# MVPSNet: Fast Generalizable Multi-view Photometric Stereo

Dongxu Zhao[1] Daniel Lichy[2] Pierre-Nicolas Perrin[1] Jan-Michael Frahm[1] Soumyadip Sengupta[1]

[1]University of North Carolina at Chapel Hill [2]University of Maryland, College Park

{dongxuz1, pnperrin, jmf, ronisen}@cs.unc.edu dlichy@umd.edu

## Abstract

*We propose a fast and generalizable solution to Multi-view Photometric Stereo (MVPS), called MVPSNet. The key to our approach is a feature extraction network that effectively combines images from the same view captured under multiple lighting conditions to extract geometric features from shading cues for stereo matching. We demonstrate these features, termed 'Light Aggregated Feature Maps' (LAFM), are effective for feature matching even in textureless regions, where traditional multi-view stereo methods often fail. Our method produces similar reconstruction results to PS-NeRF, a state-of-the-art MVPS method that optimizes a neural network per-scene, while being 411× faster (105 seconds vs. 12 hours) in inference. Additionally, we introduce a new synthetic dataset for MVPS, sMVPS, which is shown to be effective for training a generalizable MVPS method.*

## 1. Introduction

3D reconstruction of an object can be achieved either through camera viewpoint variations, Multi-view Stereo (MVS), or by lighting direction variations, Photometric Stereo (PS). Both MVS and PS have relative strengths and weaknesses. While MVS succeeds in obtaining accurate global shapes, it suffers in textureless regions due to poor feature matching, often resulting in reconstructions that lack local details. On the other hand, PS produces accurate local details, even in textureless regions, by using shading information but fails to reconstruct accurate global shapes. In this paper, we focus on the problem of Multi-view Photometric Stereo (MVPS) where both camera viewpoint and lighting direction variations are used to accurately reconstruct global and local details of a 3D shape, even in textureless regions.

3D reconstruction techniques that produce high-quality results using only viewpoint variations (MVS) rely on test-time optimization, often by training neural networks per scene [53, 64, 66]. These methods are computationally inefficient, typically taking hours of computing time on a high-end GPU for each object. Existing MVS methods [20, 56, 62] that focus on computational efficiency employ feed-forward neural networks that are efficient but fail to produce high-quality details, especially in textureless regions. Existing MVPS approaches can produce high-quality reconstructions but require computationally inefficient per-scene training or optimization [27, 28, 29, 61]. Sometimes additional manual efforts and carefully crafted refinement steps are also needed [34, 47]. In contrast, we propose an efficient feed-forward neural architecture, MVPSNet, that can generalize to unseen objects and achieve similar reconstruction quality to that of per-scene optimization techniques while being computationally efficient during inference.

We design MVPSNet by taking inspiration from various deep MVS architectures [7, 14, 18, 20, 62] that are generalizable, computationally efficient, and can operate on high-resolution images. However, these approaches often fail in textureless regions, and their reconstructed meshes often lack details. We choose the CasMVSNet [20] architecture as our feature matching module, which has been repeatedly used by various MVS pipelines [7, 14, 18] for its simplicity and efficiency, and augment it to effectively incorporate lighting variation cues for better prediction of 3D shapes. In this way, we propose a feed-forward generalizable approach to Multi-view Photometric Stereo.

We introduce a multi-scale feature representation, called Light Aggregated Feature Maps (LAFM), whose role is to extract detailed geometric features from images by utilizing lighting variations. For brevity, we define **Multi-light Images** as a collection of images taken from the same viewpoint under different directional lighting conditions. Our intuition is that LAFM can efficiently aggregate shading patterns from multi-light images, by creating an 'artificial shading texture' in the textureless region. Multi-scale LAFM will then be used to construct a sequence of cost volumes to match features across sparse views in order to

predict a depth map for a reference view. We also predict surface normals from LAFM for each viewpoint, enabling LAFM to capture features related to high-frequency local details. The predicted surface normal can be used in addition to the depth maps to produce a more detailed mesh than using the depth maps alone.

To train the proposed MVPSNet architecture, we introduce a new synthetic MVPS dataset, sMVPS dataset. It consists of shapes from sculpture dataset [57] and random compositions of primitive shapes generated by [60]. We render these shapes with spatially varying Cook-Torrance BRDF under different camera viewpoints and lighting directions. We train MVPSNet on these rendered images with ground-truth supervision over predicted depth and surface normal maps. The trained model generalizes to real-world test scenes from DiLiGenT-MV [34] dataset. We show that simply re-training CasMVSNet on our dataset improves reconstruction quality over the pre-trained model on DiLiGenT-MV by 32%, proving the effectiveness of our synthetic MVPS dataset for generalization.

We evaluate our approach on the only publicly available MVPS benchmark, the DiLiGenT-MV [34] dataset. Compared to the state-of-the-art MVPS technique, PS-NeRF [61], which optimizes a neural network per-scene, our proposed MVPSNet is $\sim$411$\times$ faster (105 seconds vs 12 hours) while producing similar reconstruction quality (L1 Chamfer distance of 0.82 vs 0.81, F-score on L2 distance of 0.985 vs 0.983). We further show that adding LAFM significantly improves reconstruction quality over CasMVSNet by 34% in L1 Chamfer distance. We also observe that refining the reconstructed mesh derived from depth maps with predicted surface normals from LAFM improves reconstruction quality as shown in Fig 3.

In summary, the key contributions of this paper include: • Light Aggregated Feature Maps (LAFM) that can efficiently utilize multi-light images to extract detailed geometric features, especially in textureless regions. The surface normal predicted from LAFM also improves mesh reconstruction quality. • A synthetic MVPS dataset for training generalizable MVPS methods, which also improves CasMVSNet by 32%. • A fast and generalizable Multi-view Photometric Stereo pipeline that is 411$\times$ faster while producing similar reconstruction accuracy compared to state-of-the-art per-scene optimization approach [61].

## 2. Related work

**Multi-view Stereo (MVS).** MVS is a 3D reconstruction technique that utilizes multiple images captured from different viewpoints. While various techniques for MVS have been proposed, one commonly used approach that is relevant to our work involves constructing cost volumes similar to Plane Sweeping Algorithm [13] and then predicting per-view depth maps [21, 25, 41, 59, 62, 63] or disparity maps [23]. To create a cost volume, features are matched across neighboring viewpoints, and the quality of the features plays a critical role in the final reconstruction quality. Traditional methods [5, 16, 17, 19, 26, 50, 54, 55] use human-defined or hand-crafted image processing operators to extract feature maps. With recent advances in deep learning, features learned from deep neural networks have been proven to be effective.

| Method | Generalizable | Mesh Reconstruction |
|---|---|---|
| PJ16 [47] | ✗ | Base mesh+displacement map |
| LZ20 [34] | ✗ | 3D points+PSR [30]+Optimization [44] |
| BKW22 [29] | ✗ | MLP+Marching Cube [40] |
| BKC22 [27] | ✗ | MLP+Marching Cube [40] |
| PS-NeRF [61] | ✗ | MLP+MISE [42] |
| BKW23 [28] | ✗ | MLP+Marching Cube [40] |
| Ours | ✓ | 3D Points+Screened Poisson [31] |

Table 1. Comparison of our method with prior MVPS methods.

The most relevant previous works are MVSNet [62] and its variations. MVSNet [62] uses homography to warp feature maps and a 3D CNN to regularize cost volumes. CasMVSNet [21] outperforms MVSNet in terms of accuracy and efficiency by building the 3D cost volume in a cascaded manner. TransMVSNet [15] builds upon CasMVSNet and adopts a transformer to consider intra-image and inter-image feature interactions, which further improves the results of CasMVSNet.

**Photometric Stereo (PS).** PS (introduced in [58]) uses lighting variation to reconstruct 3D shapes from a single viewpoint (see [52] for surveys). Calibrated PS approaches, like Chen *et al*. [10], train a neural network to predict surface normals using data with known lightings. Uncalibrated PS approaches [8, 9, 11] first predict the lighting parameters before solving for surface normals. While most PS works use a large number of images for inference, some use fewer [24,38], or even one image [6,35,36,51] (often called Shape from Shading). PS approaches are mostly based on feed-forward networks that generalize and can produce near real-time inference with low computational cost [37].

**Multi-view Photometric Stereo (MVPS).** MVPS was initially proposed in [22] by combining PS with object silhouettes to reconstruct textureless shiny objects with fine details. However, this method only works well for specific parametric BRDF models [27]. Later, Li *et al*. [34, 68] propose to get iso-depth contours from PS images and sparse 3D points using structure-from-motion, which are propagated to recover a complete 3D shape. Park *et al*. [46, 47] use a planar mesh parameterization technique to parameterize a coarse mesh from MVS and take advantage of this 2D parameter domain to perform MVPS. Some of these traditional MVPS methods achieve high-quality results, but they require an initial 3D reconstruction and their performance is sensitive to it. Besides, they consist of multiple steps, so careful execution or expert interventions are often needed to

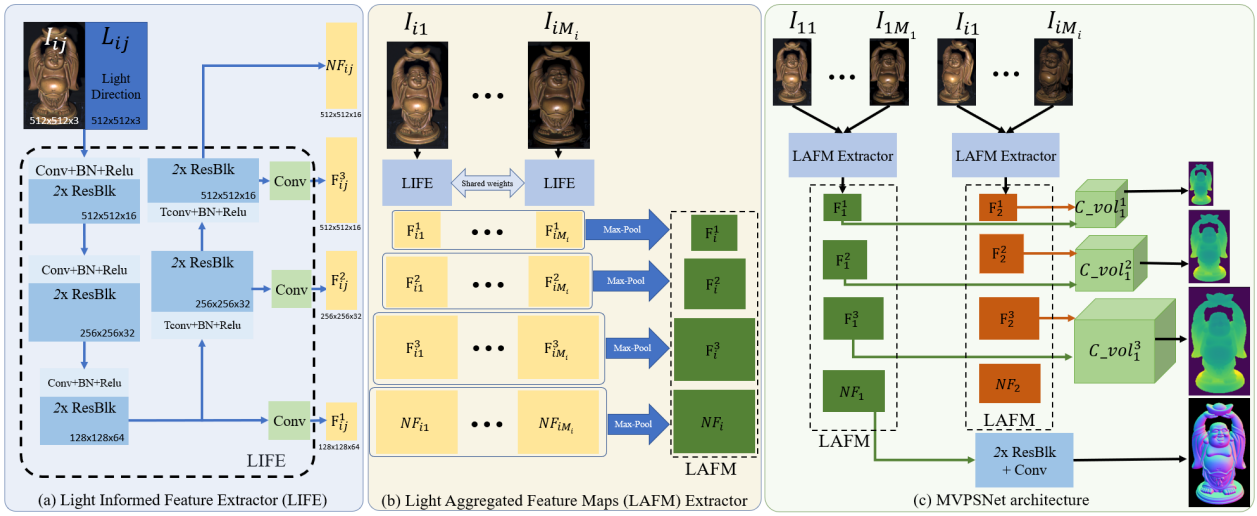


Figure 1. Overview of our network architecture. (a) Light Informed Feature Extractor (LIFE) produces a multi-scale feature representation; (b) Light Aggregated Feature Maps (LAFM) Extractor aggregates these features across images of varying lighting conditions but same view; and (c) LAFM is used to create cost volumes and predict depth maps, similar to CasMVSNet [21], in addition to normal map.

get good results [34, 47]. In contrast, our proposed method trains a neural network end-to-end and the inference only requires one feed-forward pass of the trained network.

Recently, inspired by NeRF [43], various algorithms have been proposed that optimize a neural network per-scene for MVPS. Kaya *et al*. [29] train a deep PS network first and condition the color rendering in NeRF [43] on normals predicted from PS. The reconstructed mesh, however, exhibits multiple artifacts. The authors in [27] propose to train a deep PS network and a deep MVS network extended with uncertainty estimation separately, and use these to fit the SDF represented by an MLP. To further enable reconstruction on anisotropic and glossy objects, Kaya *et al*. [28] add a neural volume rendering module to the SDF MLP in [27] to better fuse PS and MVS measurements. PS-NeRF [61] solves the task of jointly estimating the geometry, materials and lights. It first regularizes the gradient of a UNISURF [45] with estimated normals from PS, and then uses separate MLPs to explicitly model surface normal, BRDF, lights, and visibility which are optimized based on a shadow-aware differentiable rendering layer. Recent works have also used physically based differentiable rendering either inside a NeRF framework [3] or separately for optimization [65]. While per-scene optimization methods often generate precise reconstruction results, they have to optimize an individual model for each object separately, which is computationally inefficient. Thus, we propose a solution to Multi-view Photometric Stereo training on a new synthetic MVPS dataset, which is generalizable and can achieve similar results as SOTA per-scene optimization methods.

# 3. Our approach

## 3.1. Problem setup

We focus on the problem of calibrated multi-view photometric stereo, i.e. the locations of the light sources and the cameras are known a priori (calibrated prior to capture). The input data consists of a set of multi-light images of an object captured from multiple views.

Concretely, for the $i$-th view we have $M_i$ images with varying lighting directions $l_{ij}$, denoted as $I_{ij}$. We refer to the collection $\{I_{i1}, ..., I_{iM_i}\}$ as the multi-light images for the i-$th$ view. For each view, we are given camera intrinsic matrix $K_i$ and camera extrinsic parameters in the form of a rotation matrix, $R_i$, and a translation vector, $t_i$. Similar to virtually all MVS methods, we assume that we are provided with the depth range for each view.

## 3.2. Motivation

Our approach follows a long line of work in Multi-view Stereo which uses Plane Sweep to construct a cost volume and predicts a depth map aligned with a reference image. Recent advances in Plane Sweep Stereo using deep neural networks, especially CasMVSNet [21], have proven to be generalizable across scenes and can predict high-resolution reconstruction in a matter of seconds. In contrast to previous Multi-view Photometric Stereo approaches, which optimize a neural network per scene, our goal is to produce a generalizable solution. Thus we aim to build upon the Plane Sweep stereo architecture proposed in CasMVSNet [21].

CasMVSNet learns a deep image feature encoder for extracting representative features that can aid in feature matching across multiple views and create better cost vol-

umes. However, these features are often ambiguous for non-textured regions and fail to preserve the geometric details. We believe incorporating lighting variations along with viewpoint variations can lead to better features, which in turn will produce better cost volumes and depth maps.

To this end, we introduce 'Light Aggregated Feature Maps' (LAFM), whose goal is to extract detailed geometric features, even for textureless regions, by jointly learning to aggregate feature maps over all images captured from a single viewpoint and multiple lighting directions. To obtain effective features that capture geometric details we use LAFM to regress surface normals. We show that LAFM provide superior information for stereo-matching than single-lighting feature maps (as used in CasMVSNet) and thus provide a better reconstruction. We also show that surface normals predicted using LAFM can be used during mesh reconstruction to improve quality over depth maps alone.

Our approach proceeds in three stages. We first extract multi-scale feature representation from each image, along with its lighting directions, using a shared neural network, 'Light Informed Feature Extractor' (LIFE). To aggregate features extracted by LIFE across all images under the same viewpoint but different lighting conditions, we use max-pooling operation. We can then create a cost volume for the reference view by matching LAFM of the reference view with all the LAFM from neighboring views. Finally, for each reference view, we predict a depth map using cost volume regularization and a surface normal map from the LAFM. We train our system in a multi-task learning framework with supervised losses over depth and normal predictions. In the following sections, we provide the details of our MVPSNet pipeline. An overview of MVPSNet architecture is illustrated in Figure 1.

### 3.3. Light Aggregated Feature Maps (LAFM)

We introduce Light Aggregated Feature Maps (LAFM) that provide geometrically distinct multi-scale features for cost volume creation in a Plane Sweep Stereo approach. Our key observation is that the multi-light images provide us with important information for feature matching. For textureless regions, the variation in shading (including cast and attached shadows) created by different lighting directions can be interpreted as 'artificial' textures. Thus the role of LAFM for textureless regions is to capture the variation in shading as an 'artificial' texture that can be used for feature matching across different viewpoints. We also use LAFM to predict surface normal maps, enabling it to capture geometric details required for producing high-quality normal maps. Hence LAFM can capture better features for textureless regions and for reconstructing details, which were absent in the usual deep image features used in deep multi-view stereo algorithms.

We first define a multi-scale feature extractor, Light Informed Feature Extractor (LIFE), that takes an image $I_{ij}$ associated with its lighting direction $l_{ij}$ as input and produces feature maps at three different scales $F^1_{ij}, F^2_{ij}, F^3_{ij}$ at resolutions $1/4, 1/2, 1$ of the input resolution, and another feature map $NF_{ij}$ that will be used for normal prediction. The network architecture of LIFE is shown in Fig. 1(a) and will be discussed in details in the supplementary material.

$$NF_{ij}, F^1_{ij}, F^2_{ij}, F^3_{ij} = LIFE(I_{ij}, l_{ij}; \theta) \quad (1)$$

Note $l_{ij}$ is of the same resolution as $I_{ij}$ by simply repeating the same 3-dimensional lighting vector at each pixel.

Then we extract these multi-scale features for every image captured under the same viewpoint and different lighting conditions, $\{I_{i1}, ..., I_{iM_i}\}$, using the same shared encoder LIFE. Let the feature maps obtained from these images be denoted as: $\{NF_{ij}, F^1_{ij}, F^2_{ij}, F^3_{ij}\}$, $j = 1, ..., M_i$. We create 'Light Aggregated Feature Maps' (LAFM) from these multi-scale feature representations by performing a max-pooling operation for each scale. Pooling operations can handle various number of input feature maps and are order-agnostic. Moreover, max-pooling helps save the most prominent feature across all views and ignores non-activated features, which automatically handles shadows casted by directional lights [10].

$$F^s_i = \max_j F^s_{ij}, \quad \forall s = 1, 2, 3 \quad (2)$$

$$NF_i = \max_j NF_{ij}. \quad (3)$$

Thus for multi-light images we obtain LAFM as $LF_i = \{NF_i, F^1_i, F^2_i, F^3_i\}$.

The features at 3 scales $F^1_i, F^2_i, F^3_i$ are then used to build cost volumes using differentiable homography warping, which we will talk about in detail in Section 3.4. The normal feature $NF_i$ is fed into a lightweight normal regression network to predict per-view normal map, as shown in Figure 1(c). With the supervision from normal information and depth information, our LAFM benefit from the advantages of both MVS and PS which are good at global shape modeling and high-frequency component reconstruction, respectively.

### 3.4. Cost volume and depth map prediction

Given Light Aggregated Feature Maps (LAFM), $LF_i$, for each view $i$, we aim to build a cost volume for each reference view by selecting a set of source views with sufficient overlap. We adopt the multi-scale cost volume construction proposed in CasMVSNet [21], where the plane sweep is first performed at a low resolution and then at higher resolutions. Depth estimated from the previous step is used for generating depth proposals for the next step. Multi-scale cost volume reconstruction and depth map prediction follow the following steps.

**Step 1: Depth hypothesis generation.** We generate hypothesis depths for each pixel based on the lower resolution

depth estimated at the previous resolution. We store these in $h$, where

$$h(u,v,w) = Up(D^{s-1})(u,v) + \Delta_s(\frac{w}{N_s - 1} - \frac{1}{2}). \quad (4)$$

Here $h(u,v,w)$ is the $w$-th depth hypothesis at pixel $(u,v)$. $Up(D^{s-1})$ is the depth map at the previous lower resolution upsampled to the current resolution. $\Delta_s$ is the length of the depth interval we are searching at scale $s$. $N_s$ is the number of hypothesis depths at the current scale.

**Step 2: Cost volume construction.** Building cost volume is a way of robustly searching for matches between a point $(u,v)$ in the reference image $I_r$ and a point on the corresponding epipolar line in the source image $I_{s_k}$. Concretely, consider a pixel $(u,v)$ in the reference image. For every hypothesis depth $d$, we get a corresponding point in the source image on the epipolar line for $(u,v)$. We denote this point by $(u',v') = \text{warp}_{rs_k}(u,v,d)$ where

$$\begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} \sim K_{s_k} R_{s_k}^T \left[ \left( R_r K_r^{-1} d \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} + t_r \right) - t_s \right]. \quad (5)$$

We then construct a per-image volume:

$$\text{F\_vol}_i^s(u,v,w) = F_i^s(warp(u,v,h(u,v,w))), \quad (6)$$

where $i$ runs over the reference and source views i.e. $i \in \{r, s_1, ..., s_k\}$. These volumes are then aggregated into a single cost volume by taking their variance, which checks for the photo-consistency of the depth proposal $d$ for pixel $(u,v)$ in the reference image and the corresponding pixels in the warped sources images $s_k$:

$$\text{agg\_vol}^s(u,v,w) = \text{var}_i(\text{F\_vol(u,v,w)}_i^s). \quad (7)$$

**Step 3: Cost regularization.** In this step we pass the aggregated volume through a 3D convolutional network and take a softmax to convert it to a match probability, using

$$\text{prob\_vol} = \text{soft\_max}_w(\text{reg\_net}^s(\text{agg\_vol}^s)). \quad (8)$$

**Step 4: Regression.** We take the expectation of the hypothesis depths over the match probability given by the probability volume to obtain the depth at the current scale, which enables sub-pixel estimation.

$$D^s(u,v) = \sum_w \text{prob\_vol}(u,v,w) h(u,v,w). \quad (9)$$

This whole process is summarized in algorithm 1

**Algorithm 1** MVPSNet Algorithm

$NF_{ij}, F_{ij}^1, F_{ij}^2, F_{ij}^3 = \text{LIFE}(I_{ij}, l_{ij}; \theta)$
$NF_i = \max_j NF_{ij}$; $F_i^s = \max_j F_{ij}^s \quad \forall s = 1,2,3.$
$N_i = \text{normal\_regression\_net}(NF_i)$
$D^0(u,v) = (\text{max\_depth} + \text{min\_depth})/2$
**for** s = 1 to 3 **do**
  $h(u,v,w) = Up(D^{s-1})(u,v) + \Delta_s(\frac{w}{N_s - 1} - \frac{1}{2})$
  $\text{F\_vol}_i^s(u,v,w) = F_i^s(warp(u,v,h(u,v,w)))$
  $\text{agg\_vol}^s(u,v,w) = \text{var}_i(\text{F\_vol}_i^s)$
  $\text{prob\_vol} = \text{soft\_max}_w(\text{reg\_net}^s(\text{agg\_vol}^s))$
  $D^s(u,v) = \sum_w \text{prob\_vol}(u,v,w) h(u,v,w)$
**end for**

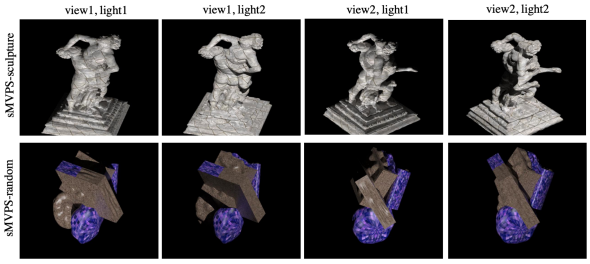


Figure 2. Example images from proposed synthetic MVPS dataset.

Once we have depth and surface normal for each view, our mesh reconstruction pipeline consists of three steps: depth filtering, lifting depth and normal maps to a point cloud, and reconstructing the mesh using Screened Poisson [31] (See supplementary materials for details).

### 3.5. Synthetic MVPS dataset

A key component of our method is that we can learn better features for stereo matching, especially in textureless regions, by learning features that incorporate multi-lighting cues. However, there is no existing MVPS dataset that is large enough for neural network training. Therefore, we generate a large-scale synthetic dataset, sMVPS, consisting of two sub-datasets, sMVPS-sculpture (800 train scenes/4 test scenes) and sMVPS-random (1000 train scenes/20 test scenes).

sMVPS-sculpture consists of objects from the sculpture dataset [57] while sMVPS-random includes objects composed of random primitives from [60]. The objects were generated following the method of [39] with spatially varying Cook-Torrance BRDF. We render images from 20 viewpoints surrounding the object approximately every 18° plus random jitter in position. For each view we render 10 randomly chosen directional light sources sampled uniformly on a 45° spherical cap centered at the camera's optical axis. In addition to the images, we render ground truth normals, depth, albedo, and roughness.

### 3.6. Training MVPSNet

We train MVPSNet with supervised loss over surface normal and depth using the ground-truth created within synthetic sMVPS dataset. For each reference view, we use 2 source views. And we randomly choose 3 lights out of 10 to train our model. The total loss is defined as:

$$L_{mvps} = \lambda_d \cdot L_d + \lambda_n \cdot L_n, \quad (10)$$

$$L_d = \sum_{s=1}^{3} \lambda_{ds} \cdot L_{ds}, \forall s = 1,2,3 \quad (11)$$

where $L_{ds}$ and $L_n$ refer to the depth loss for scale $s$ and normal loss, respectively. For loss weights, we set $\lambda_n = 1$ and $\lambda_d = 10$. The weights of each scale, $\lambda_{ds} = 1$, for $s = 1,2,3$.

| | Per-scene optimization | | | | | Generalizable | | | |
|---|---|---|---|---|---|---|---|---|---|
| Category | Manual Effort | | Standalone | | | Single-view PS | MVS | | MVPS |
| Method | PJ16 [47] | LZ20 [34] | BKW22 [29] | BKC22 [27] | PS-NeRF [61] | PS-Transformer [24] | CasMVSNet [21]- RT | TransMVSNet [15]- RT | Ours |
| BEAR | 2.54 | 0.73 | 1.01 | 1.01 | **0.76** | 3.17 | 1.47 | 1.48 | 0.80 |
| BUDDHA | 1.12 | 0.97 | 2.68 | 1.15 | **0.86** | 4.09 | 1.26 | 1.10 | 1.07 |
| COW | 1.14 | 0.39 | 1.09 | 0.76 | **0.75** | 3.04 | 1.27 | 1.05 | 0.77 |
| POT2 | 3.21 | 0.67 | 1.54 | 1.40 | **0.76** | 3.05 | 1.46 | 1.05 | 0.82 |
| READING | 1.30 | 0.66 | 1.97 | 0.84 | 0.92 | 3.60 | 0.75 | 0.76 | **0.66** |
| AVERAGE | 1.86 | 0.69 | 1.66 | 1.03 | **0.81** | 3.39 | 1.24 | 1.09 | 0.82 |
| Recon. Time/object | - | - | 7 hrs | ? | 12 hrs | ? | **22s** | 52s | 105s |

Table 2. L1 Chamfer Distance in mm (lower is better) between reconstructed mesh and GT after ICP. '-RT' denotes trained on our synthetic MVPS dataset. For non-manual methods, the best result is shown in bold, 2nd best as underline. LZ20 & PJ16 involve carefully crafted steps, manual efforts in finding correspondence, and an initial mesh or point cloud.

| | Per-scene optimization | | | | | | Generalizable | | |
|---|---|---|---|---|---|---|---|---|---|
| Category | Manual Effort | | Standalone | | | | Single-view PS | MVS | | MVPS |
| Method | PJ16 [47] | LZ20 [34] | BKW22 [29] | BKC22 [27] | BKW23* [28] | PS-NeRF [61] | PS-Transformer [24] | CasMVSNet [21]-RT | TransMVSNet [15]-RT | Ours |
| BEAR | 0.551 | 0.986 | 0.928 | 0.934 | 0.965 | **0.995** | 0.078 | 0.911 | 0.882 | 0.991 |
| BUDDHA | 0.940 | 0.936 | 0.687 | 0.926 | **0.993** | 0.983 | 0.066 | 0.919 | 0.963 | 0.958 |
| COW | 0.918 | 0.990 | 0.937 | 0.986 | 0.987 | 0.986 | 0.140 | 0.914 | 0.941 | **0.993** |
| POT2 | 0.484 | 0.985 | 0.909 | 0.889 | 0.991 | 0.991 | 0.101 | 0.901 | 0.964 | **0.994** |
| READING | 0.905 | 0.975 | 0.810 | 0.971 | 0.975 | 0.961 | 0.961 | 0.980 | 0.978 | **0.988** |
| AVERAGE | 0.760 | 0.974 | 0.854 | 0.941 | 0.982 | 0.983 | 0.269 | 0.925 | 0.946 | **0.985** |
| Recon. Time/object | - | - | 7 hrs | ? | ? | 12 hrs | ? | **22s** | 52s | 105s |

Table 3. F-score with L2 distance and 1mm threshold (higher is better) between reconstructed mesh and GT after ICP. '-RT' denotes trained on our synthetic MVPS dataset. For non-manual methods, the best result is shown in bold, 2nd best as underline. LZ20 & PJ16 involve carefully crafted steps, manual efforts in finding correspondence, and an initial mesh or point cloud. BKW23* code not available, results from the paper.

# 4. Experiments

**Dataset.** We evaluate our method and conduct ablation study on DiLiGenT-MV [34] dataset, which is the only benchmark dataset for MVPS tasks and widely used by all previous approaches. It contains images of 5 objects with diverse materials captured from 20 views. For each view, the object is illuminated by one of 96 calibrated point light sources at one time, which gives us 96 images with varying lighting conditions.

**Evaluation metrics.** We evaluate the quality of recovered meshes using L1 Chamfer distance from PyTorch3D [33, 48] and F-score [32] with L2 distance and 1mm threshold distance. The distances in both metrics are computed between the vertices of two meshes and the units are *mm*.

**Evaluation details.** Ground truth meshes and meshes of PJ16 [47] and LZ20 [34] are included in the DiLiGenT-MV dataset [34]. We thank the authors of BKW22 [29] and BKC22 [27] for providing us with their reconstructed meshes. For PS-NeRF [61] we use its publicly available mesh extraction code from stage 1 and unscale the extracted mesh to the scale of the ground truth as suggested in the code. The code or reconstructed meshes of BKW23 [28] was not released, so we only include their reported F-score on L2 distance in Table 3. To compare with PS-Transformer [24], we get normal maps from their pretrained model and integrate normals into depth maps, followed by a depth fusion step after rescaling all depth maps to the ground truth depth scales. To compare with CasMVSNet [21] and TransMVSNet [15], we consider both the pretrained models on DTU dataset [1] and models retrained on our synthetic dataset, dubbed as CasMVSNet-RT and TransMVSNet-RT. Images from 5 views are used to generate each single-view depth map and all 20 depth maps are fused together for 3D reconstruction. For ours, we take images from 5 views and 10 lightings conditions for each view, along with corresponding light directions, as input to generate single-view depth maps and all 20 depth maps are fused as other methods. Since there is no image showing the bottoms of objects in DiLiGenT-MV [34], following BKW22 [29] and BKC22 [27], we remove points that are located lower than +5 on the z-axis from all reconstructed meshes and the ground truth. Similar to most previous approaches [27, 29, 34, 47] we also perform a rigid registration using Iterative Closest Point (ICP) [2, 4, 12, 67] between the ground-truth and each reconstructed mesh for a fair comparison.

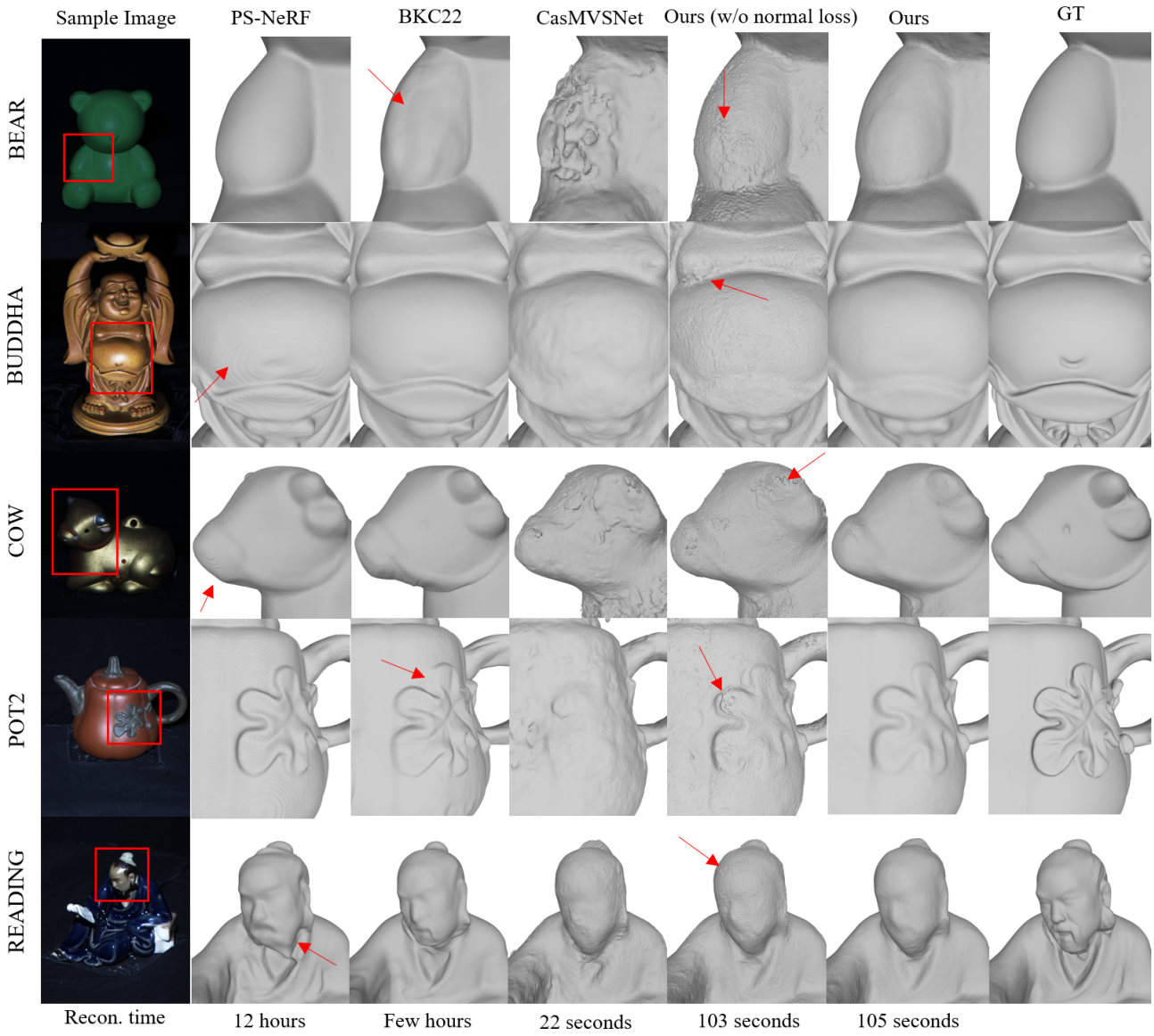


Figure 3. Qualitative comparison of our method with existing approaches (red arrows highlight artifacts) and ablation studies.

| Method | CasMVSNet | CasMVSNet-RT | Ours (train 1 light/view) | Ours (train 3 light/view) |
|---|---|---|---|---|
| BEAR | 2.00 | 1.47 | 1.31 | 0.80 |
| BUDDHA | 1.44 | 1.26 | 1.26 | 1.07 |
| COW | 2.73 | 1.27 | 1.06 | 0.77 |
| POT2 | 1.89 | 1.46 | 1.07 | 0.82 |
| READING | 1.07 | 0.75 | 0.77 | 0.66 |
| AVERAGE | 1.83 | 1.24 | 1.09 | 0.82 |

Table 4. Ablation: Results are improved by retraining CasMVSNet [20] on proposed synthetic dataset (sMVPS). LAFM help in aggregating features across lighting variations, Ours (train 3 light/view) vs Ours (train 1 light/view), and is more accurate than CasMVSNet features, Ours (train 1 light/view) vs CasMVSNet-RT.

### 4.1. Comparison with existing approaches

We compare our algorithm with approaches that require per-scene optimization or training and with feed-forward generalizable methods. The quantitative result is shown in Table 2 and 3. We also show visual comparison of meshes from representative methods in Figure 3.

(i) **Per-scene Optimization.** Per-Scene optimization methods can also be categorized into:
**(a) Manual Efforts Needed.** We compare with two traditional multi-stage MVPS methods, PJ16 [47] and LZ20 [34], which require manual efforts. We outperform PJ16 [47] with a clear margin. Although LZ20 [34] achieves better results than ours, note that both methods, PJ16 & LZ20, consist of multiple steps with carefully crafted geo-

metric modeling. Besides, they require an initial mesh [47] or point cloud [34] to build upon and their performance is sensitive to the initialization quality. When the initialization step fails in large textureless regions, LZ20 [34] incorporates manual labeling to establish correspondence across views. In contrast, our pipeline is completely automatic without any manual efforts, does not involve any carefully crafted multi-stage approach, and does not require separate hyper-parameters for individual object.

**(b) Standalone Methods.** Recent deep learning-based MVPS methods, including BKW22 [27], BKC22 [29], BKW23 [28] and PS-NeRF [61], are simpler and easier to adopt. However, like traditional methods [34, 47], they still optimize one model for each object individually, resulting in low computational efficiency. In contrast, although our model is trained only on synthetic data, we outperform some per-scene optimized methods [27, 29] and get comparable results as state-of-the-art, PS-NeRF [61]. Furthermore, even though PS-NeRF [61] recovers high-quality meshes with details, its recovered surfaces contain iso-contour pattern artifacts, *e.g*. see the red arrows in Fig. 3 on BUDDHA and COW, and sometimes incorrect shapes, *e.g*. READING. Note that, we could not report L1 Chamfer distance on BKW23 [28], since the code is unavailable, but we show in Table 3 that our method is slightly better than BKW23 on F-score.

(ii) **Generalizable.** We compare our method with two categories of generalizable methods:

**(a) Single-view Photometric Stereo.** PS-Transformer [24] is a state-of-the-art PS network, which takes multiple images with the same viewpoint but different lighting conditions as input and generates single-view normal map prediction. To get 3D reconstruction, we integrate each normal map into a depth map and fuse them together. Since the integrated depths are of arbitrary scale, we rescale them to the range of ground truth depth. Inherently, PS methods struggle with global shape modeling and it is challenging to stitch multi-view integrated depth of arbitrary scale, so PS-Transformer [24] doesn't perform well on full-view recovery.

**(b) Multi-view Stereo.** We also compare our method with CasMVSNet [21] and TransMVSNet [15], both of which use a single lighting image from each view. For fairness, we retrain both methods using our synthetic dataset with suggested hyper-parameters in original papers . We observe that lighting information can largely improve both accuracy and quality, quantitatively and qualitatively. On textureless objects, *e.g*. BEAR, and textureless regions, *e.g*. belly of BUDDHA, MVS alone gets noisy and rough surfaces. Moreover, our meshes have more high-frequency details that MVS alone may struggle with, *e.g*. the texture on POT2. This is because our LAFM are supervised with normal maps so they can learn the high-frequency components.

## 4.2. Computational efficiency

While our method outperforms some per-scene optimized methods [27, 29] and produces comparable results to state-of-the-arts [28, 61], the key advantage of our method is that it is fast, generalizable, and computationally efficient. Thus we analyze the inference time of the MVPS algorithms compared in this paper to the best of our abilities.

• LZ20 [34]: This algorithm takes *117 minutes* per object, without considering the time required for initializing a point cloud or any manual efforts. • BKW22 [29]: takes 7 hours to train per object. • BKC22 [27], BKW23 [28]: Since the authors did not mention the time required for training these algorithms it is not possible to provide an exact estimate. However, these approaches are based on MLPs, which take hours to train. • PS-NeRF [61]: takes *12 hours* to train per object. • CasMVSNet [20]:, in contrast, takes only 22 seconds per object, including obtaining depth maps for each view using 5 views (1 reference view and 4 source views) and 1 lighting per view (5 images processed for estimating a depth map), fusing depth maps from all 20 views to a point cloud, computing normals for vertices in the point cloud and adopting Screened Poisson [31] to recover a mesh from the point cloud. • MVPSNet (Ours): takes a total of *105 seconds* to create a mesh, including steps of obtaining depth maps for each view using 5 views (1 reference view and 4 source views) and 10 lightings per view (50 images processed in total), fusing depth maps from all 20 views to a point cloud, and adopting Screened Poisson [31] to recover a mesh from the point cloud.

In summary, we are around 240× faster than BKW22 [29] and around 411.4× faster than PS-NeRF [61] ignoring their mesh extraction time.

## 4.3. Ablation study

The key contribution of this work includes: (a) our synthetic sMVPS dataset, including sMVPS-sculpture and sMVPS-random, and (b) 'Light Augmented Feature Maps' (LAFM) for more accurate mesh reconstruction. Here we design experiments to analyze impacts of these contributions.

**Synthetic Data (sMVPS-sculpture and sMVPS-random).** To illustrate the effectiveness of our synthetic data, we evaluate the performance of a CasMVSNet [21] model trained on DTU dataset [1], and compare it with the CasMVSNet-RT model trained on our sMVPS dataset. The L1 Chamfer distance metric is reported in Table 4. We observe that training on our synthetic dataset improves reconstruction quality by 32.2%, proving the effectiveness of our proposed data for MVPS reconstructions. See supplementary materials for comparison between pretrained TransMVSNet [15] and TransMVSNet-RT trained on our synthetic dataset.

**Light Augmented Feature Maps (LAFM).** LAFM play

| # of lightings | 1 | 4 | 10 | 20 | 30 | 40 | 96 |
|---|---|---|---|---|---|---|---|
| 3 viewpoints | 1.237 | 0.883 | **0.855** | 0.863 | 0.865 | 0.871 | 0.892 |
| 5 viewpoints | 1.244 | 0.856 | **0.823** | 0.831 | 0.840 | 0.847 | 0.878 |

Table 5. Averaged Chamfer-L1 distance (mm) of DiLiGenT-MV. The network is trained with 3 views and 3 lightings per view.

two key roles in our approach: (i) they aggregate features from images captured with multiple lighting conditions but the same viewpoints, and (ii) they are trained with surface normal loss which helps to preserve high-frequency details in the features. The predicted surface normals are used to further refine the reconstructed mesh.

For understanding the impact of (i), we train our proposed MVPSNet with just a single lighting image per view instead of 3 lightings. Thus LAFM are only aggregated across 1 image per view. In Table 4 we observe that using a single lighting per view ('Ours (train 1 light/view)') produces worse results (1.09 vs 0.82) than using 3 lightings per view ('Ours (train 3 light/view)'). However, even using a single lighting per view produces better performance than CasMVSNet-RT (1.09 vs 1.24), which is also trained on single lighting per view. This shows that LAFM are effective in both extracting accurate information from just a single image and aggregating shading information across multiple images with varying illumination.

For understanding the impact of (ii), we train our proposed MVPSNet without any surface normal loss, 'Ours (w/o normal loss)'. We observe 'Ours (w/o normal loss)' is quantitatively comparable to 'Ours (w/normal loss)', 0.79 vs 0.82 in L1 chamfer distance and 0.985 vs 0.985 in F-score. However, in Fig. 3 we observe that the meshes produced by 'Ours (w/o normal loss)' are significantly noisier as shown with red arrows.

### 4.4. Effects on different numbers of viewpoints and lighting conditions

We also test the effects of the number of viewpoints and the number of lighting conditions used at inference time. The results of different combinations are shown in Table 5. We target at using sparse views to estimate depth maps, so we only test on 3 and 5 views. The results prove that although trained on 3 views, our model is able to utilize the extra information provided more views. For lighting conditions, the result gets better initially as images under more lighting conditions are provided, but then it gets slightly worse after around 10 lighting conditions.

### 4.5. Result on real-world captures

We include the result on real-world captures using a simple at-home setup, which requires no special equipment. As shown in Fig 4, a user captures MVPS imagery by attaching a *flashlight* with a string to the tripod to move lights in a circle around the camera. In our example, images from 5 viewpoints with 3 lighting conditions per view are captured.

We automatically calibrate camera and lighting conditions using COLMAP [49, 50] and SDPSNet [8] separately. The room isn't strictly dark and contains some ambient lighting. We generate the foreground mask with an off-the-shelf segmentation algorithm. The result shows that our method produces good reconstruction and improves over CasMVSNet [20] in both details and global shapes.

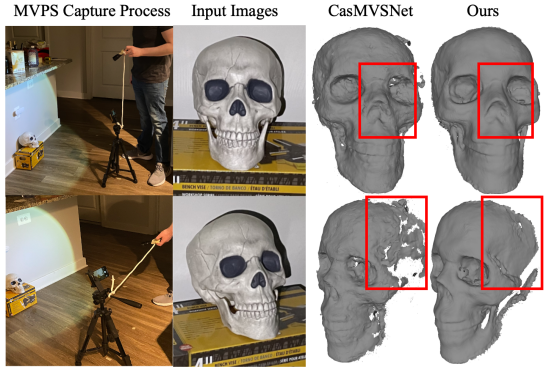


Figure 4. Results of uncalibrated in-the-wild MVPS captures.

## 5. Conclusion

In this work, we propose a fast and generalizable approach for MVPS. We introduce Light Aggregated Feature Maps that leverage shading cues from images with the same view under multiple lighting conditions to produce richer features in textureless regions. Being trained with normal estimation, LAFM also enable higher quality reconstruction than traditional MVS methods with only little compromise to speed. When trained on the synthetic sMVPS dataset we propose, our method produces results comparable to SOTA method that is about 400x slower at inference time.

## 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.

[2] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):698–700, 1987.

[3] Meghna Asthana, William AP Smith, and Patrik Huber. Neural apparent brdf fields for multiview photometric stereo. *arXiv preprint arXiv:2207.06793*, 2022.

[4] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992.

[5] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011.

[6] Mark Boss, Varun Jampani, Kihwan Kim, Hendrik P.A. Lensch, and Jan Kautz. Two-shot spatially-varying brdf and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer: Learning robust image representations via transformers and temperature-based depth for multi-view stereo. *arXiv preprint arXiv:2208.02541*, 2022.

[8] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. Self-calibrating deep photometric stereo networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8747, 2019.

[9] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee Kenneth Wong. Deep photometric stereo for non-lambertian surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[10] Guanying Chen, Kai Han, and Kwan-Yee K Wong. Ps-fcn: A flexible learning framework for photometric stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018.

[11] Guanying Chen, Michael Waechter, Boxin Shi, Kwan-Yee K Wong, and Yasuyuki Matsushita. What is learned in deep uncalibrated photometric stereo? In *European Conference on Computer Vision*, 2020.

[12] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992.

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

[14] 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.

[15] 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.

[16] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1434–1441. IEEE, 2010.

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

[18] Khang Truong Giang, Soohwan Song, and Sungho Jo. Curvature-guided dynamic scale networks for multi-view stereo. *arXiv preprint arXiv:2112.05999*, 2021.

[19] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M Seitz. Multi-view stereo for community photo collections. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[20] 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. 2019.

[21] 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.

[22] Carlos Hernandez, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):548–554, 2008.

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

[24] Satoshi Ikehata. Ps-transformer: Learning sparse photometric stereo network using self-attention mechanism. *arXiv preprint arXiv:2211.11386*, 2022.

[25] 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.

[26] Sing Bing Kang, Richard Szeliski, and Jinxiang Chai. Handling occlusions in dense multi-view stereo. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001.

[27] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Uncertainty-aware deep multi-view photometric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12601–12611, 2022.

[28] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Multi-view photometric stereo revisited. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3126–3135, 2023.

[29] Berk Kaya, Suryansh Kumar, Francesco Sarno, Vittorio Ferrari, and Luc Van Gool. Neural radiance fields approach to deep multi-view photometric stereo. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1965–1977, 2022.

[30] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, page 0, 2006.

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

[32] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[33] Christoph Lassner and Michael Zollhöfer. Pulsar: Efficient sphere-based neural rendering. *arXiv:2004.07484*, 2020.

[34] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020.

[35] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. In *SIGGRAPH Asia 2018 Technical Papers*, page 269. ACM, 2018.

[36] Daniel Lichy, Soumyadip Sengupta, and David W Jacobs. Fast light-weight near-field photometric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12612–12621, 2022.

[37] Daniel Lichy, Soumyadip Sengupta, and David W. Jacobs. Fast light-weight near-field photometric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12612–12621, June 2022.

[38] Daniel Lichy, Jiaye Wu, Soumyadip Sengupta, and David W. Jacobs. Shape and material capture at home. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6123–6133, June 2021.

[39] Daniel Lichy, Jiaye Wu, Soumyadip Sengupta, and David W Jacobs. Shape and material capture at home. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6123–6133, 2021.

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

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

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

[43] 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.

[44] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. *ACM transactions on graphics (TOG)*, 24(3):536–543, 2005.

[45] 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.

[46] Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Multiview photometric stereo using planar mesh parameterization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1161–1168, 2013.

[47] Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Robust multiview photometric stereo using planar mesh parameterization. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1591–1604, 2016.

[48] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[49] 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.

[50] 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.

[51] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018.

[52] Boxin Shi, Zhe Wu Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evalution for non-lambertian and uncalibrated photometric stereo. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 41(2):271–284, 2019.

[53] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021.

[54] Christoph Strecha, Rik Fransens, and Luc Van Gool. Wide-baseline stereo from multiple views: a probabilistic account. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[55] Christoph Strecha, Rik Fransens, and Luc Van Gool. Combined depth and outlier estimation in multi-view stereo. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2394–2401. IEEE, 2006.

[56] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. *CVPR*, 2021.

[57] Olivia Wiles and Andrew Zisserman. Silnet : Single- and multi-view reconstruction by learning from silhouettes. In *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*. BMVA Press, 2017.

[58] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980.

[59] Qingshan Xu and Wenbing Tao. Learning inverse depth regression for multi-view stereo with correlation cost volume. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12508–12515, 2020.

[60] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (TOG)*, 37(4):126, 2018.

[61] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K Wong. Ps-nerf: Neural inverse rendering for multi-view photometric stereo. *arXiv preprint arXiv:2207.11406*, 2022.

[62] 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.

[63] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[64] 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.

[65] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5453–5462, 2021.

[66] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021.

[67] Zhengyou Zhang. Iterative point matching for registration of free-form curves and surfaces. *International journal of computer vision*, 13(2):119–152, 1994.

[68] Zhenglong Zhou, Zhe Wu, and Ping Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1482–1489, 2013.