PVP-Recon: Progressive View Planning via Warping Consistency for Sparse-View Surface Reconstruction

SHENG YE, YUZE HE, MATTHIEU LIN, JENNY SHENG, and RUOYU FAN, Tsinghua University, China
YIHENG HAN, Beijing University of Technology, China
YUBIN HU, Tsinghua University, China
RAN YI, Shanghai Jiao Tong University, China
YU-HUI WEN*, Beijing Jiaotong University, China
YONG-JIN LIU*, Tsinghua University, China
WENPING WANG, Texas A&M University, USA

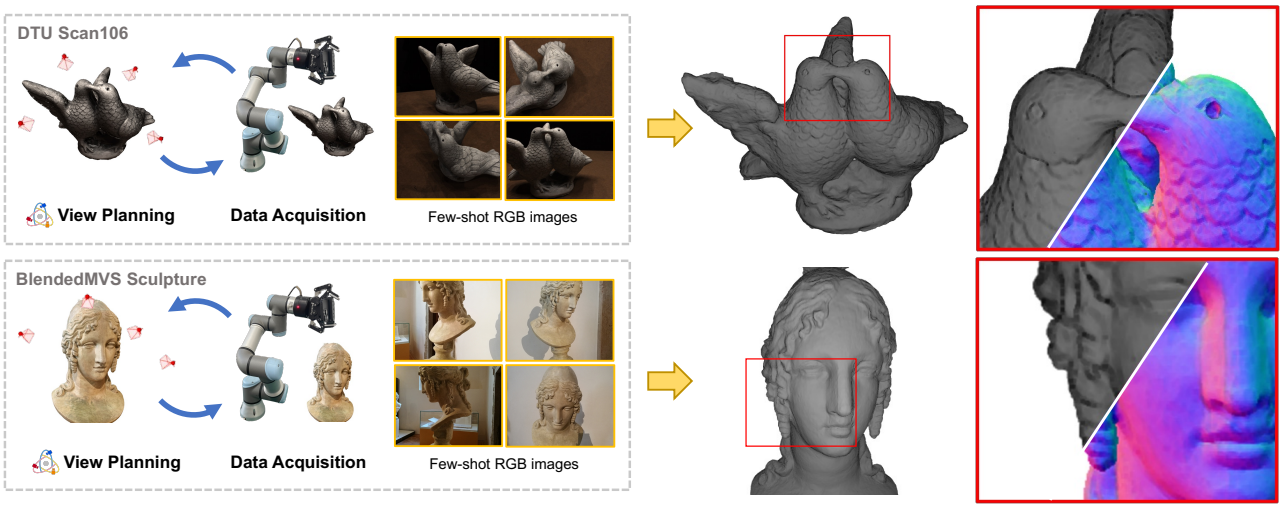

Fig. 1. We present *PVP-Recon*, a novel system that uses warping consistency to progressively determine the most informative views as input for surface reconstruction. *PVP-Recon* produces high-quality mesh surfaces with a constrained budget on the number of input views. Colors indicate surface normals.

Neural implicit representations have revolutionized dense multi-view surface reconstruction, yet their performance significantly diminishes with sparse input views. A few pioneering works have sought to tackle this challenge by leveraging additional geometric priors or multi-scene generalizability. However, they are still hindered by the imperfect choice of input views, using images under empirically determined viewpoints. We propose *PVP-Recon*, a novel and effective sparse-view surface reconstruction method that progressively plans the next best views to form an optimal set of sparse viewpoints for image capturing. *PVP-Recon* starts initial surface reconstruction with as few as 3 views and progressively adds new views which are determined based on a novel warping score that reflects the information gain of each newly added view. This progressive view planning progress is interleaved with a neural SDF-based reconstruction module that utilizes multi-resolution hash features, enhanced by a progressive training scheme and a directional Hessian loss. Quantitative and qualitative experiments on three benchmark datasets show that our system achieves high-quality reconstruction with a constrained input budget and outperforms existing baselines.

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Shape modeling**; **Mesh geometry models**;

Additional Key Words and Phrases: Neural surface reconstruction, sparse views, adaptive view planning, regularization techniques

ACM Reference Format:
Sheng Ye, Yuze He, Matthieu Lin, Jenny Sheng, Ruoyu Fan, Yiheng Han, Yubin Hu, Ran Yi, Yu-Hui Wen, Yong-Jin Liu, and Wenping Wang. 2024. PVP-Recon:

*Corresponding authors.

Authors' Contact Information: Sheng Ye, yec22@mails.tsinghua.edu.cn; Yuze He, hyz22@mails.tsinghua.edu.cn; Matthieu Lin, yh-lin21@mails.tsinghua.edu.cn; Jenny Sheng, jsheng415@outlook.com; Ruoyu Fan, fry21@mails.tsinghua.edu.cn, Tsinghua University, Beijing, China; Yiheng Han, Beijing University of Technology, Beijing, China, hanyiheng@bjut.edu.cn; Yubin Hu, Tsinghua University, Beijing, China, huyb20@mails.tsinghua.edu.cn; Ran Yi, Shanghai Jiao Tong University, Shanghai, China, ranyi@sjtu.edu.cn; Yu-Hui Wen, Beijing Jiaotong University, Beijing, China, yhwen1@bjtu.edu.cn; Yong-Jin Liu, Tsinghua University, Beijing, China, liuyongjin@tsinghua.edu.cn; Wenping Wang, Texas A&M University, College Station, USA, wenping@tamu.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 1557-7368/2024/12-ART191
https://doi.org/10.1145/3687896

Progressive View Planning via Warping Consistency for Sparse-View Surface Reconstruction. *ACM Trans. Graph.* 43, 6, Article 191 (December 2024), 12 pages. https://doi.org/10.1145/3687896

1 Introduction

3D mesh surface reconstruction from multi-view RGB images has always been an important issue in computer vision and graphics. Classic multi-view stereo (MVS) algorithms [Schönberger et al. 2016; Yao et al. 2018] are well established in the field of 3D reconstruction, but struggle to handle areas with homogeneous textures. Recently, neural implicit methods [Wang et al. 2021; Yariv et al. 2021] propose to represent the scene as a signed distance field (SDF) and optimize mesh surfaces directly from 2D image supervision through differentiable rendering. Although neural implicit methods surpass previous works in achieving high-fidelity mesh reconstruction, they typically rely on densely captured images as inputs, whose acquisition process can be time-consuming. To alleviate this reliance, some methods attempt to reconstruct 3D mesh from a predefined set of sparse views, based on geometric priors [Yu et al. 2022] or generalizable priors [Long et al. 2022]. Evidently, the surface reconstruction quality will depend on the input sparse views. However, it remains unknown how to determine a suitable set of sparse input views to achieve good surface reconstruction of a given object.

In this paper, we observe that two key factors contribute to good surface reconstruction under sparse views: (1) an effective view planning strategy to provide the most informative input images; and (2) a suitable neural representation with well-designed regularization techniques to encode the geometry. Thus, we propose *PVP-Recon*, a new sparse-view surface reconstruction system that contains two modules: (1) a view planning module for identifying most informative viewpoints for additional image capturing on the fly during surface optimization; and (2) a neural surface reconstruction module utilizing supplemented additional views to gradually improve reconstruction quality. The modules are easy to use and highly flexible, and they can be switched out in a modular manner against other existing approaches.

Different from existing sparse-view reconstruction methods, our method does *not* use a predefined set of sparse images as input. Instead, we use our view planning module to progressively identify the most informative viewpoints for capturing additional images during the reconstruction process until a high-quality surface is obtained. Hence, we have essentially solved a *next-best-view* problem [Banta et al. 2000] in the setting of image-based surface reconstruction. We propose a novel warping-based strategy, which selects the viewpoint with maximum information gain for subsequent reconstruction by calculating the cross-view rendering consistency for each candidate viewpoint. As a result, *PVP-Recon* typically ends up successfully reconstructing an object using no more than 8 images.

Our reconstruction module uses multi-resolution hash features to represent the SDF of object surfaces due to their expressiveness. Moreover, we introduce a progressive training scheme that gradually activates higher-resolution hash features, and a novel directional Hessian loss to regularize the neural SDF field and encourage the SDF under sparse views to be more smooth, which in combination make the optimization process more robust by preventing the model from quickly overfitting. The reconstruction module can provide renderings to the view planning module for decision-making, and facilitates the optimization of hash features under progressively added sparse views with the directional Hessian loss.

In summary, we make the following three contributions: (1) A novel object-level surface reconstruction system from sparse input views, where the inputs are progressively supplemented during the surface optimization process; (2) An effective view planning strategy based on image warping consistency; (3) A progressive training scheme enhanced by a directional Hessian loss to facilitate high-quality 3D reconstruction.

2 Related Works

2.1 Neural Implicit Representation

Recently, neural implicit representations have emerged as a powerful tool to encode the 3D geometry of a target object due to their compactness and remarkable performance. Occ-Net [Mescheder et al. 2019] and DeepSDF [Park et al. 2019] first propose to use neural networks to model shapes as occupancy or signed distance functions. However, they require ground-truth 3D data to supervise the learning process. Some follow-up works try to incorporate neural implicit functions and differentiable rendering to recover surfaces only with multi-view 2D image supervision. Specifically, IDR [Yariv et al. 2020] designs a differentiable rendering framework for implicit shape and appearance representations. NeuS [Wang et al. 2021] and VolSDF [Yariv et al. 2021] utilize signed distance fields to represent implicit surfaces and adopt the volume rendering technique introduced in NeRF [Mildenhall et al. 2020]. To enhance the expressive capability, Neuralangelo [Li et al. 2023] replaces the MLPs used in previous methods with multi-resolution hash grid features to encode implicit geometric functions.

2.2 Sparse View Reconstruction

Although neural implicit methods can produce remarkable meshes with dense input views, their performance drops drastically under sparse views due to the shape-radiance ambiguity [Zhang et al. 2020]. To address this challenge, a few pioneering works have been proposed, which can be divided into two main categories: regularization-based and generalization-based. Regularization-based methods utilize semantic priors [Jain et al. 2021], geometric priors [Deng et al. 2022; Yu et al. 2022], frequency priors [Yang et al. 2023], or MVS priors [Wu et al. 2023] as extra constraints, facilitating the ill-posed optimization under sparse views. Generalization-based methods aim to train on multiple scenes to gain generalization to unseen scenes. There have already been some successful attempts [Chen et al. 2021; Johari et al. 2022; Yu et al. 2021] at neural rendering. In terms of surface reconstruction, SparseNeuS [Long et al. 2022] learns generalizable priors from image features and adopts cascaded volumes for surface prediction. DiViNeT [Vora et al. 2023] learns a set of templates across different scenes, which serve as anchors in new scenes. To generate more details, VolRecon [Ren et al. 2023] represents the geometry as the Signed Ray Distance Function (SRDF) and combines multi-scale features to regress SRDF values using a ray transformer. However, neither regularization-based nor generalization-based methods consider the importance of input view planning for sparse reconstruction. Regularization-based methods typically fix a few

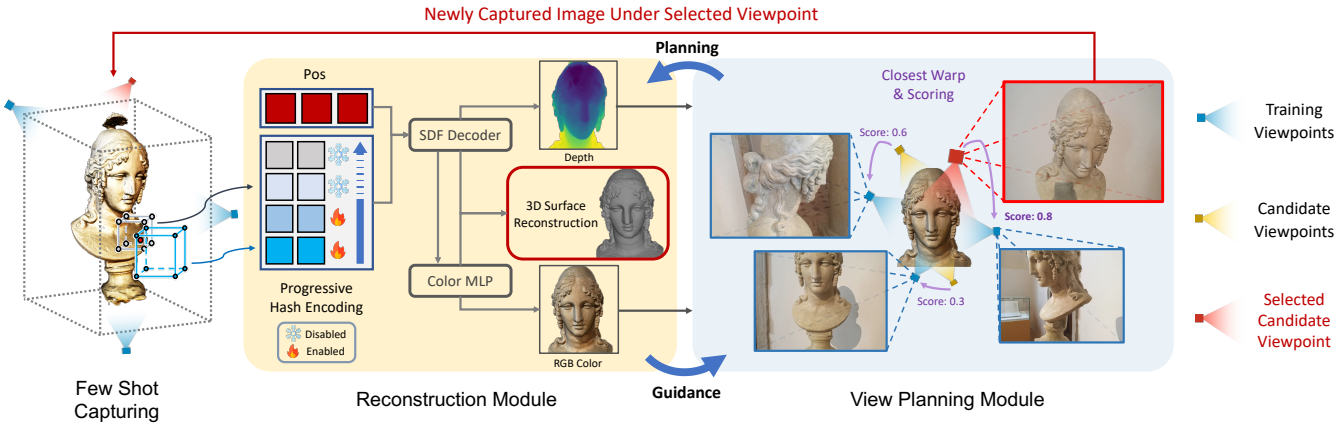

Fig. 2. Overview of our proposed system. The overall framework consists of two collaborative modules. In the reconstruction module, we use hash encoding to represent the SDF of object geometry, and adopt a training scheme that progressively enables finer levels of hash features. In the view planning module, we design a simple yet effective warping-based scoring strategy that progressively supplements the most informative views for surface reconstruction.

images as input before training, while generalization-based methods only work well under cautiously selected images with large overlap [Vora et al. 2023].

2.3 Neural Implicit View Planning

Some recent methods [Jin et al. 2023; Lee et al. 2022; Pan et al. 2022; Ran et al. 2023; Sünderhauf et al. 2023] attempt to adopt neural implicit representations for robot active planning. These methods determine the next best view based on uncertainty estimation with different policies. Most of these active planning methods [Jin et al. 2023; Pan et al. 2022; Sünderhauf et al. 2023] focus on synthesizing novel views rather than 3D surface reconstruction. Closely related to our work, two studies [Lee et al. 2022; Ran et al. 2023] also aim at view planning for 3D surface reconstruction. Lee et al. [2022] propose to calculate the density entropy along each ray as the uncertainty measurement, which is reasonable but susceptible to noise. NeurAR [Ran et al. 2023] models the emitted radiance as Gaussian distributions and evaluates the uncertainty by aggregating the variance. However, this approach may lead to degraded quality and training instability. The aforementioned methods mainly concentrate on designing uncertainty estimation strategies of implicit representations, without considering discernible bias in surface reconstruction. We propose a simple yet effective view planning strategy for high-quality object-level surface reconstruction under sparse views by evaluating the multi-view consistency.

3 Method

We propose *PVP-Recon*, a novel system consisting of a view planning module and a reconstruction module, to reconstruct high-quality object surfaces from a sparse set of RGB images. Specifically, the view planning module uses a warping-based strategy to determine the most informative viewpoint for subsequent image capture (Section 3.2). The reconstruction module adopts a *progressive hash encoding* as the geometric representation, facilitated by a novel directional Hessian loss, for reconstructing high-quality object surfaces (Section 3.3). The overall framework is illustrated in Figure 2.

We assume that view planning and surface reconstruction are mutually reinforcing. The newly acquired image from the view planning module is supplemented to the reconstruction module to help its optimization of recovering 3D surfaces. In turn, the reconstruction module provides current optimization status to guide the view planning module to make further planning decisions. We alternately plan subsequent input views and optimize the mesh surface, trying to achieve the best reconstruction results with a few images.

3.1 Preliminaries

3.1.1 Multi-resolution hash encoding. Instant-NGP [Müller et al. 2022] first proposes a multi-resolution hash encoding representation. The hash encoding partitions the space into multi-resolution grids of L levels $\{V_1, ..., V_L\}$. Using the spatial hash function [Teschner et al. 2003], each grid cell corner is mapped to a hash entry that stores a learnable feature vector. Given a 3D point $\boldsymbol{x}$, we map it to different locations $\boldsymbol{x}_l$ at each grid resolution V_l and retrieve the corresponding features $\gamma_l(\boldsymbol{x}_l) \in \mathbb{R}^F$ via trilinear interpolating hash entries at cell corners. The feature vectors of each level and auxiliary inputs $\xi \in \mathbb{R}^E$ (such as the view direction) are concatenated together to form the final encoded feature:

$$\gamma(\boldsymbol{x}) = (\xi, \gamma_1(\boldsymbol{x}_1), ..., \gamma_L(\boldsymbol{x}_L)). \tag{1}$$

The encoded feature $\gamma(\boldsymbol{x}) \in \mathbb{R}^{LF+E}$ is then passed to a shallow MLP decoder to regress the final output. Hash collisions are not handled explicitly, as the decoder is assumed to disambiguate collisions.

3.1.2 SDF-based volume rendering. NeuS [Wang et al. 2021] proposes to encode 3D scene as a signed distance field $f_g(\boldsymbol{x}; \theta_g)$ that outputs the SDF value of any specified location $\boldsymbol{x}$, and a color field $f_c(\boldsymbol{x}, \boldsymbol{d}; \theta_c)$ that outputs the view-dependent emitted radiance. Both fields can learn from 2D image supervision using the differentiable volume rendering technique. Given a ray $\boldsymbol{r}$ emitted from a posed camera, the rendering scheme first samples a set of points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$ along the ray. Then, the rendered pixel color $C(\boldsymbol{r})$

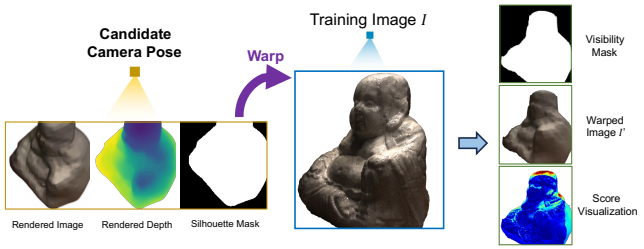

Fig. 3. Illustration of our view planning strategy on DTU scan114. For each candidate view, we render the current reconstructed mesh from this view. Then, we use the rendered depth and silhouette mask to warp the rendered RGB image into the closest training view and evaluate the warping score.

of this ray $\boldsymbol{r}$ can be calculated as:

$$C(\boldsymbol{r}) = \sum_{i=1}^{N} T_i \alpha_i f_c(\boldsymbol{x}_i, \boldsymbol{d}),\ T_i = \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{2}$$

where $\boldsymbol{d}$ is the direction vector of ray $\boldsymbol{r}$, and α_i denotes the discrete opacity of the i-th ray segment. Formally, α_i is defined as:

$$\alpha_i = \max\left(0, \frac{\Phi_\tau(f_g(\boldsymbol{x}_i)) - \Phi_\tau(f_g(\boldsymbol{x}_{i+1}))}{\Phi_\tau(f_g(\boldsymbol{x}_i))}\right), \tag{3}$$

where Φ_τ is the Sigmoid function with a learnable parameter τ. The final surface S can be extracted as the zero level-set of the SDF, *i.e.*, $S = \{\boldsymbol{x} \in \mathbb{R}^3 | f_g(\boldsymbol{x}) = 0\}$, by the Marching Cubes method.

3.2 View Planning Module With Warping Consistency

Our view planning module maintains a set of images for surface recovery, whose number dynamically increases as training proceeds. This module first discretizes the continuous camera space into a set of candidate viewpoints. Here, viewpoints refer to camera poses from which subsequent images can be captured, and can be easily obtained through uniform sampling or predefined. Then, we cluster all candidate camera poses and obtain the images under the cluster center poses as initial training images (typically 3 views). At regular training intervals, we score each remaining candidate pose and supplement the training image set with a newly captured image under the highest-scoring pose, until the view number reaches a given threshold. Note that when our view planning strategy evaluates the set of candidate camera poses, we assume that no images under these poses are already captured or available.

The main challenge is to calculate a reasonable score for each candidate camera pose that indicates its potential contribution (the amount of additional information it can provide) to future reconstruction. Our key insight is that the consistency of rendered images and depths from the reconstruction module can effectively reflect whether a specific area needs more information from an additional image. We aim to examine the multi-view consistency of renderings by drawing attention from the image-warping technique. Specifically, denote the relative transformation matrix between two camera poses as T, the camera intrinsic matrix as K, the depth map as d, then the image-warping process can be computed as:

$$p' = K\mathcal{F}^{-1}\{T\mathcal{F}\{d(p)K^{-1}p\}\}, \tag{4}$$

where p and p' are the homogeneous locations of the source image and the warped image. $\mathcal{F}$ is the homogeneous conversion from 3×1 to 4×1 vectors. To calculate the warping score for a specific candidate pose, we first render the image, depth, and silhouette mask from our reconstruction module under that pose. The rendered image is then warped to its closest training view, *i.e.,* an existing training image whose pose is closest to this candidate pose. The silhouette mask is also warped and acts as a visibility mask, indicating which areas of this training view are visible to the rendered image. The warping score $\boldsymbol{s}$ is defined as the *photometric difference* between the warped image I' and the training image I within the visibility mask M:

$$\boldsymbol{s} = \left\| M \odot (I' - I) \right\|_1. \tag{5}$$

Figure 3 illustrates our proposed planning strategy. There are two main factors contributing to a high warping score as defined in Eqn. 5. First, this photometric difference between I' and I can be caused by incomplete training of the 3D neural representation, revealing errors in current reconstructed geometry. Second, the difference between I' and I can be attributed to the multi-view inconsistency caused by occlusion, that is, there are some regions of the current reconstructed 3D model that are visible to the view under candidate pose, but not present in the training image I, suggesting that a new view under this candidate pose is likely to provide coverage to new regions that have not been covered by the existing training views. While the definition of our warping score takes both factors into account, we observe that the latter contributes more to the score. To sum up, by progressively selecting the highest-scoring viewpoints, we attempt to seek the best next views with maximum information gain to improve subsequent surface reconstruction.

3.3 Reconstruction Module With Progressive Scheme

In this module, we represent the object geometry as a signed distance field (SDF) encoded by multi-resolution hash features. Although hash features converge quickly and can capture fine-grained details, we observe that directly using hash features at all resolutions leads to noise and floating artifacts with sparse input views. This is because the expressive ability of multi-resolution hash features is too powerful. They can easily overfit to the few-shot training views, causing the optimized SDF to fall into local minima.

Inspired by FreeNeRF [Yang et al. 2023], we assume that the low-resolution hash features encode a coarse geometric shape, while the high-resolution features represent high-frequency information. Hence, to avoid quickly overfitting and unwanted artifacts, we adopt a carefully designed training scheme that progressively activates hash features. In the early stages of training, we only use low-resolution hash features to generate overly smooth geometries. In the later stages, we gradually use higher-resolution hash features to compensate for fine-grained details. This scheme can be achieved by applying a progressive activation mask:

$$\begin{aligned} \widetilde{\gamma}(\boldsymbol{x}, \psi) &= m(\psi) \odot \gamma(\boldsymbol{x}), \\ m(\psi) &= (m_0(\psi), m_1(\psi), ..., m_L(\psi)), \end{aligned} \tag{6}$$

where $m_i(\psi) = \boldsymbol{I}[i \le \psi]$, and ψ controls the bandwidth of the hash encoding. In practice, we set $\psi = \frac{i}{\Theta}L$, where i is the current training iteration, Θ is a predefined threshold, and L is the resolution level of hash features. Our progressive activation scheme shares

Table 1. Quantitative results on DTU dataset (bold means best, underline means second best). We report the Chamfer distance (mm) ↓.

Methods	Time	scan55	scan65	scan69	scan83	scan105	scan106	scan110	scan114	scan118	scan122	mean
NeuS (*random*)	83min	1.44	2.03	1.21	1.66	1.01	0.99	3.60	0.48	1.14	1.23	1.48
NeuS (*cluster*)		0.76	1.90	1.17	1.57	0.98	0.98	3.11	0.49	0.96	0.77	1.27
NeuS (*farthest*)		1.45	1.76	1.87	1.62	1.25	1.26	3.35	0.63	1.37	1.04	1.56
Neuralangelo (*random*)	512min	0.54	1.18	1.36	1.62	0.96	0.80	2.62	1.54	0.79	0.91	1.23
Neuralangelo (*cluster*)		0.52	1.43	1.15	1.43	**0.74**	**0.65**	2.98	0.99	0.88	0.70	1.15
Neuralangelo (*farthest*)		0.70	1.96	1.66	1.39	0.99	0.86	3.14	1.42	1.67	0.78	1.46
MonoSDF (*random*)	435min	0.79	1.24	1.19	1.44	0.80	3.05	1.53	0.68	1.34	2.34	1.44
MonoSDF (*cluster*)		0.89	**1.07**	1.09	1.35	0.97	2.74	1.29	0.71	1.38	1.73	1.32
MonoSDF (*farthest*)		0.98	1.53	1.39	1.37	0.91	2.64	1.90	0.86	1.68	1.45	1.47
SparseNeuS	–	0.84	1.87	1.07	1.51	1.26	1.11	1.09	0.74	1.46	1.83	1.28
VolRecon	–	0.92	1.92	1.01	1.58	0.89	1.09	1.48	0.63	1.20	1.12	1.18
Ours (*random*)	9min	0.64	1.34	0.84	1.52	0.95	1.22	1.13	0.48	0.93	0.62	0.97
Ours (*cluster*)		0.55	1.30	0.77	1.39	0.87	0.92	1.35	0.47	0.80	0.55	0.90
Ours (*farthest*)		0.76	1.38	1.22	1.43	0.89	1.05	1.47	0.49	0.91	0.51	1.01
Ours (*planning*)		**0.51**	1.15	**0.75**	**1.28**	0.84	0.91	**0.95**	**0.46**	**0.72**	**0.50**	**0.81**

some similarities with the coarse-to-fine strategy used in Li et al. [2023]. Different from theirs, our scheme focuses on solving the severe overfitting problem under sparse-view SDF optimization. This scheme plays a vital role in our framework, as we aim to achieve high-quality reconstruction without artifacts, and a well-optimized reconstruction module also provides more informative guidance to the view planning module.

3.4 Directional Hessian Loss

To produce reasonable geometry, recent works apply an Eikonal loss $\mathcal{L}_{eik}$ [Gropp et al. 2020] for regularizing the gradient of SDF to be close to one. However, we find that $\mathcal{L}_{eik}$ only considers the first-order gradient of SDF and is difficult to provide sufficient regularization, especially under challenging sparse-view settings. A higher-order gradient can provide a stronger constraint. Some studies [Li et al. 2023; Zhang et al. 2022] compute the second-order Hessian matrix and directly regularize the matrix norm to zero. Yet, we notice that the second-order gradient of SDF is not necessarily zero everywhere in space. Instead, the directional derivative of the SDF gradient along its normal direction should be zero due to the parallelism of adjacent SDF level sets near the surface. Therefore, we propose a directional Hessian loss $\mathcal{L}_{dir}$, which can constrain the second-order gradient more precisely:

$$\mathcal{L}_{dir} = exp(-\delta \cdot |f_g(\boldsymbol{x})|) \cdot \frac{|\nabla f_g(\boldsymbol{x})| - |\nabla f_g(\boldsymbol{x} + \epsilon \cdot \frac{\nabla f_g(\boldsymbol{x})}{\|\nabla f_g(\boldsymbol{x})\|})|}{\epsilon}, \quad (7)$$

where ∇ is the gradient operator, and we set the step size ϵ to equal to the minimum grid size of hash encoding; $exp(-\delta \cdot |f_g(\boldsymbol{x})|)$ is a radial basis function (RBF) with a hyperparameter δ, which encourages the loss to focus more on regions near the surface. Our proposed loss $\mathcal{L}_{dir}$ can serve as a smoothness constraint by regularizing the inconsistency of the SDF gradient, thus facilitating the ill-posed SDF optimization under sparse views.

3.5 Overall Loss

The total loss $\mathcal{L}$ used to optimize our model is:

$$\mathcal{L} = \mathcal{L}_{rgb} + w_{norm}\mathcal{L}_{norm} + w_{eik}\mathcal{L}_{eik} + w_{dir}\mathcal{L}_{dir}, \quad (8)$$

where the RGB loss $\mathcal{L}_{rgb} = \frac{1}{|\mathcal{R}|}\sum_{\boldsymbol{r}\in\mathcal{R}} \left\|C(\boldsymbol{r}) - \hat{C}(\boldsymbol{r})\right\|_1$ minimizes the difference between the rendered pixel $C(\boldsymbol{r})$ and ground-truth pixel $\hat{C}(\boldsymbol{r})$ for sampled ray $\boldsymbol{r}$. Here, $\mathcal{R}$ denotes the set of rays in each batch. To further regularize the surface, we also predict the surface normal using Omnidata [Eftekhar et al. 2021] and apply a normal loss. The normal loss $\mathcal{L}_{norm} = \frac{1}{|\mathcal{R}|}\sum_{\boldsymbol{r}\in\mathcal{R}} \left\|1 - N(\boldsymbol{r})^\top \hat{N}(\boldsymbol{r})\right\|_1$ constrains the rendered normal vector $N(\boldsymbol{r})$ to be consistent with the predicted pseudo ground-truth normal vector $\hat{N}(\boldsymbol{r})$.

4 Experiments

Implementation details. We adopt a 12-level hash encoding, with each hash entry having a channel size of 2. The total number of hash entries is 2^{19}. We progressively activate hash features with finer resolutions until $\Theta = 10,000$ iterations. We use 1 hidden layer with 64 hidden units for the SDF decoder and 2 hidden layers with 64 hidden units for the color MLP. We first train our framework using 3 clustered views for 1,000 iterations as initialization. Then, we add a new view using the proposed strategy every 1,000 iterations until the number of training views reaches a target value. The loss weights are set to $w_{norm} = w_{dir} = 0.05$, and $w_{eik} = 0.1$.

Datasets. We simulate our problem setting and conduct experiments on three datasets. Candidate viewpoints are limited to viewpoints of images in these datasets. Similar to some previous works [Long et al. 2022; Wang et al. 2021], we use representative scenes covering different aspects (*i.e.*, materials, appearances, and geometries) for evaluation. Specifically, we use 10 scenes from the DTU dataset [Jensen et al. 2014] and 5 challenging scenes (*i.e.*, Bear, Cattle, Clock, Man, Sculpture) from the BlendedMVS dataset [Yao et al. 2020] to evaluate the surface reconstruction quality. We further test on 5 scenes (we exclude scenes containing semi-transparent, hollow objects or fluids, which are not suitable for SDF-based surface reconstruction methods) from the Blender dataset [Mildenhall et al. 2020] to evaluate the novel view synthesis quality. For each dataset, we use only 10% of dense views for reconstruction. Concretely, we use 6 and 8 views for DTU and Blender, and 8-12 views for BlendedMVS.

Baselines. Our proposed view planning and reconstruction modules can be flexibly switched out in a modular manner against other

Table 2. Quantitative results on Blender dataset (bold means best, underline means second best). We report the PSNR and SSIM.

Methods	Time	Mic		Hotdog		Chair		Ficus		Materials		mean	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
NeuS (*random*)	91min	18.65	0.871	26.55	0.928	23.64	0.886	17.23	0.830	18.45	0.789	20.90	0.861
NeuS (*cluster*)		19.83	0.879	27.26	0.933	25.30	0.901	17.58	0.831	18.74	0.794	21.74	0.868
NeuS (*farthest*)		19.38	0.873	21.93	0.896	25.32	0.903	17.71	0.834	17.92	0.780	20.45	0.857
Neuralangelo (*random*)	518min	26.85	0.952	23.04	0.919	26.85	0.933	17.38	0.835	18.60	0.801	22.54	0.888
Neuralangelo (*cluster*)		26.60	0.947	25.08	0.928	**26.96**	**0.934**	17.57	0.833	**20.41**	0.819	23.32	0.892
Neuralangelo (*farthest*)		26.64	0.950	23.69	0.905	26.22	0.928	17.65	0.831	16.12	0.775	22.06	0.878
MonoSDF (*random*)	386min	23.58	0.928	24.99	0.919	20.69	0.848	17.59	0.852	19.40	0.827	21.25	0.875
MonoSDF (*cluster*)		23.29	0.927	25.24	0.924	23.61	0.882	19.52	0.863	19.65	**0.838**	22.26	0.887
MonoSDF (*farthest*)		22.99	0.925	23.39	0.902	23.38	0.883	18.89	0.861	19.67	0.830	21.66	0.880
Ours (*random*)	10min	27.65	0.951	26.97	0.935	25.16	0.910	22.36	**0.903**	18.12	0.804	24.05	0.901
Ours (*cluster*)		27.43	0.952	27.85	0.929	26.33	0.919	22.13	0.898	19.43	0.818	24.63	0.903
Ours (*farthest*)		26.45	0.950	26.24	0.917	25.46	0.911	22.30	0.895	18.25	0.807	23.74	0.896
Ours (*planning*)		**27.87**	**0.956**	**28.51**	**0.936**	26.54	0.925	**22.59**	0.899	20.23	0.832	**25.15**	**0.910**

exiting methods. For the reconstruction module, we compare with the following baselines: (1) The state-of-the-art surface reconstruction methods, including NeuS [Wang et al. 2021] and Neuralangelo [Li et al. 2023]; (2) The state-of-the-art sparse-view methods, including generalization-based SparseNeuS [Long et al. 2022], VolRecon [Ren et al. 2023] and regularization-based MonoSDF [Yu et al. 2022]. Because these methods use fixed input views before training, we apply three policies to select input views for them: *random sampling*, *cluster sampling*, *farthest sampling*. Exceptionally, for generalization-based methods, we use predefined views from their original methods for comparison, as their performance drops significantly when using other views as input. For the view planning module, we replace our proposed view planning strategy with different strategies introduced by Lee et al. [2022], NeurAR [Ran et al. 2023], and NeU-NBV [Jin et al. 2023] to show its effectiveness.

4.1 Quantitative Evaluation

To evaluate our reconstruction module, we use the same three policies to select input views and compare with baselines. In Table 1, we report Chamfer distance on the DTU dataset to measure the reconstruction accuracy. The object masks are used to remove the background for proper evaluation [Long et al. 2022]. To further evaluate image synthesis quality, we report the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) on the Blender dataset (see Table 2). On average, *PVP-Recon* achieves the lowest Chamfer distance and the highest PSNR and SSIM with significantly less training time. The results indicate that our reconstruction module generally outperforms existing works in terms of surface reconstruction and image rendering. We also notice that our reconstruction quality can be further improved after applying the view planning module (*i.e.*, *planning*).

For further evaluation of the view planning module, we compare our proposed strategy with three representative strategies in the field of robotics. Lee et al. [2022] calculate the density entropy along each ray and use this entropy as a measure for view planning. Yet, density struggles to provide the most valuable information for geometric reconstruction. Also, calculating entropy for each ray results in a heavy computational burden. NeurAR [Ran et al. 2023] and NeU-NBV [Jin et al. 2023] model the emitted radiance as Gaussian distributions and plans subsequent views using the distribution variance. Nevertheless, expanding the radiance from a specific value to a distribution induces randomness, which makes the optimization difficult and degrades the reconstruction quality. In contrast, our strategy is simple yet effective as it directly verifies the multi-view consistency via image warping. We implement and incorporate these strategies into our framework and conduct experiments on the DTU dataset. The entropy, variance maps, and warping scores are all calculated at 150×200 pixel resolution. Table 3 shows the averaged Chamfer distance and averaged running time for each round of view planning. Our strategy consistently outperforms other strategies under different input views with comparable running time.

Table 3. Averaged Chamfer distance and running time of different view planning strategies on 10 scenes of the DTU dataset.

Strategy Type	4 views	5 views	6 views	Time
Entropy [Lee et al. 2022]	1.20	1.07	0.93	59s
NeurAR [Ran et al. 2023]	1.51	1.36	1.23	**6.5s**
NeU-NBV [Jin et al. 2023]	1.35	1.25	1.12	11.1s
Ours	**0.99**	**0.92**	**0.81**	7.8s

4.2 Qualitative Evaluation

We visualize the reconstructed meshes and conduct qualitative comparisons (shown in Figure 4). For each baseline, we choose the best results from three view selection policies. Neuralangelo [Li et al. 2023] and MonoSDF [Yu et al. 2022] struggle to generate accurate geometries for textureless regions (Scan 110) or delicate structures (Scan 118). VolRecon [Ren et al. 2023] generates missing or noisy results. Moreover, baselines fail to generate good results on challenging scenes of the BlendedMVS dataset. In contrast, *PVP-Recon* can reconstruct both smooth surfaces and detailed structures.

Although our primary goal is surface reconstruction, we also evaluate rendering quality using the Blender dataset, which provides a separate test set for evaluating novel view synthesis. Figure 5 shows that *PVP-Recon* also outperforms baselines in terms of rendering quality. NeuS [Wang et al. 2021] and MonoSDF [Yu et al. 2022] synthesize blurry images because of the low-frequency bias

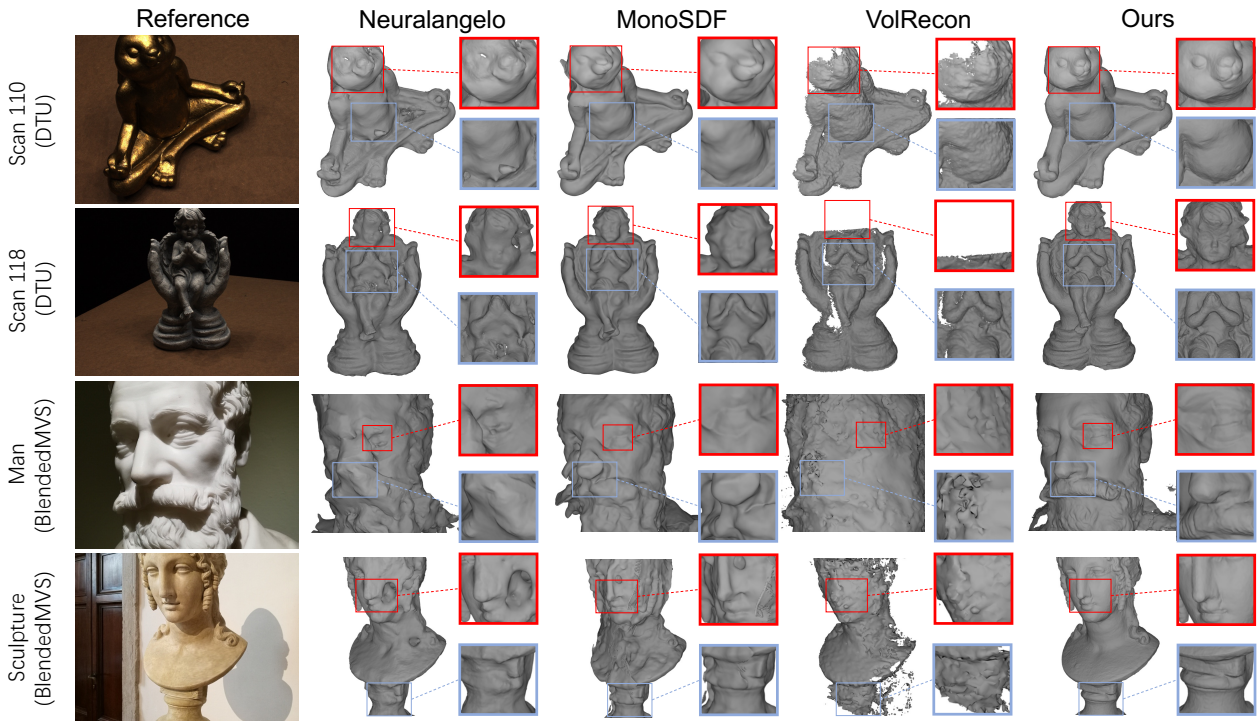

Fig. 4. Comparison of surface reconstruction on DTU and BlendedMVS datasets. Our method generates the most accurate and detailed results.

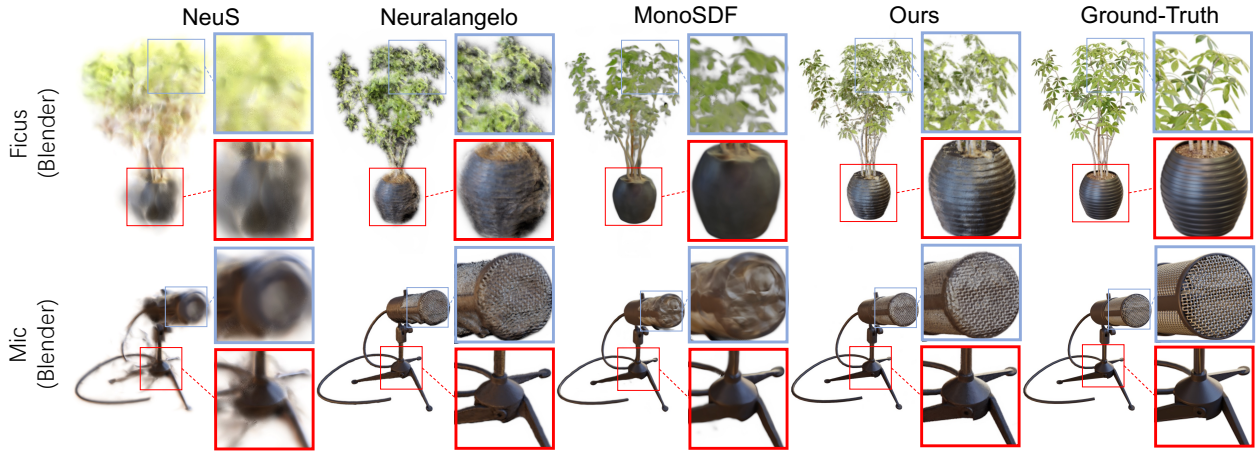

Fig. 5. Comparison of image rendering on Blender dataset. Our method can generate high-quality renderings with richer details.

of the MLP networks. Neuralangelo [Li et al. 2023] captures more details but produces incorrect textures in complex areas. This is due to its ill-posed optimization of hash features under sparse views. By utilizing our proposed training scheme and the directional Hessian loss $\mathcal{L}_{dir}$, our model is better optimized and recovers sharper and more visually appealing rendering results.

In Figure 6, we show the evolution process of the reconstructed mesh under progressively added input views. The recovered mesh becomes more precise and richer in details, as our view planning module can supplement the most informative views for surface reconstruction. We also visualize the selected cameras, corresponding RGB images, and final reconstruction results of different planning strategies for comparison in Figure 7. Input views planned by entropy-based strategy [Lee et al. 2022] tend to neglect edge regions, leading to incomplete surfaces of the target object. Variance-based strategies (*i.e.*, NeurAR [Ran et al. 2023] and NeU-NBV [Jin et al. 2023]) can plan reasonable input views, but introduce randomness into the surface optimization process, making the recovered meshes

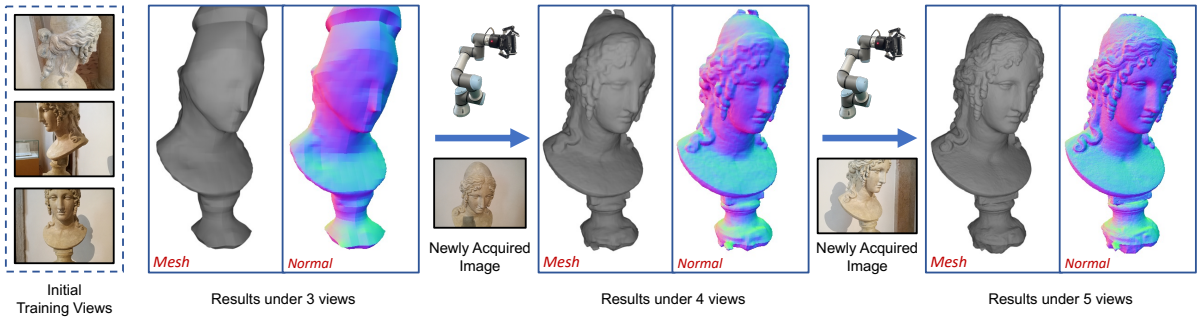

Fig. 6. We show how the reconstructed mesh of BlendedMVS Sculpture changes as new input views are progressively added. Note that the newly added views provide beneficial information that makes the reconstruction result more precise and detailed.

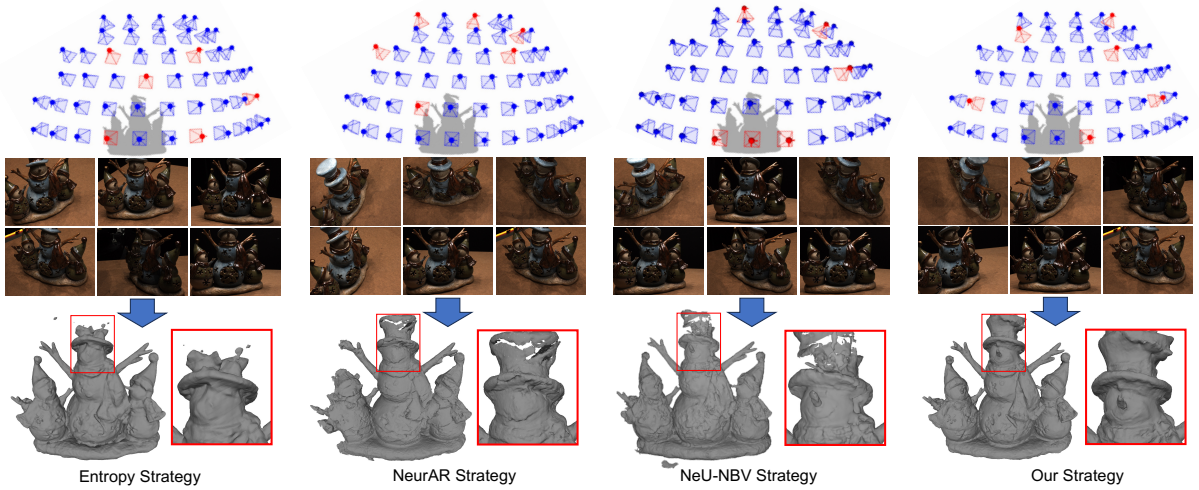

Fig. 7. We visualize the selected cameras, corresponding RGB inputs, and final recovered meshes (DTU scan69) of different view planning strategies.

Table 4. Ablation results (averaged Chamfer distance) on the DTU dataset.

Configs	*w/o prog*	*w/o* $\mathcal{L}_{dir}$	*full setup*
farthest sampling	1.53	1.16	**1.01**
view planning	1.45	0.91	**0.81**

bumpy. Our warping-based strategy achieves better capture coverage and results in high-quality and delicate surfaces.

4.3 Ablation Study

We conduct more ablation experiments on the DTU dataset. We remove the following key components separately: (1) progressive training scheme (*w/o prog*); (2) directional Hessian loss (*w/o* $\mathcal{L}_{dir}$). Note that we evaluate the effectiveness of these components both with and without (*i.e.*, *farthest sampling*) the view planning module, to avoid entangling the evaluation with changes in the input views. Table 4 and Figure 8 show the ablation results. In both settings, the full setup achieves the best results, and the reconstruction quality deteriorates when either component is removed. Specifically, the progressive scheme prevents the reconstruction optimization from quickly falling into local minima. Without this scheme (*w/o prog*), the model fails to capture scene details and produces messy geometries. Intuitively, our proposed loss $\mathcal{L}_{dir}$ constrains the gradient of SDF and acts as a consistency prior by regularizing abrupt changes in surface curvature. Therefore, the reconstructed meshes are non-smooth and inconsistent when this loss is removed (*w/o* $\mathcal{L}_{dir}$). As shown in Figure 8, our full setup generates the most accurate and complete surfaces with fine-grained details.

4.4 Different Numbers of Input Views

We evaluate the reconstruction quality under different numbers of input views. Figure 9 shows the Chamfer distance (C.D.) variation curves for three representative scenes in the DTU dataset. With fewer input views, the Chamfer distance decreases significantly each time a new image is added. Our *PVP-Recon* generally outperforms NeU-NBV [Jin et al. 2023] strategy under different numbers of views. We also report the reconstruction results of Neuralangelo [Li et al. 2023] using all dense images (64 views) of a scene. Note that we can

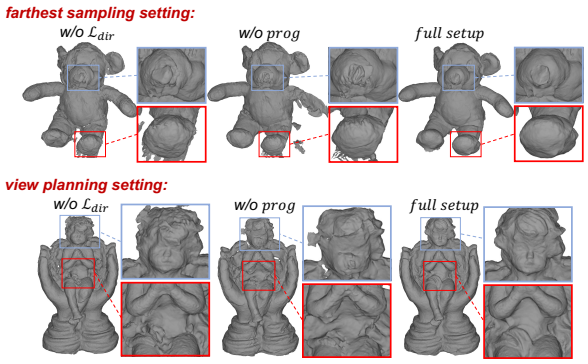

Fig. 8. Ablation visualizations (DTU scan105 and scan118). Note that Artifacts will appear when any of our proposed components are removed.

Table 5. Integrating our view planning module into different 3D reconstruction methods can achieve lower averaged Chamfer distance on the DTU dataset than predefining a fixed set of views before training.

Methods	predefined views	view planning
NeuS [Wang et al. 2021]	1.27	**1.15**
MonoSDF [Yu et al. 2022]	1.44	**1.36**

achieve comparable or even better results with fewer input images, which demonstrates the effectiveness of our system.

4.5 Flexibility of the View Planning Module

Our warping-based view planning module is flexible and can be combined with different 3D reconstruction methods. We incorporate the view planning module into NeuS [Wang et al. 2021] and MonoSDF [Yu et al. 2022] framework. Figure 10 and Table 5 show that, compared with predefining a fixed set of input views using the *cluster sampling*, our incorporated view planning module can further help NeuS and MonoSDF to reconstruct more accurate and complete surfaces, and achieve lower Chamfer distance.

4.6 Real-World Robotic Application

PVP-Recon alternately optimizes the surface and plans subsequent input views based on current optimization status. Therefore, *PVP-Recon* is well suited for active reconstruction in the field of robotics and can be applied in practice. In Figure 11, we show meshes reconstructed by *PVP-Recon* from sparse images progressively captured by a robotic arm. *PVP-Recon* can still generate accurate and high-quality 3D mesh surfaces in real-world robotic scenarios.

5 Limitations and Conclusion

Limitations. Currently, *PVP-Recon* takes 8 seconds each time for view planning and 10 minutes for the overall reconstruction. Future direction includes achieving further acceleration with CUDA implementation for applications with high requirements on real-time performance. Moreover, *PVP-Recon* now focuses on object-level scene reconstruction. In the future, we aim to combine our method with mobile robots or drones to achieve full 3D reconstruction of large, unbounded, and non-object-centric scenarios.

Conclusion. In this paper, we propose *PVP-Recon*, a novel sparse-view surface reconstruction system that progressively plans the next best views to form an optimal set of input images. Specifically, we design a scoring strategy that checks the multi-view consistency to seek the most informative images for further training. To stabilize the implicit surface optimization under sparse views, we also introduce a progressive training scheme and a directional Hessian loss. Extensive experiments on three datasets show that our method outperforms previous reconstruction baselines and active planning strategies. Furthermore, we demonstrate that *PVP-Recon* is well suited for active reconstruction task in the field of robotics, and our proposed view planning module can also be incorporated into other frameworks to facilitate better surface reconstruction.

Acknowledgments

This work was supported by Beijing Science and Technology plan project (Z231100005923029), the Natural Science Foundation of China (62332019, U2336214), and the Talent Fund of Beijing Jiaotong University (2023XKRC045).

References

Joseph E Banta, LR Wong, Christophe Dumont, and Mongi A Abidi. 2000. A next-best-view system for autonomous 3-D object reconstruction. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 30, 5 (2000), 589–598.

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. 2021. MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, US, 14104–14113.

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. 2022. Depth-supervised NeRF: Fewer Views and Faster Training for Free. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, US, 12872–12881.

Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. 2021. Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets From 3D Scans. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, US, 10786–10796.

Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. 2020. Implicit Geometric Regularization for Learning Shapes. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Vol. 119. PMLR, London, UK, 3789–3799.

Ajay Jain, Matthew Tancik, and Pieter Abbeel. 2021. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Canada, October 10-17, 2021*. IEEE, US, 5865–5874.

Rasmus Ramsbøl Jensen, Anders Lindbjerg Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. 2014. Large Scale Multi-view Stereopsis Evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, Washington, DC, 406–413.

Liren Jin, Xieyuanli Chen, Julius Rückin, and Marija Popovic. 2023. NeU-NBV: Next Best View Planning Using Uncertainty Estimation in Image-Based Neural Rendering. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, US, 11305–11312.

Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. 2022. GeoNeRF: Generalizing NeRF with Geometry Priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, US, 18344–18347.

Soomin Lee, Le Chen, Jiahao Wang, Alexander Liniger, Suryansh Kumar, and Fisher Yu. 2022. Uncertainty Guided Policy for Active Robotic 3D Reconstruction Using Neural Radiance Fields. *IEEE Robotics Autom. Lett.* 7, 4 (2022), 12070–12077.

Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H. Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. 2023. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, US, 8456–8465.

Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. 2022. SparseNeuS: Fast Generalizable Neural Surface Reconstruction from Sparse Views. In

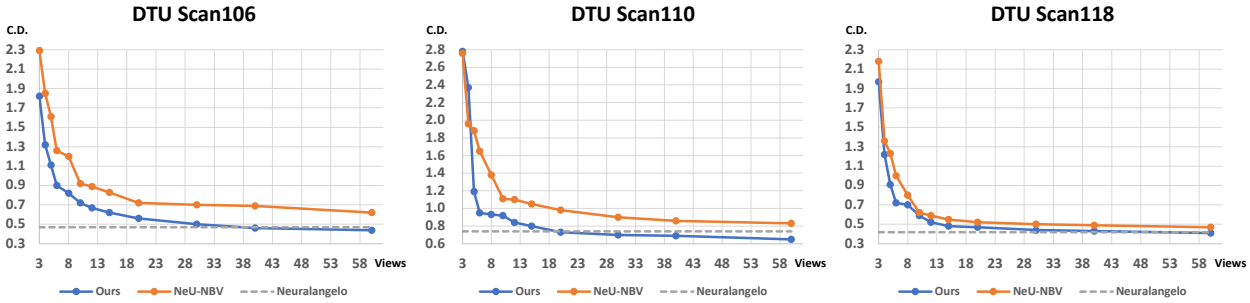

Fig. 9. Reconstruction results (the Chamfer distance, C.D. ↓) of ours and [Jin et al. 2023; Li et al. 2023] using different numbers of input views on DTU dataset.

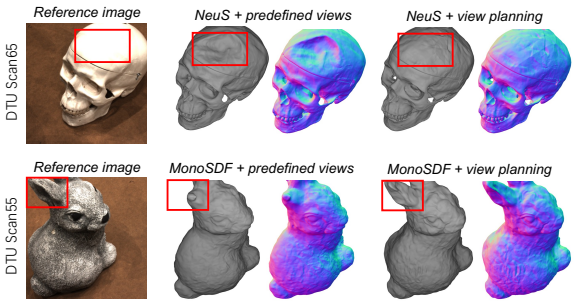

Fig. 10. Our proposed view planning module can be flexibly incorporated into different 3D reconstruction methods and help to reconstruct more accurate and complete surfaces.

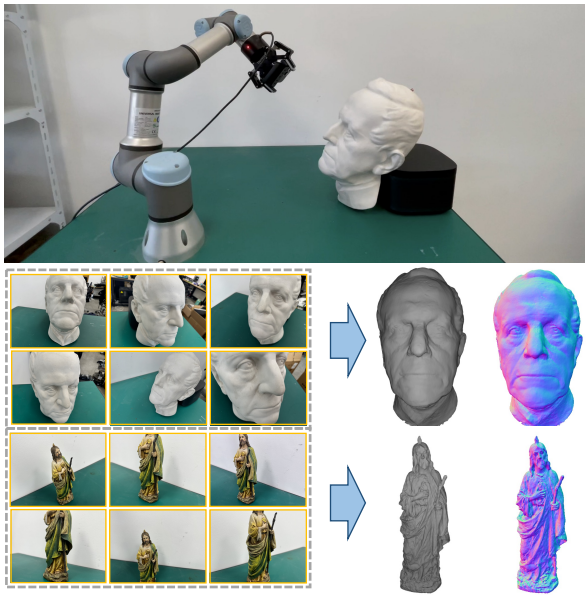

Fig. 11. The meshes reconstructed by *PVP-Recon* from sparse-view images progressively captured by a robotic arm.

ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII, Vol. 13692. Springer, Heidelberg, Germany, 210–227.

Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, New York, NY, 4460–4470.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, Vol. 12346. Springer, Heidelberg, Germany, 405–421.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.* 41, 4 (2022), 102:1–102:15.

Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. 2022. ActiveNeRF: Learning Where to See with Uncertainty Estimation. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, Vol. 13693. Springer, Heidelberg, Germany, 230–246.

Jeong Joon Park, Peter R. Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. 2019. DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, New York, NY, 165–174.

Yunlong Ran, Jing Zeng, Shibo He, Jiming Chen, Lincheng Li, Yingfeng Chen, Gimhee Lee, and Qi Ye. 2023. NeurAR: Neural Uncertainty for Autonomous 3D Reconstruction With Implicit Neural Representations. *IEEE Robotics Autom. Lett.* 8, 2 (2023), 1125–1132.

Yufan Ren, Fangjinhua Wang, Tong Zhang, Marc Pollefeys, and Sabine Süsstrunk. 2023. VolRecon: Volume Rendering of Signed Ray Distance Functions for Generalizable Multi-View Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*. IEEE, US, 16685–16695.

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *Computer Vision - ECCV 2016 - Proceedings, Part III*, Vol. 9907. Springer, Heidelberg, Germany, 501–518.

Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. 2023. Density-aware NeRF Ensembles: Quantifying Predictive Uncertainty in Neural Radiance Fields. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*. IEEE, US, 9370–9376.

Matthias Teschner, Bruno Heidelberger, Matthias Müller, Danat Pomerantes, and Markus H. Gross. 2003. Optimized Spatial Hashing for Collision Detection of Deformable Objects. In *8th International Fall Workshop on Vision, Modeling, and Visualization, VMV 2003, München, Germany, November 19-21, 2003*. Aka GmbH, Munich, Germany, 47–54.

Aditya Vora, Akshay Gadi Patil, and Hao Zhang. 2023. DiViNeT: 3D Reconstruction from Disparate Views via Neural Template Regularization. *arXiv* abs/2306.04699 (2023), 1–21.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Neural Information Processing Systems 34: NeurIPS 2021, December 6-14, 2021, virtual*. Curran Associates, New York, NY, 27171–27183.

Haoyu Wu, Alexandros Graikos, and Dimitris Samaras. 2023. S-VolSDF: Sparse Multi-View Stereo Regularization of Neural Implicit Surfaces. *arXiv* abs/2303.17712 (2023), 1–13.

Jiawei Yang, Marco Pavone, and Yue Wang. 2023. FreeNeRF: Improving Few-Shot Neural Rendering with Free Frequency Regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, US, 8254–8263.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. 2018. MVSNet: Depth Inference for Unstructured Multi-view Stereo. In *Computer Vision - ECCV 2018 -*

15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII, Vol. 11212. Springer, Heidelberg, Germany, 785–801.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. 2020. BlendedMVS: A Large-Scale Dataset for Generalized Multi-View Stereo Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, New York, NY, 1787–1796.

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume Rendering of Neural Implicit Surfaces. In *Neural Information Processing Systems 34: NeurIPS 2021, December 6-14, 2021, virtual*. Curran Associates, New York, NY, 4805–4815.

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. 2020. Multiview Neural Surface Reconstruction by Disentangling Geometry and Appearance. In *Neural Information Processing Systems 33: NeurIPS 2020, December 6-12, 2020, virtual*. Curran Associates, New York, NY, 1–11.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelNeRF: Neural Radiance Fields From One or Few Images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, New York, NY, 4578–4587.

Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. 2022. MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction. In *NeurIPS*. OpenReview.net, US, 1–15.

Jingyang Zhang, Yao Yao, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. 2022. Critical Regularizations for Neural Surface Reconstruction in the Wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, US, 6260–6269.

Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. NeRF++: Analyzing and Improving Neural Radiance Fields. *arXiv* abs/2010.07492 (2020), 1–9.

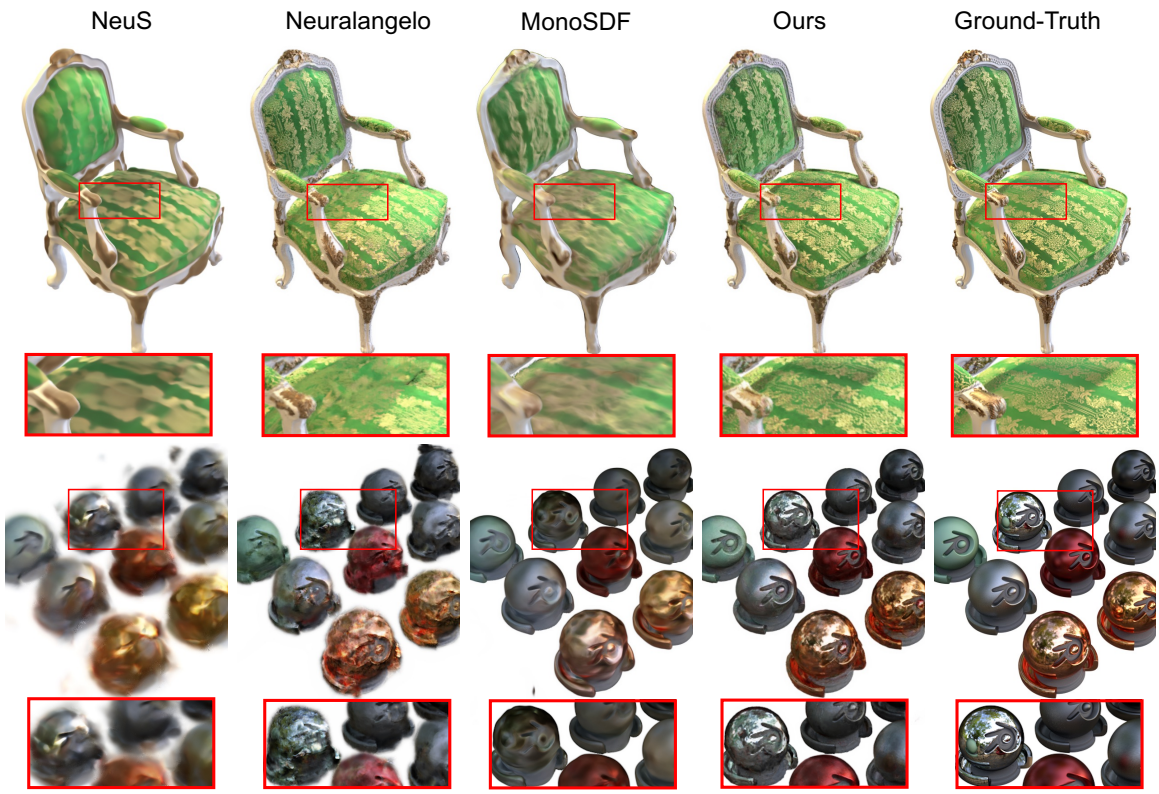

Fig. 12. Additional comparison results of image rendering on Blender. Our approach can generate high-quality renderings with richer details.

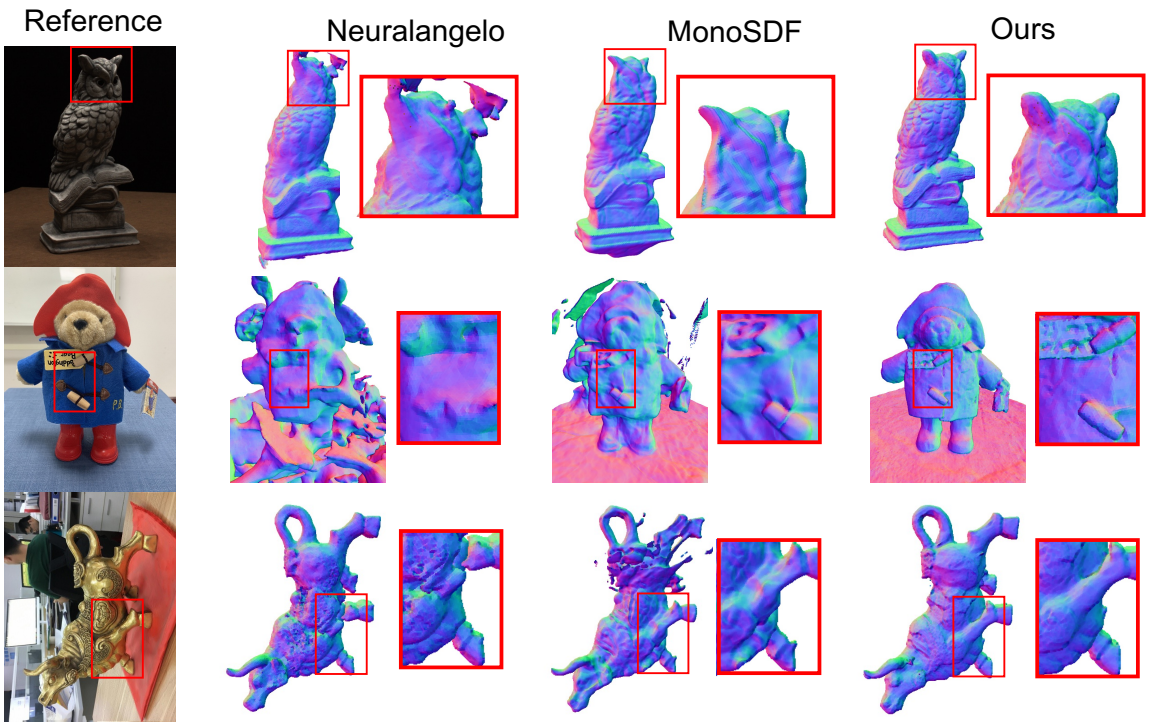

Fig. 13. Additional comparison results of surface normal on DTU and BlendedMVS. Note that our method produces more accurate and delicate results.