SpaRP: Fast 3D Object Reconstruction and Pose Estimation from Sparse Views

Chao Xu[1,2†], Ang Li[3], Linghao Chen[2,4†], Yulin Liu[5], Ruoxi Shi[2,5†], Hao Su[2,5‡], and Minghua Liu[2,5†‡]

[1] UCLA [2] Hillbot Inc. [3] Stanford University
[4] Zhejiang University [5] UC San Diego

Fig. 1: SpaRP handles open-world 3D reconstruction and pose estimation from unposed sparse-view images, delivering results in approximately 20 seconds.

Abstract. Open-world 3D generation has recently attracted considerable attention. While many single-image-to-3D methods have yielded visually appealing outcomes, they often lack sufficient controllability and tend to produce hallucinated regions that may not align with users' expectations. In this paper, we explore an important scenario in which the input consists of one or a few unposed 2D images of a single object, with little or no overlap. We propose a novel method, SpaRP, to reconstruct a 3D textured mesh and estimate the relative camera poses for these

† This work is done while the author is an intern at Hillbot Inc. ‡ Equal advisory.

sparse-view images. SpaRP distills knowledge from 2D diffusion models and finetunes them to implicitly deduce the 3D spatial relationships between the sparse views. The diffusion model is trained to jointly predict surrogate representations for camera poses and multi-view images of the object under known poses, integrating all information from the input sparse views. These predictions are then leveraged to accomplish 3D reconstruction and pose estimation, and the reconstructed 3D model can be used to further refine the camera poses of input views. Through extensive experiments on three datasets, we demonstrate that our method not only significantly outperforms baseline methods in terms of 3D reconstruction quality and pose prediction accuracy but also exhibits strong efficiency. It requires only about 20 seconds to produce a textured mesh and camera poses for the input views. Project page: `https://chaoxu.xyz/sparp`.

Keywords: Sparse-View 3D Reconstruction · Pose Estimation · Open-World Generation

1 Introduction

3D object reconstruction is a long-standing problem with applications spanning 3D content creation, augmented reality, virtual reality, and robotics, among others. Although traditional photogrammetry [1, 57, 64] and recent neural field methods [43, 77, 91] have made significant strides in reconstructing high-fidelity geometry and appearance, they typically require dense view inputs. However, in many practical scenarios, such as in e-commerce and consumer capture situations, acquiring a comprehensive set of high-resolution images along with precise camera data is not always feasible.

On the other end of the spectrum, the tasks of converting a single image to 3D and text to 3D have recently seen substantial progress [28,30,32,46,62,80], thanks to the rich priors embedded in 2D diffusion models [50, 53, 54] and pre-training on extensive 3D datasets [9]. These methods may achieve high-quality geometry and texture that matches the input view, but they also introduce ambiguities in the regions not visible in the input image (such as the back view). Although these methods attempt to hallucinate reasonable interpretations of these invisible areas, the generated regions may not always align with users' expectations, and users often lack sufficient control over these ambiguous regions.

In this paper, we explore a critical scenario where the input consists of one or a few unposed 2D images of a single object. The images are captured from arbitrarily distributed camera poses, often with little to no overlap. We tackle both the 3D reconstruction and pose estimation of input images under this sparse view setting. Note that, in dense view setting, traditional Structure-from-Motion (SfM) solvers (e.g., COLMAP [58]) are typically employed for pose estimation. However, with sparse view inputs, these solvers often become unreliable and tend to fail due to insufficient overlapping visual cues. This issue is the main reason why existing sparse view reconstruction methods [22, 39, 97] generally require known camera poses as input. While some recent methods have attempted pose-free reconstruction and pose estimation for sparse views [17, 18, 29, 63, 95], they are usually trained on a predefined small set of object categories and exhibit poor generalization to unseen object categories.

In response, we propose an innovative class-agnostic approach called SpaRP, capable of processing arbitrary object categories with unposed sparse views. Our inspiration comes from recent breakthroughs in open-domain single-image-to-3D methods. They leverage 2D diffusion models (e.g., Stable Diffusion [53]) to generate novel viewpoints of an object [35], and even consistent multi-view images from a single input image [28, 36, 38, 60, 62], by finetuning the diffusion models with corresponding multi-view image pairs. These discoveries imply that 2D diffusion models harbor rich priors concerning 3D objects. Instead of merely producing multi-view images, we contemplate leveraging 2D diffusion models to examine a set of unposed input images from sparse viewpoints, infer their spatial interrelationships, and recover relative camera poses and underlying 3D shapes.

Specifically, we finetune a 2D diffusion model [53] to process sparse input views by compositing them into a single image for conditioning. The diffusion model is concurrently tuned to deduce the relative poses of the input images and the underlying 3D objects. For the relative pose estimation branch, instead of outputting camera poses as scalars, we task 2D diffusion models to produce a surrogate representation: the NOCS maps [74] that embed pixel-wise correspondences across different views and are more suitable for 2D diffusion models. From these maps, we extract the relative camera poses for the sparse views using the traditional PnP algorithm [2], assuming known camera intrinsics. For the reconstruction branch, the diffusion model is tasked to produce multi-view images of the object from fixed known camera poses, covering the entire 3D object. This task requires the models to incorporate all information from input sparse views and hallucinate invisible regions. We then feed the generated images with fixed known poses into a pre-trained 3D reconstruction module [32] to create a textured 3D mesh. We can further refine the estimated camera poses by aligning the input views with the generated mesh through differentiable rendering [26].

We train SpaRP on the Objaverse [9] dataset with 1–6 unposed input views. Unlike some previous methods that rely on costly per-shape optimization [83], our method delivers 3D textured meshes along with camera poses in a much more efficient manner, requiring only ∼16 seconds. As shown in Fig. 1, our approach can faithfully generate 3D assets that closely follow the reference unposed images, effectively overcoming the ambiguity issue of single-image-to-3D. Extensive evaluation on three datasets demonstrates the superior performance of our method over baselines in reconstructing 3D meshes with vivid appearance and high-fidelity geometry, alongside precise pose estimation of the input images.

2 Related Work

2.1 Sparse-View 3D Reconstruction

Reconstructing 3D objects from sparse-view images is challenging due to the lack of visual correspondence and clues. When a small baseline between images is assumed, several methods [4, 19, 24, 37, 39, 51, 52, 70, 76, 79, 89, 93] have pretrained generalizable models to infer surface positions by establishing pixel correspondences and learning generalizable priors across scenes. However, these methods often fail to produce satisfactory results when the sparse-view images

have a large baseline. Some studies have attempted to alleviate the dependence on dense views by incorporating priors or adding regularization [16, 22, 45, 61] into the NeRF optimization process. Others have employed 2D diffusion priors to generate novel-view images as additional input for the NeRF model [3,21,68,97]. For example, ReconFusion [84] trains a NeRF from sparse-view images and uses a denoising UNet to infer some novel view images as support for the NeRF model. EscherNet [23] utilizes Stable Diffusion for novel view synthesis and designs a camera positional encoding module to yield more consistent images. Furthermore, some recent works [36, 38, 62] have integrated specialized loss functions and additional modalities as inputs into NeRF-based per-scene optimization.

In contrast to these methods, our approach does not require camera poses for the input sparse views. It is not limited to small baselines and is capable of generating 360-degree meshes. Furthermore, without the need for per-shape optimization, our method can quickly produce both textured meshes and camera poses in about 20 seconds.

2.2 Pose-Free Reconstruction

Unlike the methods mentioned above, which assume known camera poses, many studies have aimed to solve the pose-free reconstruction challenge. When provided with dense images, some approaches [31,81,86] jointly optimize the NeRF representation along with camera parameters. However, due to the highly non-convex nature of this optimization problem, such methods are susceptible to initial pose guesses and can become trapped in local minima. This issue worsens when input images are sparse, with increasing ambiguity and reduced constraint availability. In response, numerous proposals have attempted to enhance optimization robustness. For example, SpaRF [71] uses dense image matches as explicit optimization constraints, while FvOR [90] starts with coarse predictions of camera poses and alternated updates between shape and pose.

In contrast to the optimization-based methods, there is a body of research proposing generalizable solutions for this problem. VideoAE [25] infers scene geometry from the first frame in a video series and estimates camera poses relative to that frame, which allows for warping scene geometry to decode new viewpoints. SparsePose [63] first regresses and then iteratively refines camera poses. FORGE [17] designs neural networks to infer initial camera poses, fuse multi-view features, and decode spatial densities and colors. GRNN [72] offers a GRU-based reconstruction method estimating the relative pose for each input view against a global feature volume. The RelPose series [29,95] use probabilistic modeling for relative rotation estimation between images. Other works [18, 55] eschew explicit camera pose estimations, instead employing transformers to encode input views into latent scene representations for novel view synthesis.

More recently, leveraging large vision models and diffusion models, which have shown significant promise, new efforts have emerged for camera pose estimation. PoseDiffusion [75] implements a diffusion model guided by 2D keypoint matches to estimate poses. PF-LRM [78] adapts the LRM model [14] to predict a point cloud for each input image, then utilizes differentiable PnP for pose estimation. iFusion [83] employs an optimization pipeline to assess relative elevations

and azimuths. It utilizes Zero123 [35] predictions as a basis and optimizes the relative pose between two images by minimizing the reconstruction loss between the predicted and target images.

In contrast to these existing approaches, our proposal capitalizes on the extensive priors inherent in pre-trained 2D diffusion models, thereby providing exceptional generalizability to handle a diverse range of open-world categories. Our method predicts camera poses and 3D mesh geometry in a single feedforward pass, negating the need for per-shape optimization.

2.3 Open-World 3D Generation

Open-world single-image-to-3D and text-to-3D tasks have recently undergone significant advancements. Recent 2D generative models [50, 53, 54] and vision-language models [48] have supplied valuable priors about the 3D world, sparking a surge in research on 3D generation. Notably, models such as DreamFusion [46], Magic3D [30], and ProlificDreamer [80] have pioneered a line of approach to per-shape optimization [5, 6, 10, 15, 27, 40–42, 44, 47, 49, 59, 66, 67, 73, 87, 88, 94]. These models optimize a 3D representation (e.g., NeRF) for each unique text or image input, utilizing the 2D prior models for gradient guidance. Although they produce impressive results, these methods are hampered by prolonged optimization times, often extending to several hours, and "multi-face issue" problems.

Moreover, beyond optimization-based methods, exemplified by Zero123 [35], numerous recent studies have investigated the employment of pre-trained 2D diffusion models for synthesizing novel views from single images or text [36, 38, 60, 62, 82, 92]. They have introduced varied strategies to foster 3D-consistent multi-view generation. The resulting multi-view images can then serve for 3D reconstruction, utilizing either optimization-based methods [36, 38, 62] or feed-forward models [28, 32, 34].

While most existing works focus on single-image-to-3D or text-to-3D, they often hallucinate regions that are invisible in the input image, which provides users with limited control over those areas. In this paper, we seek to broaden the input to encompass unposed sparse views and address both the 3D reconstruction and pose estimation challenges in a time-efficient way—within tens of seconds.

3 Method

Given n unposed input images $\{\mathbf{I}_i \mid i = 1, \ldots, n;\ 1 \leq n \leq 6\}$, which illustrate a single object from arbitrary categories, we predict their relative camera poses $\boldsymbol{\xi}_{ij}$ and reconstruct the 3D model $\mathcal{M}$ of the object. As illustrated in Fig. 2, we first finetune a 2D diffusion model [53] to process the unposed sparse input images (Sec. 3.1). The 2D diffusion model is responsible for jointly generating grid images for both the NOCS maps of the input views, as well as multi-view images with known camera poses. We use the predicted NOCS maps to estimate the camera poses for the input views (Sec. 3.2). The resulting multi-view images are fed into a two-stage 3D diffusion model for a coarse-to-fine generation of a 3D textured mesh (Sec. 3.3). This joint training strategy allows the two branches to complement each other. It enhances the understanding of both the input

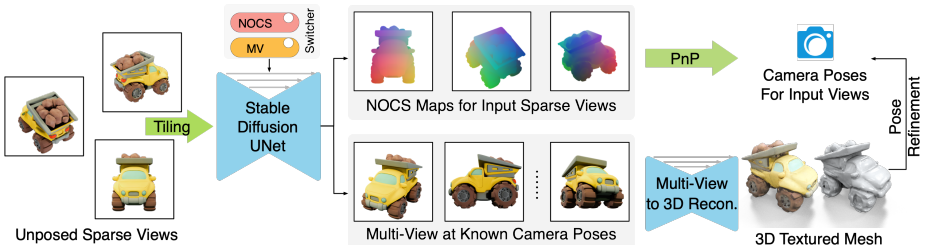

Fig. 2: Pipeline Overview of SpaRP. We begin by taking a sparse set of unposed images as input, which we tile into a single composite image. This composite image is subsequently provided to the Stable Diffusion UNet to serve as the conditioning input. The 2D diffusion model is simultaneously finetuned to predict NOCS maps for the input sparse views and multi-view images under known camera poses. From the NOCS maps, we extract the camera poses corresponding to the input views. The multi-view images are then processed by a reconstruction module to generate textured 3D meshes. Optionally, the camera poses can be further refined using the generated mesh for improved accuracy.

sparse views and the intrinsic properties of the 3D objects, thereby improving the performance of both pose estimation and 3D reconstruction. Optionally, the generated 3D mesh can also be used to further refine the camera poses (Sec. 3.4).

3.1 Tiling Sparse View Images as Input Condition

Recently, numerous studies have shown that 2D diffusion models not only possess robust open-world capabilities but also learn rich 3D geometric priors. For instance, Stable Diffusion [53], can be finetuned to include camera view control [28, 35, 36, 38, 60, 62], enabling it to predict novel views of objects—a task that necessitates significant 3D spatial reasoning. Consequently, we are inspired to utilize the rich priors inherent in 2D diffusion models for the tasks of sparse view 3D reconstruction and pose estimation.

Unlike most existing approaches that use a single RGB image as the condition and focus on synthesizing multi-view images, our goal is to take a sparse set of input images and stimulate Stable Diffusion to infer the spatial relationships among the input views implicitly. To accomplish this, given $1 \sim 6$ sparse views from arbitrary camera poses, we tile them into a 3×2 multi-view grid, as illustrated in Fig. 3 (c). The image in the first grid cell determines a canonical frame (to be discussed later), while the order of the other views is inconsequential. When there are fewer than 6 sparse views, we use empty padding for the remaining grid cells. This composite image then serves as the condition for Stable Diffusion, which is expected to assimilate all information from the input sparse views during the diffusion process.

We employ Stable Diffusion 2.1 as our base model. To adapt the original text-conditioning to our tiled multi-view image condition, we follow [60] to apply both local and global conditioning strategies. For local conditioning, we use the reference-only attention mechanism [96], where we process the reference tiled image with the denoising UNet model and append the attention keys and values

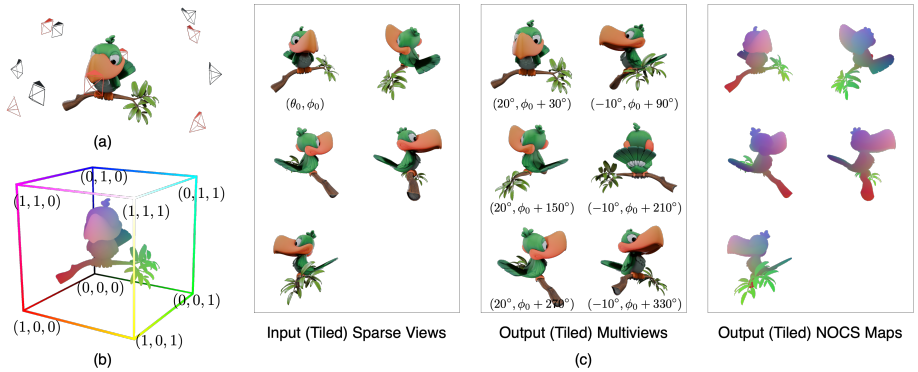

Fig. 3: (a) Regardless of the poses of the sparse input views (in black), the output multiviews are uniformly distributed (in red) and encompass the entire 3D object. (b) The Normalized Object Coordinate Space (NOCS) of the object, whose orientation is aligned with the azimuth of the first input view. (c) An example of input and output tiled images. The elevation and azimuth of the first input view are denoted by θ_0 and ϕ_0, respectively. The camera poses of the output multiview images are determined by ϕ_0. The output NOCS maps correspond to the input sparse views, and the orientation of the coordinate frame is also determined by ϕ_0.

from this image to corresponding layers in the denoising model for the target images. This mechanism facilitates implicit yet effective interactions between the diffusion model and the sparse views. For global conditioning, we integrate the mean-pooled CLIP embedding of all input images—modulated by learnable token weights—into the diffusion process, enhancing the model's ability to grasp the overarching semantics and structure of the sparse views.

As depicted in Figs. 2 and 3, our objective is to concurrently generate grid images for both NOCS maps of the input views and multi-view images from known camera poses. To achieve this, we utilize a domain switcher [38] that enables flexible toggling between the two domains. The switcher consists of two learnable embeddings, one for each domain, which are then injected into the UNet of the stable diffusion models by being added to its time embedding.

3.2 Image-to-NOCS Diffusion as a Pose Estimator

Conventional Structure-from-Motion (SfM) solvers, such as COLMAP [56], rely on feature matching for pose estimation. However, in scenarios with sparse views, there may be little to no overlap between input views. The lack of sufficient visual correspondence cues often renders the solvers unreliable and prone to failure. Consequently, instead of relying on local correspondences, we leverage the rich semantic priors embedded in 2D diffusion models for pose estimation.

One of the primary challenges is to enable 2D diffusion models to output camera poses. While camera poses can be represented in various scalar formats (e.g., 6-dimensional vector, four-by-four matrix, etc.), they are not native representations for a 2D diffusion model to generate. Inspired by recent works demonstrating that 2D diffusion models can be used to predict normal maps [38]—a

domain different from natural images—we propose using a surrogate representation: the Normalized Object Coordinate Space (NOCS) [74]. We finetune Stable Diffusion to predict NOCS maps for each input view.

As depicted in Fig. 3(b), a NOCS frame is determined for each set of input sparse view images and the underlying 3D object. Specifically, the 3D shape is normalized into a unit cube, i.e., $x, y, z \in [0, 1]$. The shape's upward axis aligns with the dataset's inherent upward axis of the 3D object, typically the gravity axis. Predicting the object's forward-facing direction may be ambiguous, so we rotate the 3D shape in the NOCS frame to align its forward direction (zero azimuth) with that of the first input view, thus unambiguously establishing the NOCS frame. For each input view, we then render a NOCS map, where each 2D pixel (r,g,b) represents the corresponding 3D point's position (x,y,z) in the defined NOCS frame, as shown in Fig. 3(c). These NOCS maps align with the operational domain of 2D diffusion models, similar to the normal maps in previous work [38].

To facilitate interactions between NOCS maps from different views and generate more 3D-consistent NOCS maps, we tile all NOCS maps into a 3×2 grid image as the input condition (see Sec. 3.1), following the same tiling order and the empty padding convention. We finetune Stable Diffusion to generate these multi-view tiled NOCS maps, so the 2D diffusion model can attend to both the input sparse views and their NOCS maps during the diffusion process.

After generating the NOCS maps for the input sparse views, we employ a traditional Perspective-n-Point (PnP) solver [2] to compute the poses $\{\boldsymbol{\xi}_i^{\text{pnp}}\}$ from the NOCS frame to the camera frames of each input view by minimizing the reprojection error:

$$\boldsymbol{\xi}_i^{\text{pnp}} = \arg \min_{\boldsymbol{\xi}_i \in \text{SE}(3)} \sum_{j=1}^{m_i} \|\mathbf{p}_{i,j} - \text{proj}(\mathbf{q}_{i,j}, \boldsymbol{\xi}_i)\|_2^2, \tag{1}$$

where $\mathbf{p}_{i,j}$ represents the jth pixel's location in the ith NOCS map; $\mathbf{q}_{i,j}$ is the corresponding 3D point location in the NOCS frame; m_i is the number of pixels for the ith view, and proj is the perspective projection operation. Note that the PnP algorithm assumes known camera intrinsics and optimizes only for the camera extrinsics. A RANSAC scheme is applied during the PnP computation for outlier removal, enhancing the robustness of the pose prediction to boundary noises and errors from the 2D diffusion model. As all NOCS maps share a common NOCS frame, we can thus determine the relative camera poses between views i and i' through $\boldsymbol{\xi}_i^{-1}\boldsymbol{\xi}_{i'}$.

3.3 Multi-View Prediction for 3D Reconstruction

We follow the paradigm of recent single-image-to-3D methods [28,32] by initially generating multi-view images and subsequently using a feed-forward 3D reconstruction module to convert these images into a 3D representation. It is noteworthy that the input sparse views might not encompass the entire 3D objects, nor provide adequate information for 3D reconstruction. Therefore, we propose to predict multi-view images at uniformly distributed camera poses first, and then use these predicted images for 3D reconstruction.

Unlike traditional novel view synthesis [35], our approach employs a fixed camera configuration for target multi-views. As depicted in Fig. 3, our target multi-view images consist of six views with alternating $20°$ and $-10°$ elevations, and $60°$-spaced azimuths relative to the first input view. Although the elevation angles are set absolutely, the azimuth angles are relative to the azimuth of the first input sparse view to resolve the ambiguity in face-forwarding directions. Furthermore, We maintain consistent camera intrinsics across target views, independent of input views. These strategies mitigate challenges in predicting camera intrinsics and elevation during the 3D reconstruction process. Existing methods hindered by this issue may be sensitive to intrinsic variations and often depend on predicting [34] or requiring user-specified [36, 65, 66] input image elevations.

Similar to NOCS map prediction, we tile all six views into a 3×2 grid image and finetune Stable Diffusion to generate this tiled image. The 2D diffusion model, conditioned on the input sparse views, aims to incorporate all information from input views, deduce the underlying 3D objects, and predict the multi-view images at the predetermined camera poses. Although the predicted poses of input sparse views (Sec. 3.2) are not directly employed in the 3D reconstruction, the joint training of NOCS prediction and multi-view prediction branches implicitly complement each other and boost the performance of both tasks.

Upon generating the multi-view images at known camera poses, we utilize the multi-view to 3D reconstruction module proposed in [32] to lift these images to 3D. The reconstruction module adopts a two-stage coarse-to-fine approach, which involves initially extracting the 2D features of the generated multi-view images, aggregating them with the known camera poses, and constructing a 3D cost volume. This 3D cost volume acts as the condition for the 3D diffusion networks. In the coarse stage, a low-resolution 64^3 3D occupancy volume is produced. This is subsequently refined to yield a high-resolution 128^3 SDF (Signed Distance Field) volume with colors. Finally, a textured mesh is derived from the SDF volume employing the marching cubes algorithm.

3.4 Pose Refinement with Reconstructed 3D Model

In Section 3.2, we finetune diffusion models for NOCS map prediction and camera pose estimation. However, due to the hallucinatory and stochastic nature of diffusion models, unavoidable errors may exist. The generated 3D mesh $\mathcal{M}$, though not perfect, provides a multi-view consistent and explicit 3D structure. We can further refine the coarse poses predicted from the NOCS maps by leveraging the reconstructed 3D shape.

Pose Refinement via Differentiable Rendering. Starting with initial poses $\{\boldsymbol{\xi}_i^{\text{pnp}}\}$ extracted from the predicted NOCS maps, we refine them through differentiable rendering [26]. Specifically, we render the generated mesh $\mathcal{M}$ at optimizing camera poses $\boldsymbol{\xi}_i$. We minimize the rendering loss between the rendered image $\mathbf{I}_i^r = \mathcal{R}(\mathcal{M}, \boldsymbol{\xi}_i)$ and the input image $\mathbf{I}_i$ to obtain the optimally fitted camera pose $\boldsymbol{\xi}_i^*$. The optimization process can be formulated as:

$$\boldsymbol{\xi}_i^* = \arg \min_{\boldsymbol{\xi}_i \in \text{SE}(3)} (\lambda \cdot \mathcal{L}_{\text{mask}}(\mathbf{I}_i^r, \mathbf{I}_i) + \mu \cdot \mathcal{L}_{\text{rgb}}(\mathbf{I}_i^r, \mathbf{I}_i)), \tag{2}$$

Table 1: Evaluation Results for Pose Estimation. We compare our method with RelPose++ [29], FORGE [17], and iFusion [83] on three unseen datasets: OmniObject3D [85], GSO [12], and ABO [7]. 500 objects are sampled for each dataset.

Dataset	Method	Rot. Err↓	Acc.@15° ↑	Acc.@30° ↑	Trans. Err↓	Time↓
GSO [12]	RelPose++	103.24	0.011	0.033	4.84	3.6s
	FORGE	111.40	0.004	0.020	4.21	440s
	iFusion (n_{init}=1)	95.15	0.208	0.258	3.65	64s
	iFusion (n_{init}=4)	8.61	0.651	0.759	0.49	256s
	Ours (w/o refine)	13.02	0.537	0.616	0.58	10s
	Ours (n_{init}=1)	9.87	0.563	0.617	0.42	27s
	Ours (n_{init}=4)	**5.28**	**0.750**	**0.787**	**0.23**	57s
OO3D [85]	RelPose++	105.05	0.008	0.046	7.38	3.6s
	FORGE	99.27	0.014	0.063	7.27	440s
	iFusion (n_{init}=1)	91.15	0.166	0.271	4.77	64s
	iFusion (n_{init}=4)	15.08	0.498	0.721	1.12	256s
	Ours (w/o refine)	14.75	0.508	0.725	0.90	10s
	Ours (n_{init}=1)	13.40	0.544	0.730	0.89	27s
	Ours (n_{init}=4)	**10.07**	**0.668**	**0.849**	**0.63**	57s
ABO [7]	RelPose++	103.14	0.017	0.039	5.01	3.6s
	FORGE	110.64	0.005	0.023	4.18	440s
	iFusion (n_{init}=1)	96.65	0.186	0.219	3.88	64s
	iFusion (n_{init}=4)	8.55	0.578	0.631	0.68	256s
	Ours (w/o refine)	10.87	0.554	0.597	0.49	10s
	Ours (n_{init}=1)	9.30	0.565	0.600	0.43	27s
	Ours (n_{init}=4)	**5.80**	**0.675**	**0.701**	**0.27**	57s

where $\mathcal{L}_{\text{mask}}$ and $\mathcal{L}_{\text{rgb}}$ are the cross-entropy and MSE losses computed for the foreground masks and the RGB values, respectively, and λ and μ are two weighting coefficients. The refinement process is lightweight and can be completed in just one second, given the generated mesh.

Mixture of Experts (MoE). The NOCS pose predictions are inherently stochastic and may not produce an accurate pose in a single pass. For instance, with objects possessing certain symmetries, the diffusion model may predict only one of the possible symmetric poses. We employ a Mixture of Experts (MoE) strategy to further refine the pose, which is simple but effective. Specifically, we generate multiple NOCS maps for each input view using different seeds. We then select the pose that minimizes the rendering loss based on the refinement results with the generated 3D mesh. This technique effectively reduces pose estimation error, as quantitatively validated by the ablation study in the supplementary materials.

4 Experiments

4.1 Evaluation Settings

Training Datasets and Details. We train our models on a curated subset of 100k shapes from the Objaverse dataset [9]. Considering the variable quality of the original Objaverse dataset, we opted to filter out higher-quality data by initially manually annotating 8,000 3D objects based on overall geometry quality and texture preferences. Subsequently, we train MLP models for quality rating classification and texture score regression, utilizing their multimodal features [33]. Based on the predictions of these models, we select shapes that are rated as high-quality and have top texture scores. Further details about the data filtering process are included in the supplementary materials.

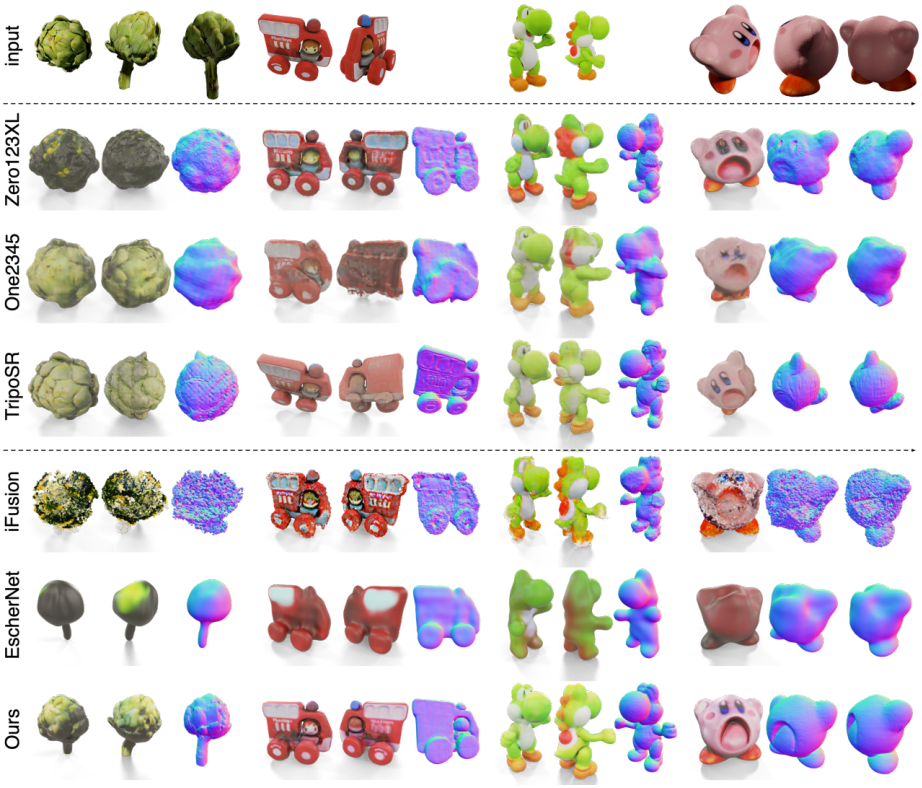

Fig. 4: Qualitative Results on 3D Reconstruction. Zero123XL [8], One2345 [34], and TripoSR [69] are single-image-to-3D methods, each utilizing only the first input image. iFusion [83], EscherNet [23], and our approach take all input images (the first row). Textured meshes and mesh normal renderings are shown. Shapes come from the OmniObject3D [85] and GSO [12] datasets.

For each 3D shape, we render 10 sets of images using BlenderProc [11]. Each set comprises 6 input images, 6 output multi-view images, and 6 NOCS maps. To mimic real-world conditions and ensure model robustness, we randomly sample camera intrinsics and extrinsics, as well as environment maps for the input images. For the output multi-view images, their intrinsics remain constant, while the extrinsics are derived from a fixed delta pose and the azimuth of the input images. Each set of input and output images shares the same environment map. During training, we randomly selected between 1 to 6 views as sparse input views, with the first view of each set always being included. We train the model utilizing 8 A100 GPUs for approximately 3 days.

Baselines. For 3D reconstruction, we compare our method with both state-of-the-art single-image-to-3D and sparse-view-to-3D baselines. Single-view-to-3D methods we evaluate include optimization-based approaches, such as Zero123 XL [8], SyncDreamer [36], and DreamGaussian [66], as well as feed-forward methods like One-2-3-45 [34] and Shap-E [20]. For sparse-view methods, we consider two recent open-source approaches as baselines: iFusion [83] and EscherNet [23].

Table 2: Quantitative Comparison on 3D Reconstruction. Evaluated on the complete GSO [12] dataset, which contains 1,030 3D objects. Five single-image-to-3D methods [8, 20, 34, 36, 66] and two sparse-view methods [23, 83] are compared.

n_{input}	Method	F-Score (%)↑	CLIP-Sim↑	PSNR↑	LPIPS↓	Time↓
1	Zero123 XL [8]	91.6	73.1	18.16	0.136	20min
	Shap-E [20]	91.8	73.1	18.96	0.140	27s
	One-2-3-45 [34]	90.4	70.8	19.07	0.133	45s
	SyncDreamer [36]	84.8	68.9	16.86	0.145	6min
	DreamGaussian [66]	81.0	68.4	17.88	0.147	2min
	Ours	**95.7**	**78.2**	**19.87**	**0.124**	**16s**
6	iFusion [83]	88.5	66.7	16.2	0.151	28min
	EscherNet [23]	94.8	65.9	16.6	0.139	9min
	Ours	**96.9**	**78.1**	**19.3**	**0.123**	**16s**

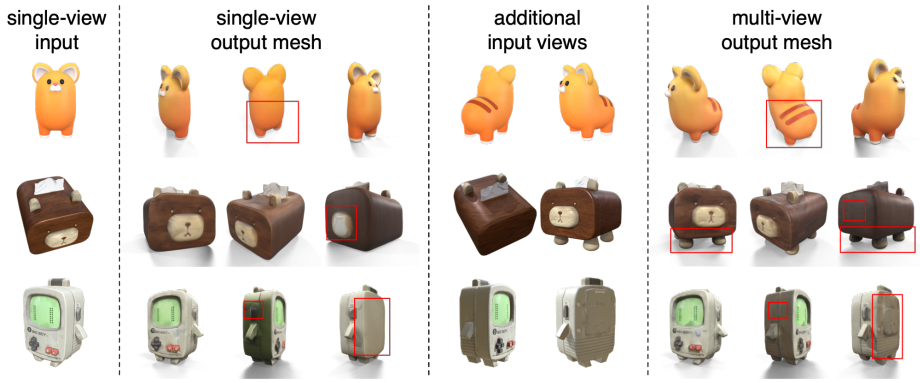

Fig. 5: Single-View vs. Sparse-View for 3D Reconstruction. We compare the results of our method when using single-view and sparse-view inputs.

We utilize ThreeStudio [13]'s implementation for Zero123 XL and the official implementations for the other baselines. Specifically, for iFusion, we use their official reconstruction pipeline integrated with Zero123 XL.

For sparse-view pose estimation, we compare our method with state-of-the-art approaches including RelPose++ [29], FORGE [17], and iFusion [83]. The latter two are optimization-based while [29] is a feed-forward method.

Evaluation Datasets. For 3D reconstruction, we evaluate the methods on the entire GSO [12] dataset, which comprises 1,030 3D shapes; none of these shapes were seen during our training. For each 3D shape, we randomly render six views as input images. For single-image-to-3D methods, a fixed-view image is taken as input following [34]. We carefully align the predictions with the ground truth meshes before calculating the metrics. Please refer to the supplementary materials for detailed information on shape alignment and the evaluation metrics.

For pose estimation, we evaluate the approaches on three datasets: OmniObject3D [85] and GSO [12], both captured from real scans, and ABO [7], a synthetic dataset created by artists. For each dataset, we randomly choose 500 objects and render five random sparse views per shape. We follow iFusion [83] to report the rotation accuracy and the median error in rotation and translation across all image pairs. More details are provided in the supplementary materials.

Table 3: Ablation Study on the Number of Input Views. Evaluated on the GSO dataset [12].

n views	Rot. Err↓	Acc.@5° ↑	Trans. Err↓	F-Score (%)↑	CLIP-Sim↑	PSNR↑
1	–	–	–	89.1	74.9	17.7
2	8.56	0.32	0.42	93.3	76.5	18.3
4	6.03	0.43	0.28	96.0	77.6	19.0
6	**5.28**	**0.48**	**0.25**	**96.9**	**78.1**	**19.3**

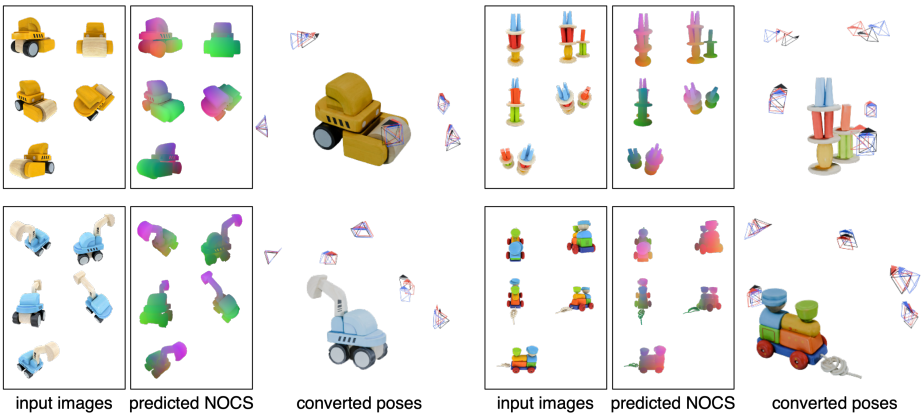

Fig. 6: Ablation Study on Pose Refinement. We showcase the input images, predicted NOCS maps, and converted poses. The ground truth poses are in black, while the predicted poses before and after refinement are in blue and red, respectively.

4.2 Experiment Results

Pose Prediction. We report the pose estimation results in Tab. 1, where it is evident that SpaRP outperforms all baseline methods by a significant margin. It is worth noting that RelPose++ [29] and FORGE [17] struggle to yield satisfactory results for our open-world evaluation images. iFusion, an optimization-based approach, is prone to becoming trapped in local minima. With only one initial pose ($n_{init} = 1$), it also fails to produce adequate results. In contrast, our method leverages priors from 2D diffusion models and can generate acceptable results in a single forward pass. Even without any additional refinement (w/o refine), our method can already produce results similar to iFusion with four initial poses ($n_{init} = 4$), while being far more efficient, requiring just 1/25 of the runtime. With the integration of further refinement through a mixture of experts, our method achieves even better performance.

3D Reconstruction. We present the qualitative results in Fig. 4. With only a single-view input, single-image-to-3D methods fail to produce meshes that faithfully match the entire structure and details of the ground truth mesh. For instance, most single-view baseline methods are unable to reconstruct the stems of the artichoke, the back of the firetruck, the red saddle on Yoshi, and the two separate legs of Kirby standing on the ground. In contrast, sparse-view methods yield results that are much closer to the ground truth by incorporating information from multiple sparse views. Compared to iFusion, EscherNet, our method generates meshes with higher-quality geometry and textures that more

Table 4: Effect of Joint Training. Evaluated on 500 objects from GSO [12].

Method	Rot. Err↓	Acc.@15° ↑	Acc.@30° ↑	Trans. Err↓	F-Score (%)↑	CLIP-Sim↑	PSNR↑
Separate	9.87	0.563	0.617	0.42	96.81	78.36	19.03
Joint	**8.57**	**0.601**	**0.677**	**0.37**	**97.12**	**78.90**	**19.42**

accurately match the input sparse views. We report the quantitative results in Tab. 2, where our method significantly outperforms both single-view-to-3D and sparse-view approaches in terms of both 2D and 3D metrics. Moreover, our method exhibits superior efficiency, being much faster than the baseline methods.

4.3 Analysis

Single View vs. Sparse Views. In Fig. 5, we present the results obtained by our method when provided with single-view and sparse-view inputs. With a single-view input, our method can still generate reasonable results, yet it may not accurately capture the structures and details of the regions that are not visible. Our method demonstrates the capability to effectively integrate information from all sparse-view inputs provided.

Number of Views. In Tab. 3, we quantitatively showcase the impact of the number of views on both 3D reconstruction and pose estimation. We observe that incorporating more input views enables the 2D diffusion network to better grasp their spatial relationships and underlying 3D objects, boosting both tasks.

Pose Refinement. While the predicted NOCS maps can be directly converted into camera poses, we have found that these poses can be further refined through alignment with the generated 3D meshes. Fig. 6 showcases the predicted poses before and after refinement. Although both are generally very close to the ground truth poses, refinement can further reduce the error.

Number of Experts. We employ a mixture-of-experts strategy to address the ambiguity issues related to NOCS prediction for symmetric objects. By using this strategy and increasing the number of experts, there is a substantial increase in pose estimation accuracy. Please refer to the supplementary material for quantitative ablation studies.

Joint Training. We finetune 2D diffusion models to jointly predict NOCS maps and multi-view images from sparse, unposed views by leveraging a domain switcher. As shown in Tab. 4, this joint training strategy enables the two branches to implicitly interact and complement each other, enhancing the interpretation of both the input sparse views and the intrinsic properties of the 3D objects, which in turn improves the performance of each task.

5 Conclusion

We present SpaRP, a novel method for 3D reconstruction and pose estimation using unposed sparse-view images. Our method leverages rich priors embedded in 2D diffusion models and exhibits strong open-world generalizability. Without the need for per-shape optimization, it can deliver high-quality textured meshes, along with accurate camera poses, in approximately 20 seconds.

Acknowledgements: We thank Chong Zeng, Xinyue Wei for the discussion and help with data processing, and Peng Wang for providing the evaluation set. We also extend our thanks to all annotators for their meticulous annotations.

References

1. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
2. Bradski, G.: Perspective-n-point (pnp) pose computation (the opencv library). `https://docs.opencv.org/4.x/d5/d1f/calib3d_solvePnP.html` (2000)
3. Chan, E.R., Nagano, K., Chan, M.A., Bergman, A.W., Park, J.J., Levy, A., Aittala, M., De Mello, S., Karras, T., Wetzstein, G.: Genvs: Generative novel view synthesis with 3d-aware diffusion models (2023)
4. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14124–14133 (2021)
5. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873 (2023)
6. Chen, Z., Wang, F., Liu, H.: Text-to-3d using gaussian splatting. arXiv preprint arXiv:2309.16585 (2023)
7. Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., et al.: Abo: Dataset and benchmarks for real-world 3d object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21126–21136 (2022)
8. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663 (2023)
9. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
10. Deng, C., Jiang, C., Qi, C.R., Yan, X., Zhou, Y., Guibas, L., Anguelov, D., et al.: Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20637–20647 (2023)
11. Denninger, M., Sundermeyer, M., Winkelbauer, D., Zidan, Y., Olefir, D., Elbadrawy, M., Lodhi, A., Katam, H.: Blenderproc. arXiv preprint arXiv:1911.01911 (2019)
12. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022)
13. Guo, Y.C., Liu, Y.T., Shao, R., Laforte, C., Voleti, V., Luo, G., Chen, C.H., Zou, Z.X., Wang, C., Cao, Y.P., Zhang, S.H.: threestudio: A unified framework for 3d content generation. `https://github.com/threestudio-project/threestudio` (2023)

14. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. arXiv preprint arXiv:2311.04400 (2023)
15. Jain, A., Mildenhall, B., Barron, J.T., Abbeel, P., Poole, B.: Zero-shot text-guided object generation with dream fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 867–876 (2022)
16. Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5885–5894 (2021)
17. Jiang, H., Jiang, Z., Grauman, K., Zhu, Y.: Few-view object reconstruction with unknown categories and camera poses. arXiv preprint arXiv:2212.04492 (2022)
18. Jiang, H., Jiang, Z., Zhao, Y., Huang, Q.: Leap: Liberate sparse-view 3d modeling from camera poses. arXiv preprint arXiv:2310.01410 (2023)
19. Johari, M.M., Lepoittevin, Y., Fleuret, F.: Geonerf: Generalizing nerf with geometry priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18365–18375 (2022)
20. Jun, H., Nichol, A.: Shap-e: Generating conditional 3d implicit functions. arXiv preprint arXiv:2305.02463 (2023)
21. Karnewar, A., Vedaldi, A., Novotny, D., Mitra, N.J.: Holodiffusion: Training a 3d diffusion model using 2d images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18423–18433 (2023)
22. Kim, M., Seo, S., Han, B.: Infonerf: Ray entropy minimization for few-shot neural volume rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12912–12921 (2022)
23. Kong, X., Liu, S., Lyu, X., Taher, M., Qi, X., Davison, A.J.: Eschernet: A generative model for scalable view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9503–9513 (2024)
24. Kulhánek, J., Derner, E., Sattler, T., Babuška, R.: Viewformer: Nerf-free neural rendering from few images using transformers. In: European Conference on Computer Vision. pp. 198–216. Springer (2022)
25. Lai, Z., Liu, S., Efros, A.A., Wang, X.: Video autoencoder: self-supervised disentanglement of static 3d structure and motion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9730–9740 (2021)
26. Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., Aila, T.: Modular primitives for high-performance differentiable rendering. ACM Transactions on Graphics **39**(6) (2020)
27. Lee, H.H., Chang, A.X.: Understanding pure clip guidance for voxel grid nerf models. arXiv preprint arXiv:2209.15172 (2022)
28. Li, J., Tan, H., Zhang, K., Xu, Z., Luan, F., Xu, Y., Hong, Y., Sunkavalli, K., Shakhnarovich, G., Bi, S.: Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. arXiv preprint arXiv:2311.06214 (2023)
29. Lin, A., Zhang, J.Y., Ramanan, D., Tulsiani, S.: Relpose++: Recovering 6d poses from sparse-view observations. arXiv preprint arXiv:2305.04926 (2023)
30. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 300–309 (2023)
31. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5741–5751 (2021)

32. Liu, M., Shi, R., Chen, L., Zhang, Z., Xu, C., Wei, X., Chen, H., Zeng, C., Gu, J., Su, H.: One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10072–10083 (2024)
33. Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., Su, H.: Openshape: Scaling up 3d shape representation towards open-world understanding. Advances in Neural Information Processing Systems **36** (2024)
34. Liu, M., Xu, C., Jin, H., Chen, L., Varma T, M., Xu, Z., Su, H.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems **36** (2024)
35. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023)
36. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)
37. Liu, Y., Peng, S., Liu, L., Wang, Q., Wang, P., Theobalt, C., Zhou, X., Wang, W.: Neural rays for occlusion-aware image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7824–7833 (2022)
38. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9970–9980 (2024)
39. Long, X., Lin, C., Wang, P., Komura, T., Wang, W.: Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In: European Conference on Computer Vision. pp. 210–227. Springer (2022)
40. Melas-Kyriazi, L., Laina, I., Rupprecht, C., Vedaldi, A.: Realfusion: 360deg reconstruction of any object from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8446–8455 (2023)
41. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12663–12673 (2023)
42. Michel, O., Bar-On, R., Liu, R., Benaim, S., Hanocka, R.: Text2mesh: Text-driven neural stylization for meshes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13492–13502 (2022)
43. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
44. Mohammad Khalid, N., Xie, T., Belilovsky, E., Popa, T.: Clip-mesh: Generating textured meshes from text using pretrained image-text models. In: SIGGRAPH Asia 2022 conference papers. pp. 1–8 (2022)
45. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022)
46. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)

47. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., et al.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. arXiv preprint arXiv:2306.17843 (2023)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
49. Raj, A., Kaza, S., Poole, B., Niemeyer, M., Ruiz, N., Mildenhall, B., Zada, S., Aberman, K., Rubinstein, M., Barron, J., et al.: Dreambooth3d: Subject-driven text-to-3d generation. arXiv preprint arXiv:2303.13508 (2023)
50. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
51. Rematas, K., Martin-Brualla, R., Ferrari, V.: Sharf: Shape-conditioned radiance fields from a single view. arXiv preprint arXiv:2102.08860 (2021)
52. Ren, Y., Zhang, T., Pollefeys, M., Süsstrunk, S., Wang, F.: Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16685–16695 (2023)
53. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
54. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)
55. Sajjadi, M.S., Meyer, H., Pot, E., Bergmann, U., Greff, K., Radwan, N., Vora, S., Lučić, M., Duckworth, D., Dosovitskiy, A., et al.: Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6229–6238 (2022)
56. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
57. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: 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. pp. 501–518. Springer (2016)
58. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
59. Seo, J., Jang, W., Kwak, M.S., Ko, J., Kim, H., Kim, J., Kim, J.H., Lee, J., Kim, S.: Let 2d diffusion model know 3d-consistency for robust text-to-3d generation. arXiv preprint arXiv:2303.07937 (2023)
60. Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., Chen, L., Zeng, C., Su, H.: Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv:2310.15110 (2023)
61. Shi, R., Wei, X., Wang, C., Su, H.: Zerorf: Fast sparse view 360 {\deg} reconstruction with zero pretraining. arXiv preprint arXiv:2312.09249 (2023)
62. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023)

63. Sinha, S., Zhang, J.Y., Tagliasacchi, A., Gilitschenski, I., Lindell, D.B.: Sparsepose: Sparse-view camera pose regression and refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21349–21359 (2023)
64. Stereopsis, R.M.: Accurate, dense, and robust multiview stereopsis. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE **32**(8) (2010)
65. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. arXiv preprint arXiv:2402.05054 (2024)
66. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
67. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. arXiv preprint arXiv:2303.14184 (2023)
68. Tewari, A., Yin, T., Cazenavette, G., Rezchikov, S., Tenenbaum, J.B., Durand, F., Freeman, W.T., Sitzmann, V.: Diffusion with forward models: Solving stochastic inverse problems without direct supervision. arXiv preprint arXiv:2306.11719 (2023)
69. Tochilkin, D., Pankratz, D., Liu, Z., Huang, Z., Letts, A., Li, Y., Liang, D., Laforte, C., Jampani, V., Cao, Y.P.: Triposr: Fast 3d object reconstruction from a single image (2024)
70. Trevithick, A., Yang, B.: Grf: Learning a general radiance field for 3d representation and rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15182–15192 (2021)
71. Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: Sparf: Neural radiance fields from sparse and noisy poses. In: CVF Conference on Computer Vision and Pattern Recognition, CVPR. vol. 1 (2023)
72. Tung, H.Y.F., Cheng, R., Fragkiadaki, K.: Learning spatial common sense with geometry-aware recurrent networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2595–2603 (2019)
73. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12619–12629 (2023)
74. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2642–2651 (2019)
75. Wang, J., Rupprecht, C., Novotny, D.: Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9773–9783 (2023)
76. Wang, P., Chen, X., Chen, T., Venugopalan, S., Wang, Z., et al.: Is attention all nerf needs? arXiv preprint arXiv:2207.13298 (2022)
77. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)
78. Wang, P., Tan, H., Bi, S., Xu, Y., Luan, F., Sunkavalli, K., Wang, W., Xu, Z., Zhang, K.: Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. arXiv preprint arXiv:2311.12024 (2023)

79. Wang, Q., Wang, Z., Genova, K., Srinivasan, P.P., Zhou, H., Barron, J.T., Martin-Brualla, R., Snavely, N., Funkhouser, T.: Ibrnet: Learning multi-view image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2021)
80. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. arXiv preprint arXiv:2305.16213 (2023)
81. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: Nerf–: Neural radiance fields without known camera parameters. arXiv preprint arXiv:2102.07064 (2021)
82. Weng, H., Yang, T., Wang, J., Li, Y., Zhang, T., Chen, C., Zhang, L.: Consistent123: Improve consistency for one image to 3d object synthesis. arXiv preprint arXiv:2310.08092 (2023)
83. Wu, C.H., Chen, Y.C., Solarte, B., Yuan, L., Sun, M.: ifusion: Inverting diffusion for pose-free reconstruction from sparse views (2023)
84. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21551–21561 (2024)
85. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 803–814 (2023)
86. Xia, Y., Tang, H., Timofte, R., Van Gool, L.: Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction. arXiv preprint arXiv:2210.04553 (2022)
87. Xu, D., Jiang, Y., Wang, P., Fan, Z., Wang, Y., Wang, Z.: Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360deg views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4479–4489 (2023)
88. Xu, J., Wang, X., Cheng, W., Cao, Y.P., Shan, Y., Qie, X., Gao, S.: Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20908–20918 (2023)
89. Yang, H., Hong, L., Li, A., Hu, T., Li, Z., Lee, G.H., Wang, L.: Contranerf: Generalizable neural radiance fields for synthetic-to-real novel view synthesis via contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16508–16517 (2023)
90. Yang, Z., Ren, Z., Bautista, M.A., Zhang, Z., Shan, Q., Huang, Q.: Fvor: Robust joint shape and pose optimization for few-view object reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2497–2507 (2022)
91. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. Advances in Neural Information Processing Systems **34**, 4805–4815 (2021)
92. Ye, J., Wang, P., Li, K., Shi, Y., Wang, H.: Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. arXiv preprint arXiv:2310.03020 (2023)
93. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)

94. Yu, C., Zhou, Q., Li, J., Zhang, Z., Wang, Z., Wang, F.: Points-to-3d: Bridging the gap between sparse points and shape-controllable text-to-3d generation. arXiv preprint arXiv:2307.13908 (2023)
95. Zhang, J.Y., Ramanan, D., Tulsiani, S.: Relpose: Predicting probabilistic relative rotation for single objects in the wild. In: European Conference on Computer Vision. pp. 592–611. Springer (2022)
96. Zhang, L.: Reference-only control. `https://github.com/Mikubill/sd-webui-controlnet/discussions/1236` (2023)
97. Zhou, Z., Tulsiani, S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12588–12597 (2023)