methods offer fast inference and exhibit interesting generalization abilities. Optimization based methods are alternative or complementary solutions that can provide better precision and generalization abilities, as shown in our experiments in Sec. 5.6.

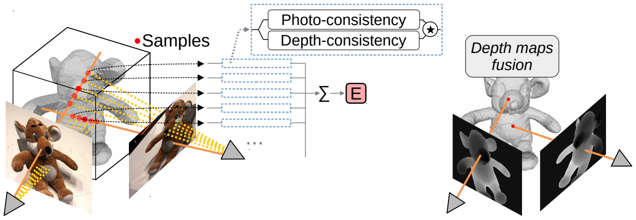

Figure 2. Overview. Left: Given calibrated RGB images and an initial coarse reconstruction, our method optimizes depth maps using a volumetric shape energy $E$ evaluated at samples along viewing lines. Right: The optimized depth maps are further fused into a surface model.

### 2.2. Differentiable Rendering

Another line of works has explored differentiable rendering approaches. Many of these works are used for novel view rendering applications; however several also consider 3D shape geometry reconstruction, often as part of the new image generation process. They build on a rendering that is differentiable and henceforth enable shape model optimization by differentiating the discrepancy between generated and observed images. These methods were originally applied to various shape representations including meshes [19, 26, 33, 48], volumetric grids [14, 25, 43, 63, 77] or even point clouds [21, 24, 42]. In association with deep learning, new neural implicit representations have also emerged. Their continuous nature and light memory requirements are attractive and they have been successfully applied to different tasks including 3D reconstruction [37,47,50,55,56,67] or geometry and appearance representations [16,38,45,60,62]. Most of these methods solve for 3D shape inference and require 3D supervision, however recent works combine implicit representations with differentiable renderering and solve therefore for shape optimization with 2D image supervision. They roughly belong to two categories depending on the shape representation they consider.

**Volume-based methods** use a volumetric renderer with 3D samples and estimate for each sample a density as well as a color conditioned on a viewing direction. Colors and densities of the samples are integrated along viewing rays to obtain image colors that can be compared with the observed pixel colors. The pioneer work of Mildenhall *et al.* [39] opened up this research area with impressive results on novel view synthesis. The quality of the associated geometry, as encoded with densities, is however not perfect and often noisy. Several works have followed that target generalization to new data [22, 75], dynamic scenes [32, 51] or propose new formulations based on Signed Distance Functions to improve the estimated geometry [65, 71]. Darmon *et al.* [9] propose a finetuning strategy based on image warpings to take advantage of high-frequency texture. A main limitation of methods based on neural volumetric rendering lies in the optimization time complexity which often plagues the shape modeling process. To address this limitation various strategies have been investigated: a divide-and-conquer strategy [52], more efficient sampling [5], traditional volumetric representations to directly optimize densities and colors inside octrees [74], or voxel grids [73] and efficient multi-resolution hash tables [40]. Dellaert *et al.* [11] provide more details. These strategies dramatically decrease the optimization time while maintaining very good results for novel view rendering. However the estimated geometry often still lacks precision as the methods are not primarily intended to perform surface reconstruction.

**Surface-based methods** [29, 34, 44, 72] address this issue with a surface renderer that estimates 3D locations and their colors, where viewing rays enter the surface. By making the shape surface explicit, these approaches obtain usually better geometries. Nevertheless, they are more prone to local minima during the optimization since gradients are only computed near the estimated surface as opposed to volumetric strategies. An interesting hybrid approach [46], proposes to combine the advantages of both volumetric and surface rendering and shows good surface reconstructions. Comparisons to state-of-the-art methods are provided in Sec. 5.

# 3. Method

Our method takes as input $N$ calibrated color images $\mathcal{I} = I_{j \in [1,N]}$ and assumes $N$ initial associated depth maps $\mathcal{D} = D_{j \in [1,N]}$ that can be obtained using an initial coarse reconstruction, *e.g.* based on pre-segmented image silhouettes. It optimizes depth values along pixel viewing lines by considering a photo-consistency criterion that is evaluated in 3D over an implicit volumetric shape representation. Final shape surfaces are thus obtained by fusing depth maps, as in *e.g.* [8, 15]. The main features of the method are:

- Shape representation (Sec 3.1): depth maps determine the signed distances, along pixel viewing rays, that define our volumetric shape representation with the SRDF. Parameterizing with depths offers several advantages: it better accounts for the geometric context by materializing the shape surface; it enables pixel accuracy regardless of the image resolution; it allows for coarse to fine strategies in addition to parallelization with groups of views.

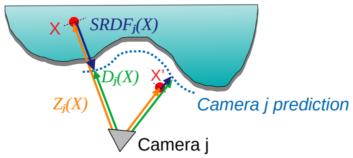

Figure 3. For any 3D point $X$, its ray signed distance $SRDF_j(X) = D_j(X) - Z_j(X)$ with respect to camera $j$ is the signed shortest distance from $X$ to the surface, as predicted by camera $j$, along the corresponding viewing line.

- Energy function (Sec 3.2): our shape energy function is evaluated at sample locations along viewing lines and involves multiple depth maps simultaneously, therefore enforcing spatial consistency. It focuses on the geometry and avoids potential ambiguous estimation of the appearance.

- Photometric prior (Sec 3.3): the photo-consistency hypothesis evaluated by the energy function along a viewing line can be diverse. We propose a criterion that is learned over ground truth 3D data, such as DTU [23]. We also experiment a baseline unsupervised criterion that builds on the median color.

Fig. 2 shows a visual overview of our method.

## 3.1. Signed Ray Distance Function

Our shape representation is a volumetric signed distance function parameterized by depths along viewing rays. This is inspired by signed distance functions (SDF) and shares some similarities with more recent works on signed directional distance functions (SDDF) [78]. Unlike traditional surface-based representations, such a function is differentiable at any point in the 3D observation volume.

Instead of considering the shortest distances along any direction, as in standard SDF or in a fixed direction, as in SDDF [78], we define, for a given 3D point $X$, its $N$ signed distances with respect to cameras $j \in [1, N]$ as the signed distances of $X$ to its nearest neighbor on the surface as predicted by camera $j$ along the viewing ray passing through $X$. We denote the distance for $X$ and camera $j$ by the *Signed Ray Distance Function (SRDF)*, as shown in Fig. 3:

$$SRDF(X, D_j) = SRDF_j(X) = D_j(X) - Z_j(X), \quad (1)$$

where $D_j(X)$ is the depth at the the projection of $X$ in depth map $D_j$ and $Z_j(X)$ the distance from $X$ to camera $j$.

## 3.2. Volumetric Shape Energy

The intuition behind our volumetric energy function is that photometric observations across different views should be consistent on the surface and not elsewhere. Importantly such a behavior should be shared by the SRDF predictions across views that should also consistently identify zero distances for points on the surface and non consistent distances elsewhere. Given this principle, illustrated in Fig. 4, a computational strategy is to look at the correlation between these 2 signals, the observed photo-consistency and the predicted SRDF consistencies, and to maximise it at 3D sample locations $\{X\}$ in the observation space (see Fig. 5). For this purpose, we introduce the following consistency energy:

$$E(\{X\}, \mathcal{D}, \mathcal{I}) = \sum_X C_{SRDF}(X, \mathcal{D})\ C_\Phi(X, \mathcal{I}), \tag{2}$$

where $\{X\}$ are the 3D sample locations, $C_{SRDF}(X, \mathcal{D})$ and $C_\Phi(X, \mathcal{I})$ represent measurements of consistency among the predicted SRDFs values $SRDF_{j\in[1,N]}(X)$ and among the observed photometric observations $\Phi_{j\in[1,N]}(X)$, respectively, at location $X$. Both are functions that return values between 0 and 1 that characterize consistency at $X$. We detail below the SRDF consistency measure $C_{SRDF}(X)$. The photo-consistency measure $C_\Phi(X)$ is discussed in Sec. 3.3. The above energy $E$ is differentiable with respect to the predicted depths values $\mathcal{D}$ and computed in practice at several sample locations along each viewing ray of each camera, which enforces SRDFs to consistently predict surface points over all cameras.

**SRDF consistency** From the observation that SRDF consistency is only achieved when $X$ is on the surface, *i.e.* when $SRDF_j(X) = 0$ for all non occluded cameras $j$, we define:

$$C_{SRDF}(X) = \prod_{j=1}^{N} \Big( exp\Big( -\frac{SRDF_j(X)^2}{\sigma_d} \Big) + \Gamma_{SRDF} \Big), \tag{3}$$

where the ray signed distances are transformed into probabilities using an exponential which is maximal when $SRDF_j(X) = 0$. $\Gamma_{SRDF}$ is a constant that prevents the product over all cameras to cancel out in case of inconsistencies caused by camera occlusions. It can be interpreted as the probability of the $SRDF_j$ value at $X$ knowing camera $j$ is occluded, which can be set as constant for all values. $\sigma_d$ is a hyper parameter that controls how fast probabilities decrease with distances to the surface. It should be noted here that the above energy term $C_{SDRF}$ is a product over views at a 3D point $X$ and not a sum, hence gradients w.r.t. depth values are not independent at $X$, which forces distances to become consistent across views, as shown in the ablation test provided in the supplemental.

## 3.3. Photometric Consistency

Our model is agnostic to the photo-consistency measure $C_\Phi(X)$ that is chosen. In practice we have considered 2 instances of $C_\Phi(X)$: a baseline prior that relies on the traditional Lambertian assumption for the observed surface and a learned version trained with ground truth 3D data.

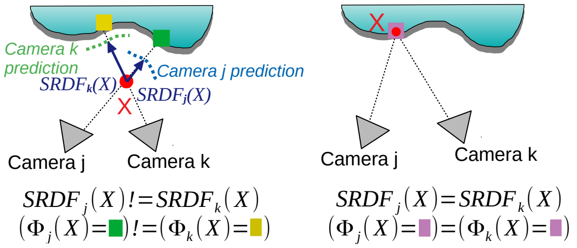

Figure 4. Inconsistency (left) and consistency (right) of the ray signed distances $SRDF_{j,k}(X)$ and of the photometric information $\Phi_{j,k}(X)$ at $X$ with respect to cameras $j$ and $k$.

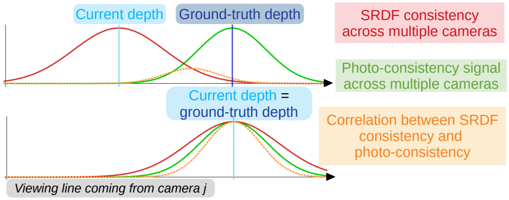

Figure 5. The SRDF consistency (red) and photo-consistency signals (green) along a viewing line. Their cross correlation will be maximal when the current predicted depth aligns with the ground truth depth.

**Baseline Prior** assumes a Lambertian surface and therefore similar photometric observations for points on the observed surface for all non-occluded viewpoints. While ignoring non diffuse surface reflections, the assumption has been widely used in image based 3D modelling, especially by MVS strategies. The associated consistency measure we propose accounts for the distance to the median observed value. Under the Lambertian assumption all observed appearances from non-occluded viewpoints are equal. Assuming that there are fewer occluded viewpoints than visible ones, we define the photo-consistency as:

$$C_\Phi(X) = \prod_{j=1}^{N} \Big( exp\Big( -\frac{\|\Phi_j(X) - \widetilde{\Phi}(X, \mathcal{I}))\|^2}{\sigma_c} \Big) + \Gamma_\Phi \Big), \tag{4}$$

where $\Phi_j(X)$ is photometric observation of $X$ in image $j$, typically a RGB color, and $\widetilde{\Phi}(X, \mathcal{I}))$ is the median value of the observations at $X$ over all images. Similarly to Eq. 3, $\Gamma_\Phi$ is a constant that prevents the product over all cameras to cancel out in case of occlusion and $\sigma_c$ is a hyper-parameter. Sec. 5 shows that this photometric prior yields state-of-the-art results on synthetic 3D data for which the Lambertian assumption holds but is naturally less successful on real data.

**Learned Prior**. In order to better handle real images that are noisy, and for which the Lambertian assumption is often not satisfied, we have experimented a data driven approach. Inspired by previous works [18, 31], we cast the problem as a binary classification between points that are photo-consistent across multiple views and points that are not, and train a network for that purpose.

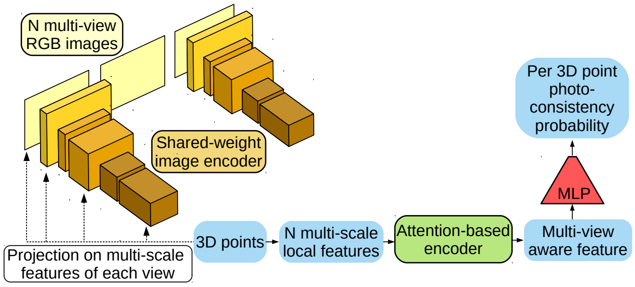
N multi-view RGB images
Shared-weight image encoder
Projection on multi-scale features of each view
3D points
N multi-scale local features
Attention-based encoder
Multi-view aware feature
MLP
Per 3D point photo-consistency probability

Figure 6. Proposed architecture to learn photo-consistency.

As shown in Fig. 6, the network architecture aims to match the local appearance of a 3D point in different views and outputs a photo-consistency score between 0 and 1. This module is independent of the number of cameras, provides good results on real data, and generalizes well as demonstrated in Sec. 5. Please refer to supplementary material for more details about the architecture.

# 4. Implementation

## 4.1. Optimization Pipeline

To allow for efficient processing, we define $G$ groups of cameras, that are optimized in parallel. Since our approach optimizes geometry based on appearance matching, it is advantageous to minimize occlusions. For this reason, we heuristically choose to gather cameras that are close to each other. For DTU and Renderpeople data, in which cameras follow a relatively standard placement and observe the front part of the objects, we use the Euclidean distance between camera positions to calculate distances between pairs of cameras and form distinct camera groups. For DTU, each group contains 7 cameras, and for Renderpeople, we form one group with 10 cameras and one with 9. For the real captured data with irregular camera placement and wider baselines, we follow the more elaborated strategy of MVSNet [68] that computes a score based on sparse 3D points obtained with COLMAP [57] and a piece-wise Gaussian function. We create one group for each camera containing the 6 closest cameras based on the computed scores.

For the depth maps associated to a camera group, at each epoch, we iterate over all rays $r_j^i$ corresponding to foreground pixels $i$ of cameras $j$, as defined by pre-segmented silhouettes, and sample points along $r_j^i$ around the current depth estimation $d_j^i$. This sampling is parameterized by two parameters: an offset $o$ that defines an interval for the sampling around the current depth $[d_j^i - o; d_j^i + o]$, and the density of the sampling which represents the number of points that we sample uniformly in that interval. Ideally, the real depth $\hat{d_j^i}$ should be contained inside the interval $[d_j^i - o; d_j^i + o]$, this to help the appearance guide the geometry optimization. We define a coarse-to-fine strategy for the sampling that aims to capture the ground truth depth in the interval. The sampling density can be adjusted in the same way but decreasing the size of the sampling interval already indirectly increases its density, so in practice we keep the sampling density constant.

Our shape energy, described in Sec. 3.2, is computed over all the samples from all the rays of each camera. The gradients are computed using Pytorch autodiff [49] and back-propagated to update depth maps.

## 4.2. Photo-consistency Network

To train the photo-consistency network, we use the DTU Robot Image Data Sets [23] composed of 124 scans of objects. For each scan, there are: 49 or 64 images under 8 different illuminations settings; camera calibration and ground truth point cloud obtained from structured light. We select 15 test objects and remove all the scans that contain these objects from the training set. This results in 79 training scans. Next, we reconstruct a surface from the ground truth point cloud using the Screened Poisson algorithm [28] and surface trimming of 9.5. From the reconstructed meshes, we render ground truth depth maps and use them to sample points on the surface (positive samples) and points that are either in front or behind the surface (negative samples). We make sure to keep a balanced sampling strategy with an equal number of positive and negative samples.

To encourage the network to remain invariant to the number of cameras, at each training iteration, we randomly select a subset of $K$ cameras from the total $N$ cameras. Matching appearances between cameras too far from each other leads to inconsistencies as a result of the potentially high number of occlusions. To remedy this, we create the camera groups using a soft nearest neighbour approach. We randomly select a first camera, compute its $K'$ closest cameras with $K < K' < N$, and randomly select $K - 1$ cameras from them. In practice, $N = 49$ or 64 and we choose $K \in [4, 10]$ and $K' = min(2K, 15)'$.

# 5. Experimental Results

To assess our method, we conduct an evaluation on multi-view 3D shape reconstructions. First, we introduce the existing methods that we consider as baselines. Then, we present the datasets as well as the evaluation metrics. We provide quantitative and qualitative comparisons against the current state of the art on real images using our learned prior for photo-consistency. We further show that our method, combined with a baseline prior for photo-consistency, provides good reconstruction results under the Lambertian surface assumption. Finally, we demonstrate better generalization abilities of our method compared to deep MVS inference-based methods and that the latter can serve as an initialization. The values of our hyper-parameters for each experiment are available in the supplementary material.

### 5.1. Datasets and Metrics

To evaluate our method on real multi-view images with complex lighting, we use 15 test objects from the DTU Data Sets [23] and BlendedMVS [70]. Note again that BlendedMVS is not used to train our learned photo-consistency prior. For DTU, we use existing corresponding background masks [72]. To test our method with the baseline prior for photo-consistency, we render multi-view images from Renderpeople [3] meshes. This dataset provides detailed textured meshes obtained from 3D scans of dressed humans and corrected by artists. We render 19 high-resolution images (2048x2048) that mostly show the frontal part of the human. For the quantitative evaluation with DTU, we use a Python implementation [1] of the official evaluation procedure of DTU. The accuracy and completeness metrics, with Chamfer distances in $mm$, are computed w.r.t. ground truth point clouds obtained from structured light. Finally, to evaluate generalization, we also experiment with images from a large scale hemispherical multi-view setup with 65 cameras of various focal lengths.

### 5.2. Baseline Methods

To assess our approach, we evaluate the geometry against state-of-the-art methods of 3 categories: classic MVS, deep MVS and differential rendering based methods. First, COLMAP [57] and ACMMP [66] are classic MVS methods that have been widely used and demonstrate strong performances for MVS reconstruction. Among all the deep MVS methods, we consider two of the most efficient methods: PatchmatchNet [64] and CasMVSNet [17], for which code is available. Finally, for the differentiable rendering based methods, we consider IDR [72], which was one of the first works that combines a differentiable surface renderer with a neural implicit representation. It requires accurate masks but handles specular surfaces and has shown impressive reconstruction results. We also compare with two more recent works that use volumetric rendering and provide impressive reconstruction results, NeuS [65] and NeuralWarp [9].

For the evaluation on DTU using all the available views (49 or 64 depending on the scan), we retrain PatchmatchNet and CasMVSNet as their pre-trained models use a different train/test split. We use the pre-trained models for IDR, NeuS (with the mask loss) and NeuralWarp.

To recover meshes with our method, we use a post-processing step with a bilateral filter on the optimized depth maps, a TSDF Fusion [8] method and a mesh cleaning based on the input masks. For COLMAP, ACMMP, PatchmatchNet and CasMVSNet we try to use the same TSDF Fusion method [8] as much as possible. For differentiable rendering based methods (IDR, NeuS and NeuralWarp), the implicit representation is evaluated in a 3D grid of size $512^3$ and Marching Cubes [35] is applied.

### 5.3. Multi-view Reconstruction from Real Data

**Qualitative results**. In Fig. 7, we show comparisons between our method and the baselines. While IDR, NeuS and NeuralWarp produce high quality details, they show some artifacts or misleading parts: some regions of the fruits (1st row), near the right arm of the figurine (2nd row) or at the separation between the belly and the legs of the statue (3rd row). In contrast, our method provides a high level of details without failing on these difficult parts. For IDR and NeuS, the appearance prediction probably compensates for the wrong geometry during the optimization, however our approach exhibits more robustness by focusing on the geometry. Our method produces visual results comparable to COLMAP, ACMMP, PatchmatchNet and CasMVSNet with high fidelity and reduced noise. As shown in Fig. 1 and 8 our method provides similar results with high quality details on BlendedMVS, using the same photoconsistency prior trained on DTU. Note that in Fig. 1, we use a coarse initial reconstruction obtained with PatchmatchNet.

**Quantitative results**. In our quantitative evaluation, results are computed on the meshes obtained with the differentiable rendering based methods, and directly on the point clouds fused from the depthmaps for COLMAP, ACMMP, PatchmatchNet, CasMVSNet and our method. On average, our method clearly outperforms methods based on differentiable rendering (IDR, NeuS and NeuralWarp) in terms of accuracy and completeness. Our method also demonstrates an improvement over classic MVS methods COLMAP and ACMMP. Compared to deep MVS methods that are trained end-to-end on DTU, our approach is on-par, though better on combined accuracy and completeness, while training only a small neural network for photo-consistency that is used as prior in the optimization. Note that quantitative results for PatchmatchNet and CasMVSNet are not as high as in their paper since the training set is not the same and, in contrast to their train/test split, we remove all the scans from the training set in which a test object is seen.

**Runtime analysis**. Table 1 shows that our method is competitive with MVS methods COLMAP and ACMMP in terms of runtime. We note that it could further be parallelized by optimizing different groups of depthmaps on different GPUs, or even servers, which could significantly reduce the computation time. COLMAP, ACMMP and our method run significantly faster than methods based on differentiable rendering and neural implicit representations that require several hours. As expected, deep MVS methods, like PatchmatchNet and CasMVSNet that perform only inference, are inherently significantly faster than optimization based methods.

### 5.4. Reconstruction from Synthetic Data

To demonstrate the validity of our volumetric shape energy, we experiment with the baseline photo-consistency

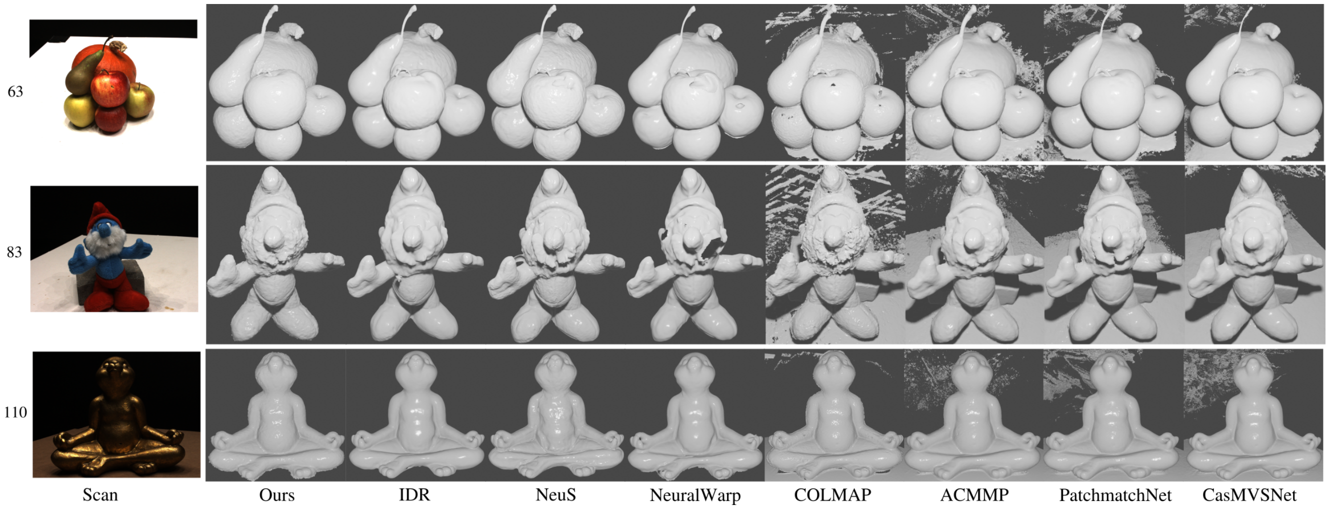

Figure 7. Qualitative comparisons with state-of-the-art methods.

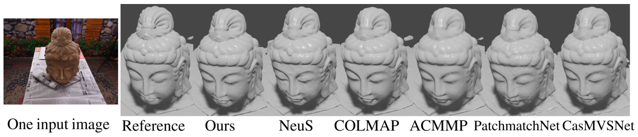

Figure 8. Qualitative comparison using 36 images of a model from BlendedMVS [70].

| Methods | Use masks | Chamfer Distance ↓ | Accuracy ↓ | Completeness ↓ | Time |
|---|---|---|---|---|---|
| IDR [72] | ✓ | 0.89 | 1.02 | 0.79 | 8 h |
| NeuS [65] | ✓ | 0.77 | 0.85 | 0.68 | 14 h |
| NeuralWarp [9] | ✗ | 0.69 | 0.68 | 0.69 | 17 h |
| COLMAP [57] | ✗ | 0.49 | 0.40 | 0.58 | 38 min |
| ACMMP [66] | ✗ | 0.42 | 0.46 | 0.39 | 32 min |
| PatchmatchNet [64] | ✗ | 0.40 | 0.44 | **0.35** | 1 min |
| CasMVSNet [17] | ✗ | 0.38 | **0.34** | 0.43 | 3 min |
| Ours | ✓ | **0.36** | 0.37 | 0.36 | 48 min |

Table 1. Quantitative evaluation on DTU [23] (49 or 64 images per model). Best scores are in **bold** and second best are underlined. Last column: runtime comparison (single GPU, scan 83).

prior defined in Sec. 3.3. We use 19 synthetic images from Renderpeople [3] and compare to COLMAP, ACMMP, IDR, NeuS, PatchmatchNet, and CasMVSNet.

Fig. 9 shows that our method is able to reconstruct very accurate and detailed meshes. COLMAP and ACMMP's results are less detailed and more noisy (*e.g*. COLMAP's bottom row). IDR and NeuS also lack details and even fail to correctly reconstruct the geometry of the jacket on the first row because of the checkered texture. In that case, optimizing both the geometry and the color clearly leads to the wrong geometry. PatchmatchNet and CasMVSNet also work well, leading to results with little noise (*e.g*. feet on the first row) and slightly less pronounced details compared to our method (*e.g*. wrinkles on the scarf and sweater on the first row and on the upper part of the dress on the bottom row). The qualitative results are confirmed by the accuracy and completeness metrics computed between each reconstruction and the ground truth mesh.

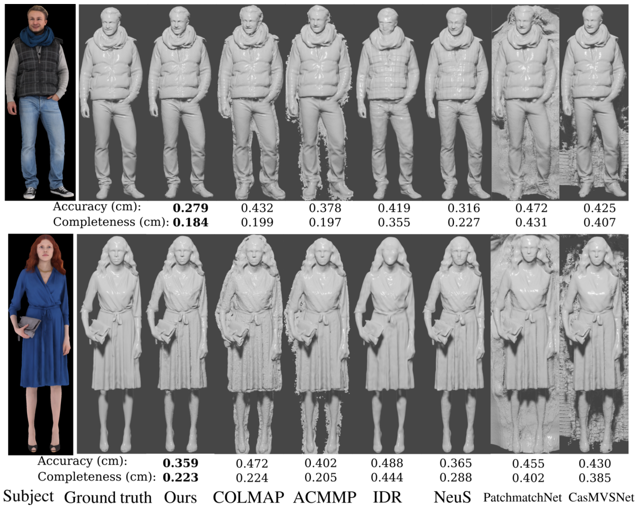
Accuracy (cm): 0.279 0.432 0.378 0.419 0.316 0.472 0.425
Completeness (cm): 0.184 0.199 0.197 0.355 0.227 0.431 0.407
Accuracy (cm): 0.359 0.472 0.402 0.488 0.365 0.455 0.430
Completeness (cm): 0.223 0.224 0.205 0.444 0.288 0.402 0.385
Subject
Ground truth
Ours
COLMAP
ACMMP
IDR
NeuS
PatchmatchNet
CasMVSNet

Figure 9. Qualitative and quantitative results with 19 images (Renderpeople [3]) and using the baseline photo-consistency prior.

### 5.5. Reconstruction from Real Captured Data

To further evaluate the generalization ability of our method, we apply it on human capture data. We use images from the hemispherical multi-camera platform Kinovis [2], composed of 65 cameras of various focal lengths. This setup is designed to capture humans moving in a large scene so the setting is significantly different from DTU with more distant cameras and significantly wider baselines.

Similarly to the previous experiments, we compare to COLMAP, ACMMP, PatchmatchNet and CasMVSNet. For time reasons, we only compare with one optimization method based on differentiable rendering. We choose Neus as it performs better than IDR on DTU and is much faster than NeuralWarp, which requires an expensive two stage

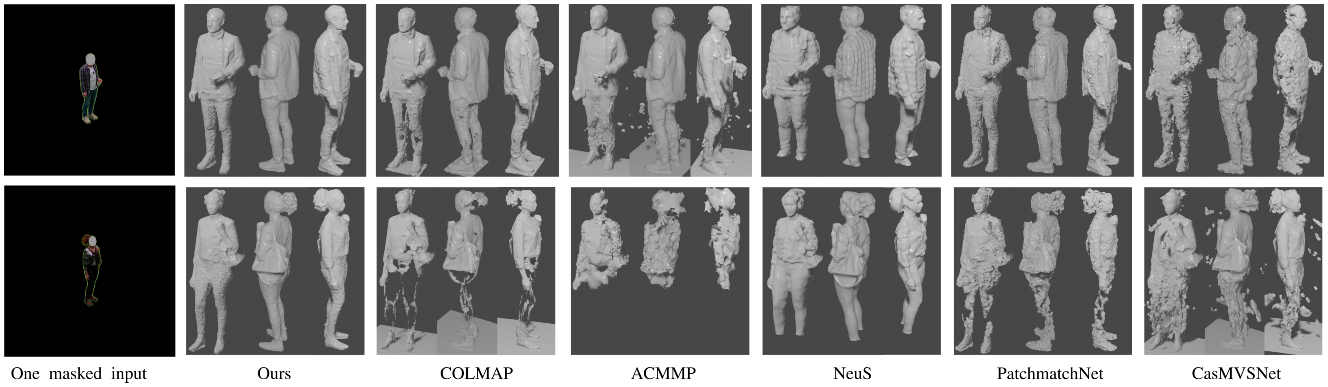

Figure 10. Left: One input image from a multi-camera platform. Right: Qualitative comparison using 65 images.

optimization. Note that PatchmatchNet, CasMVSNet and our learned photo-consistency prior are all trained on the same training set of DTU.

Qualitatively, COLMAP performs well with the $1st$ model (top row), despite some holes in the legs. However, it has difficulties with the dark pants and the hair on the $2nd$ model (bottom row). For ACMMP, a single iteration was used for optimization due to RAM limitation, even with 64GB. Its results are less precise. NeuS reconstructs a watertight surface but lacks high-frequency details (*e.g.* faces on both rows) and exhibits poor geometries at different locations due to appearance ambiguities. The deep MVS methods PatchmatchNet and CasMVSNet partially succeed with the top example but fail with the bottom one. This illustrates that end-to-end learning based methods face generalization issues when the inference scenario is substantially different from the training one, *i.e.* DTU. Our method shows detailed surfaces with limited noise, even on some difficult parts, *e.g.* the dark pants in the bottom row. This demonstrates the benefit of a weaker prior with local photoconsistency, embedded in a global optimization framework.

### 5.6. Finetuning inference-based results

Deep MVS methods like PatchmatchNet and CasMVSNet have the advantage of very fast inference but, as shown in Sec. 5.5, tend to generalize poorly. From this observation, we experiment in this section the combination of an inference-based method with our optimization-based method. As shown in Fig. 11, the result of PatchmatchNet can be used as initialization for our optimization instead of a coarse visual hull. Finetuning the results from PatchmatchNet leads to results with more details and less noise, but nevertheless fails to recover from large errors, *e.g.* top of the head, top of the back, and left hip.

## 6. Conclusion

We have presented a strategy that combines depth optimization, as performed in the latest MVS strategies, with volumetric representations successfully used in more recent methods based on differentiable rendering. Building on signed distances, our SRDF representation allows to optimize multi-view depthmaps consistently by correlating depth prediction with photometric observations along viewing rays. Experiments on real and synthetic data demonstrate the efficiency of our method compared to classic MVS, deep MVS and differentiable rendering based methods. We also demonstrate the good applicability of our method with a learned photo-consistency prior that generalizes well on data very different from the training set. As future work, photo-consistency priors can be explored to improve generalization, with *e.g.* data augmentation, or efficiency using specific architectures for very fast inference.

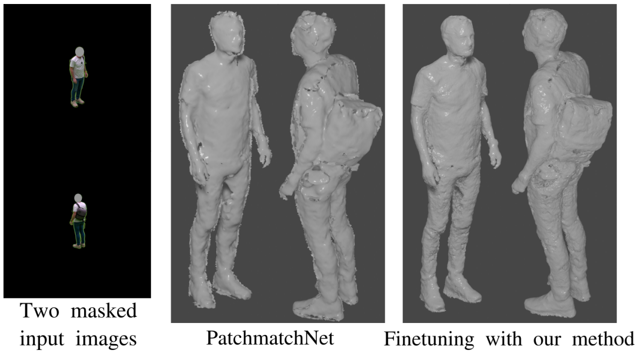

Figure 11. Finetuning the result of PatchmatchNet.

## 7. Acknowledgements

This work was supported by a French government funding managed by the National Research Agency under the Investments for the Future program (PIA) with the grant ANR-21-ESRE-0030 (CONTINUUM project). We also thank Laurence Boissieux and Julien Pansiot from the Kinovis platform at Inria Grenoble, our volunteer subjects for help with the 3D data acquisition and Jean-Sébastien Franco and Mathieu Marsot for helpful discussions.

## References

[1] Dtueval-python. https://github.com/jzhangbs/DTUeval-python. 6

[2] Kinovis platform. https://kinovis.inria.fr/inria-platform/. 7

[3] Renderpeople, 2018. https://renderpeople.com/3d-people/. 2, 6, 7

[4] Motilal Agrawal and Larry S Davis. A probabilistic framework for surface reconstruction from multiple images. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–II. IEEE, 2001. 2

[5] Relja Arandjelović and Andrew Zisserman. Nerf in detail: Learning to sample for view synthesis. *arXiv preprint arXiv:2106.05264*, 2021. 3

[6] Adrian Broadhurst, Tom W Drummond, and Roberto Cipolla. A probabilistic framework for space carving. In *IEEE International Conference on Computer Vision*, volume 1, pages 388–393. IEEE, 2001. 2

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

[8] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Annual Conference on Computer Graphics and Interactive Techniques*, pages 303–312, 1996. 2, 3, 6

[9] 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. 3, 6, 7

[10] Jeremy S De Bonet and Paul Viola. Poxels: Probabilistic voxelized volume reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 418–425, 1999. 2

[11] Frank Dellaert and Lin Yen-Chen. Neural volume rendering: Nerf and beyond. *arXiv preprint arXiv:2101.05204*, 2020. 3

[12] Simon Donne and Andreas Geiger. Learning non-volumetric depth fusion using successive reprojections. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7634–7643, 2019. 2

[13] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2009. 2

[14] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3d shape induction from 2d views of multiple objects. In *International Conference on 3D Vision*, pages 402–411. IEEE, 2017. 2

[15] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Gipuma: Massively parallel multi-view stereo reconstruction. *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e. V*, 25(361-369):2, 2016. 2, 3

[16] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *IEEE/CVF International Conference on Computer Vision*, pages 7154–7164, 2019. 2

[17] 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, 6, 7

[18] Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, and Konrad Schindler. Learned multi-patch similarity. In *IEEE International Conference on Computer Vision*, pages 1586–1594, 2017. 2, 4

[19] Paul Henderson and Vittorio Ferrari. Learning to generate and reconstruct 3d meshes with only 2d supervision. *arXiv preprint arXiv:1807.09259*, 2018. 2

[20] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018. 2

[21] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[22] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. 3

[23] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014. 2, 3, 5, 6, 7

[24] Li Jiang, Shaoshuai Shi, Xiaojuan Qi, and Jiaya Jia. Gal: Geometric adversarial loss for single-view 3d-object reconstruction. In *European Conference on Computer Vision*, pages 802–816, 2018. 2

[25] Danilo Jimenez Rezende, SM Eslami, Shakir Mohamed, Peter Battaglia, Max Jaderberg, and Nicolas Heess. Unsupervised learning of 3d structure from images. *Advances in Neural Information Processing Systems*, 29, 2016. 2

[26] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2018. 2

[27] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Eurographics Symposium on Geometry Processing*, volume 7, 2006. 2

[28] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013. 2, 5

[29] Petr Kellnhofer, Lars C Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4287–4297, 2021. 3

[30] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3):199–218, 2000. 2

[31] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In *European Conference on Computer Vision*, pages 781–796, 2018. 2, 4

[32] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis. *arXiv preprint arXiv:2103.02597*, 2021. 3

[33] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019. 2

[34] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020. 3

[35] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *SIGGRAPH*, pages 163–169, 1987. 6

[36] Paul Merrell, Amir Akbarzadeh, Liang Wang, Philippos Mordohai, Jan-Michael Frahm, Ruigang Yang, David Nistér, and Marc Pollefeys. Real-time visibility-based fusion of depth maps. In *IEEE International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[37] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 2

[38] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *IEEE/CVF International Conference on Computer Vision*, pages 4743–4752, 2019. 2

[39] 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*, pages 405–421. Springer, 2020. 3

[40] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 3

[41] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *European Conference on Computer Vision*, pages 414–431. Springer, 2020. 2

[42] KL Navaneet, Priyanka Mandikal, Mayank Agarwal, and R Venkatesh Babu. Capnet: Continuous approximation projection for 3d point cloud reconstruction using 2d supervision. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 8819–8826, 2019. 2

[43] Thu H Nguyen-Phuoc, Chuan Li, Stephen Balaban, and Yongliang Yang. Rendernet: A deep convolutional network for differentiable rendering from 3d shapes. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[44] Michael Niemeyer, Lars M. Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. *CoRR*, abs/1912.07372, 2019. 3

[45] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2019. 2

[46] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 3

[47] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 2

[48] N Passalis, S Pedrazzi, R Babuska, W Burgard, D Dias, F Ferro, M Gabbouj, O Green, A Iosifidis, E Kayacan, et al. Opendr: An open toolkit for enabling high performance, low footprint deep learning for robotics. *arXiv preprint arXiv:2203.00403*, 2022. 2

[49] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[50] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020. 2

[51] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 3

[52] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021. 3

[53] Alexander Rich, Noah Stier, Pradeep Sen, and Tobias Höllerer. 3dvnet: Multi-view depth prediction and volumetric refinement. In *International Conference on 3D Vision*, pages 700–709. IEEE, 2021. 2

[54] Gernot Riegler, Ali Osman Ulusoy, Horst Bischof, and Andreas Geiger. Octnetfusion: Learning depth fusion from data. In *International Conference on 3D Vision*, pages 57–66. IEEE, 2017. 2

[55] 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 *IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2

[56] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2

[57] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for

unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 2, 5, 6, 7

[58] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *IEEE Conference on Computer Vision and Pattern Recognition*, 1:519–528, 2006. 1

[59] Steven M Seitz and Charles R Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):151–173, 1999. 2

[60] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[61] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 2

[62] 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021. 2

[63] Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2626–2634, 2017. 2

[64] 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. 2, 6, 7

[65] 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. 3, 6, 7

[66] Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5483–5492, 2019. 2, 6, 7

[67] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[68] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 2, 5

[69] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5525–5534, 2019. 2

[70] 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. 1, 2, 6, 7

[71] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. volume 34, 2021. 3

[72] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems*, volume 33, 2020. 3, 6, 7

[73] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021. 3

[74] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 3

[75] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 3

[76] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4353–4361, 2015. 2

[77] Jun-Yan Zhu, Zhoutong Zhang, Chengkai Zhang, Jiajun Wu, Antonio Torralba, Josh Tenenbaum, and Bill Freeman. Visual object networks: Image generation with disentangled 3d representations. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[78] Ehsan Zobeidi and Nikolay Atanasov. A deep signed directional distance function for object shape representation. *arXiv preprint arXiv:2107.11024*, 2021. 3