iFusion: Inverting Diffusion for Pose-Free Reconstruction from Sparse Views

Chin-Hsuan Wu*[1] Yen-Chun Chen[2] Bolivar Solarte[1] Lu Yuan[2] Min Sun[1,3]
[1]National Tsing Hua University [2]Microsoft [3]Amazon
chinhsuanwu.github.io/ifusion

Figure 1. **Demonstration on real-world 3D reconstruction.** With only two casually taken photos **without camera poses**, *iFusion* can reconstruct plausible 3D assets. The top row example is taken from DreamBooth3D [52], and we took photos for the cat statue by ourselves.

Abstract

We present iFusion, a novel 3D object reconstruction framework that requires only two views with unknown camera poses. While single-view reconstruction yields visually appealing results, it can deviate significantly from the actual object, especially on unseen sides. Additional views improve reconstruction fidelity but necessitate known camera poses. However, assuming the availability of pose may be unrealistic, and existing pose estimators fail in sparse-view scenarios. To address this, we harness a pre-trained novel view synthesis diffusion model, which embeds implicit knowledge about the geometry and appearance of diverse objects. Our strategy unfolds in three steps: (1) We invert the diffusion model for camera pose estimation instead of synthesizing novel views. (2) The diffusion model is fine-tuned using provided views and estimated poses, turned into a novel view synthesizer tailored for the target object. (3) Leveraging registered views and the fine-tuned diffusion model, we reconstruct the 3D object. Experiments demonstrate strong performance in both pose estimation and novel view synthesis. Moreover, iFusion seamlessly integrates with various reconstruction methods and enhances them.

1. Introduction

Reconstructing objects from sparse views poses a significant challenge yet holds paramount importance for various applications, including 3D content creation, augmented reality, virtual reality, and robotics. Recent breakthroughs, guided by pre-trained models, have facilitated visually plausible reconstructions from a single view, without requiring the camera pose [32, 33, 38, 49, 66, 67, 82]. However, the reconstructed assets might not precisely capture the actual objects due to the inherent single-view ambiguity, *e.g.*, the object's side opposite to the camera can only be imagined. Furthermore, multiple potential 3D structures could correspond to the same input image.

On the other hand, sparse-view methods assume the availability of an accurate camera pose for each view [3, 19, 35, 65, 90]. To meet this requirement, a Structure-from-Motion (SfM) pre-processing, *e.g.*, COLMAP [58], is typically employed. Paradoxically, these methods demand a substantial number of images, usually more than 50 in practice, for reliable pose estimation. Recent learning-based pose estimation [28, 61, 85, 86] and pose-free reconstruction [21, 22] have sought to alleviate this issue. However, they still require a minimum of five input views to achieve favorable results and are primarily demonstrated on objects with simple geometry or within a constrained set of object categories. A generic framework for pose-free, sparse-view 3D reconstruction is still lacking, posing a significant obstacle to real-world applications with casually captured photos. We hereby raise the research question: How can one ensure the *reconstruction fidelity* of diverse objects using *extremely sparse* views *without camera poses*?

The key is a sparse-view pose estimator. Our motiva-

*Part of this work was done as a research intern at Microsoft.

tion stems from a recent novel view synthesis diffusion model, namely Zero123 [33], which is pre-trained on the most extensive 3D object dataset to date [8]. Given a reference view image, Zero123 can generate a novel view (query view) from a specified pose (Fig. 2, left). We thus hypothesize that Zero123 can be effectively used for pose estimation, with an intuition that a well-estimated pose fed into Zero123 will produce an image similar to the query view. Conversely, if the query view is provided as input, the model should be able to infer the pose by generating an image that best matches the query view. This strategy shares a similar concept with Textual Inversion [14], which finds the token that generates the image through Text-to-Image models. In our case, we recover the camera pose that generates the viewpoint through Zero123. Following this idea, we repurpose Zero123 by inverting it to take the two views and estimate the relative camera transformation (Fig. 2, right). More specifically, we adopt an analysis-by-synthesis paradigm [7, 47, 81] that optimizes the transformation by minimizing the difference between the denoised latent visual features, *i.e.*, Zero123's output image feature map, and the query view's feature. Empirically, the proposed approach achieves strong pose estimation with as few as 2 views, even outperforming existing approaches' results with 5 views.

Well-estimated poses also open up a new opportunity. Using the given views registered with poses, a mini-dataset can be constructed to further fine-tune Zero123 and customize the diffusion model for synthesizing the target object's novel views. Specifically, we can form a set of (reference view, camera pose, query view) triplets from the given sparse views and fine-tune Zero123. To accelerate training and prevent overfitting, we use Low-Rank Adaptaion (LoRA) [17] to fine-tune the diffusion model, a recognized technique for customizing diffusion models.[1] Experiments demonstrate that this step significantly improves novel view synthesis, achieving an average increase of **+3.6** in PSNR across two datasets, and is beneficial to the final reconstruction. Note that our approach shares a similar spirit with test-time training [64], test-time adaptation [68], and self-training [59, 77]. Like test-time training and adaptation, we align the model to the test distribution based on test inputs (given views) but without test labels (novel views). Analogous to self-training, we synthesize additional labels (camera poses) using the learning model itself. To the best of our knowledge, the above combination is new for diffusion-based 3D reconstruction.

To this end, we introduce *iFusion*, a novel framework that reconstructs diverse 3D objects with sparse, pose-free views. First, the pose estimation is achieved by **i**nverting the Zero123 dif**Fusion** model, as described earlier. With the estimated camera pose, an object-specific improvement on Zero123's novel view synthesis capability is performed, which can be further utilized as additional reconstruction guidance. Finally, for reconstructing the 3D asset, any differentiable renderer can be plugged in, including NeRFs [40] and the recently proposed 3D Gaussian Splatting [25]. It is noteworthy that our framework does not assume any specific reconstruction pipeline, and experimental results demonstrate that *iFusion* is readily applicable to four different single-view reconstruction methods. Improved geometric fidelity is observed with a significant **+7.2%** increase in volume IoU, showcasing the necessity of additional views for reliable 3D reconstruction.

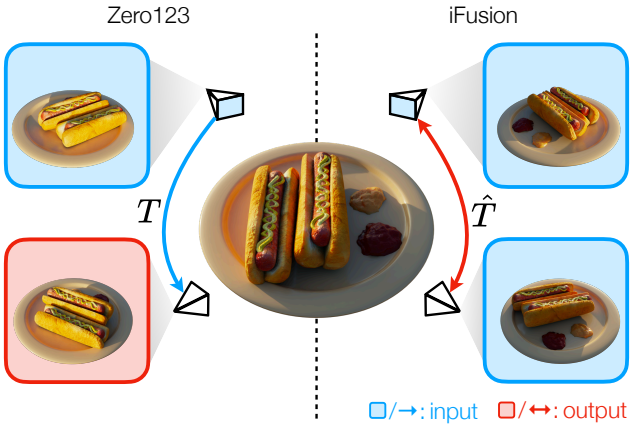

Figure 2. ***Zero123 vs. iFusion.*** Unlike Zero123 [33] (left), which synthesizes an object's novel view given an image and a transformation T, *iFusion* (right) instead optimizes an unknown relative transformation $\hat{T}$ from two given views.

Our contributions are summarized as follows:

1. We propose a novel camera pose estimator that significantly outperforms existing methods in terms of both accuracy and required number of input views, while being effective for diverse objects.
2. A self-training and test-time training inspired fine-tuning stage is innovated. This stage results in a much stronger novel view synthesis diffusion model, which plays a crucial role in guiding the reconstruction process.
3. For the first time, we escalate diffusion-based single-view reconstruction to multi-view for enhanced fidelity with merely two pose-free images.

2. Preliminary

iFusion repurposes a novel view synthesizing diffusion model for camera pose prediction. To prepare readers with the necessary backgrounds, we briefly introduce the basics of Diffusion Models (DM) and how they can be used for novel view synthesis. Next, we summarize a recently popular approach to utilize DM for 3D reconstruction, which we integrate into *iFusion* to allow reconstruction.

[1] https://github.com/cloneofsimo/lora

Diffusion Models. Diffusion models [16, 62, 63] are a class of deep generative models that has become the mainstream approach for high-fidelity visual synthesis. In image generation, they work by "diffusing" an image by adding noise over repeated steps, and then a deep neural network is trained to predict the applied step-wise noise from a corrupted image. This allows the reversion of the diffusion process, thus an image can be generated from a random noise by iterative denoising using the trained noise predicting network. More specifically, Ho et al. [16] formulated the diffusion process in the following analytical form:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t}\epsilon, \quad t \in [0, 1, \dots, \mathcal{T}], \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ denotes the Gaussian noise and hyper-parameter α_t denotes the noise schedule. For the reverse process, the noise predictor is denoted as $\epsilon_\theta(x_t, t)$, where θ is the set of trainable parameters. Instead of directly modeling the RGB pixel values x, a widely used diffusion model, Stable Diffusion (SD),[2] applies the Latent Diffusion Model (LDM) [54] to model the latent feature maps z. The encoding and reconstruction of images is done via a pre-trained VQ-VAE: $z = \mathcal{E}(x)$, and $x = \mathcal{D}(z)$. Moreover, DM may optionally take conditional inputs c, *e.g.*, texts, bounding box layouts, and depth maps. For instance, the standalone SD takes texts as the condition c and enables text-to-image generation (T2I). Formally, the training loss of the prediction network can be written as:

$$\mathcal{L}(x, c) = \mathbb{E}_{z,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right], \tag{2}$$

where $\|\cdot\|_2$ denotes the L2 norm.

Diffusion Models for Novel View Synthesis. The original Stable Diffusion was trained on web-scale image-text pairs[3] for text-to-image generation. Recently, Liu et al. [33] proposed Zero123 to further fine-tune SD on Objaverse [8], a large-scale 3D assets dataset, for object-centric novel view synthesis. Given an image at the reference viewpoint x^r and the **r**eference-to-**q**uery **t**ransformation $T_{r\to q} \in \mathrm{SE}(3)$, the model synthesizes the desired query view x^q with condition $c(x^r, T_{r\to q})$. This is formulated as a DM and shares the same training objective as Eq. (2).

3D Reconstruction via Score Distillation Sampling. Recent studies [20, 41, 48, 69] indicated that large-scale pre-trained 2D vision models [51, 54, 56] implicitly encapsulate rich 3D geometric prior. Notably, DreamFusion [48] introduced the Score Distillation Sampling (SDS) to facilitate 3D generation guided by a pre-trained 2D DM. Let $x = \mathcal{R}_\psi(T)$ be the rendered image at viewpoint $T \in \mathrm{SE}(3)$, where $\mathcal{R}$ is a differentiable renderer parameterized by ψ, *e.g.*, Neural Radiance Fields (NeRFs) [40] or 3D Gaussian Splatting [25]. Given a denoising network ϵ_θ, SDS optimizes the renderer ψ by minimizing the residuals between the predicted noise and the added noise, thereby producing the gradients:

$$\nabla_\psi \mathcal{L}_{SDS}(x, c) = \mathbb{E}_{z,\epsilon,t}\left[(\epsilon_\theta(z_t, t, c) - \epsilon)\frac{\partial z}{\partial \psi}\right]. \tag{3}$$

3. Method

Figure 3 presents an overview of the *iFusion* framework. The key of our pose-free reconstruction framework is the sparse-view pose estimator shown in Fig. 3 (a). By inverting the diffusion model, accurate poses can be estimated. Next, the registered views are leveraged to customized the novel view synthesis model for the target object as in Fig. 3 (b). Finally, 3D reconstruction can be done using the registered views, and the customized diffusion model serves as the guidance, shown in Fig. 3 (c).

3.1. Diffusion as a Pose Estimator

The goal is to recover the relative camera pose $T_{r\to q}$ from a reference view x^r to the query view x^q, leveraging the pre-trained diffusion model ϵ_θ. Intuitively, a model trained for a task involving camera poses could potentially be used in reverse: to retrieve or estimate the camera pose from given inputs, as evident in [7, 47, 81]. Hence, rather than optimizing DM parameters θ to reconstruct x^q given $c(x^r, T_{r\to q})$ as in the training stage described in Eq. (2) and Sec. 2, we solve the inverse problem by freezing θ and optimizing $\hat{T}_{r\to q}$ to reconstruct x^q:

$$\hat{T}_{r\to q} = \underset{T\in \mathrm{SE}(3)}{\operatorname{argmin}} \mathcal{L}(x^q, c(x^r, T)). \tag{4}$$

To minimize Eq. (4), we query a view in its latent space $z_t \sim \mathcal{E}(x^q)$ using Eq. (1), followed by denoising z_t to $\hat{z}_{t-1}$ conditioned on $c(x^r, \hat{T}_{r\to q})$. Finally, we compute the residuals for backpropagation of the transformation's gradient $\nabla \hat{T}_{r\to q}$. To ensure that the estimated pose $\hat{T}_{r\to q}$ continue to lie on the SE(3) manifold during the gradient-based optimization, we parameterize the pose $T_{r\to q} = \exp(\hat{\xi})$, where $\xi \in \mathbb{R}^6$ is the twist coordinates of the Lie algebra $\mathfrak{se}(3)$ associated with the Lie group SE(3) [37], following Engel et al. [11]. Therefore, we reformulate Eq. (4) as follows:

$$\xi_{r\to q} = \underset{\xi\in \mathfrak{se}(3)}{\operatorname{argmin}} \mathcal{L}(x^q, c(x^r, \exp(\hat{\xi}))). \tag{5}$$

Note that Eq. (5) can further be constrained by the inverse transformation defined by the same vector representation, *i.e.*, $T_{q\to r} = \exp(-\hat{\xi})$. We therefore obtain:

$$\xi_{r\to q} = \underset{\xi\in \mathfrak{se}(3)}{\operatorname{argmin}} \mathcal{L}(x^q, c(x^r, \exp(\hat{\xi}))) + \mathcal{L}(x^r, c(x^q, \exp(-\hat{\xi}))). \tag{6}$$

[2] https://github.com/CompVis/stable-diffusion
[3] https://laion.ai/blog/laion-aesthetics/

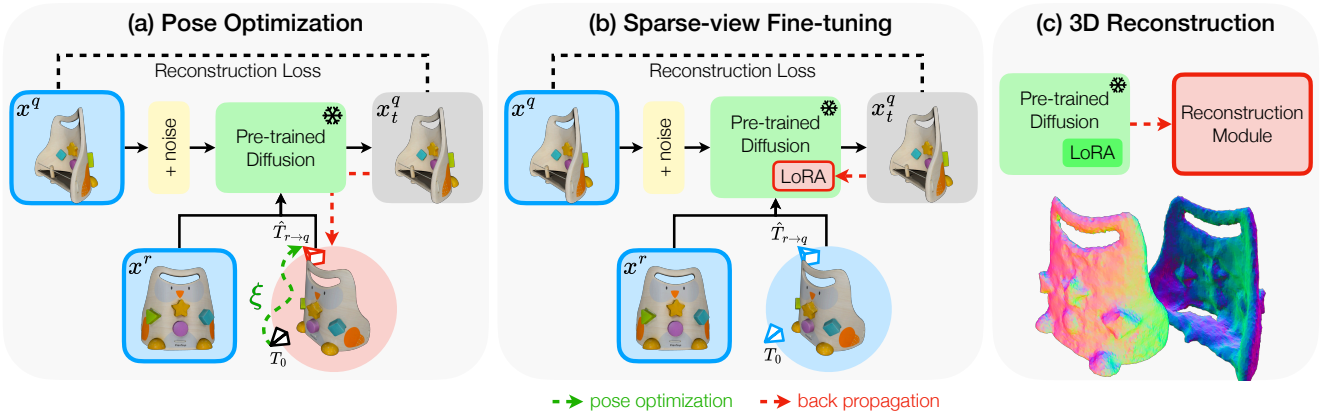

Figure 3. **iFusion framework.** (a) Given as few as two pose-free images (x^r, x^q), we estimate the pose $\hat{T}_{r\to q}$ from T_0 to optimally reconstruct the input view through the frozen diffusion model. (b) Based on $\hat{T}_{r\to q}$, we efficiently fine-tune the diffusion model by LoRA [17] to customize the model to synthesize novel views of the given object with enhanced fidelity. (c) Conditioned on $\hat{T}_{r\to q}$ and the refined diffusion model, we optimize a reconstruction module to perform sparse view 3D reconstruction.

In practice, we initialize our optimization from four distinct canonical poses relative to the reference view, *i.e.*, front, left, right, and back, designated as T_0. This helps reduce the possibility of stucking at a local minima during the optimization. The final estimated camera pose can be denoted as follows:

$$\hat{T}_{r\to q} = T_0 \cdot \exp(\hat{\xi}_{r\to q}). \tag{7}$$

Furthermore, taking inspiration from Huang et al. [18], instead of sampling the timestep t from a uniform distribution as in training, we linearly decrease t. This adjustment aligns with diffusion models' coarse-to-fine progressive optimization and has been empirically observed to lead to more stable optimization.

3.2. From Single-View to Multi-View

Even with a fairly accurate estimated pose $\hat{T}_{r\to q}$, there is still no guarantee that the diffusion model generates the pixel-exact query image x^q. We propose to close the gap by further fine-tuning the DM with the given views and estimated poses. However, due to limited training samples, naively optimizing all trainable parameters θ is inefficient and may jeopardize the pre-trained model. To this end, we incorporate LoRA [17], injecting thin trainable layers ϕ to the attention module in the U-Net ϵ_θ while freezing the pre-trained θ. The objective in Eq. (2) is reformulated as follows:

$$\mathcal{L}_\phi(x, c) = \mathbb{E}_{z,\epsilon,t}\left[\|\epsilon - \epsilon_{\theta,\phi}(z_t, t, c)\|_2^2\right], \tag{8}$$

where $(x, c) \in \left\{\left(x^q, (x^r, \hat{T}_{r\to q})\right), \left(x^r, (x^q, \hat{T}_{q\to r})\right)\right\}$. In other words, the fine-tuning process adapts the DM to generate the query view x^q from the condition $c(x^r, \hat{T}_{r\to q})$, and vice versa, for a specific object. Empirically, this LoRA fine-tuning effectively customize the DM to generate novel views different from x^r and x^q of the target object, despite the small number of training samples and parameters ϕ, and the inherent noise from the estimated poses.

While the original Zero123 only conditions on a single view, we have multiple images available along with their relative transformations in a sparse-view setting.[4] This raises the question: How can we better utilize these additional views for improved generation quality? To address this, we employ a simple stochastic conditioning strategy inspired by Watson et al. [75]. The key concept is that all given views should collectively shape the final output. More specifically, we randomly sample a registered view as the input condition at each denoising timestep. Empirically, this stochastic multi-view conditioning (MVC) significantly improves the novel view synthesis results compared to naively using the nearest view as the condition. Moreover, the final reconstruction quality is also improved.

3.3. From Sparse Views to 3D Reconstruction

There are two primary lines of existing literature for 3D object reconstruction via diffusion, namely image-based reconstruction [32, 34] and SDS-based generation [30, 48, 49, 67]. To integrate our proposed technique with the image-based approaches, we may simply generate multi-view images using the fine-tuned model obtained from Eq. (8) with stochastic multi-view conditioning outlined in Sec. 3.2, and then feed them as the training data to the differentiable renderer, *e.g.*, NeRF [40] and NeuS [71]. For SDS-based methods, in addition to Eq. (3), we further incorporate the recon-

[4] We mainly formulate the two-view setting (x^r and x^q). Multi-view settings are achieved via treating all distinct image pairs as query-reference pairs and estimating the pose transform for each pair.

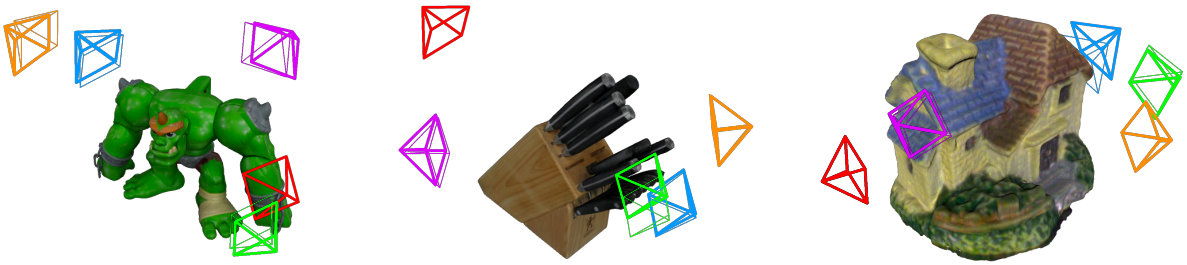

Figure 4. **Qualitative results on pose estimation.** We visualize the predicted poses (thin) alongside the ground truth (bold), using the same color, while the reference views are plotted in red. *iFusion* accurately predicts poses even on the opposite side of the reference view (red), emphasizing its effectiveness in leveraging the strong prior knowledge embedded in Zero123 [33].

struction loss on the registered input views:

$$\mathcal{L}_{rec} = \left\| x - \mathcal{R}_{\psi}(\hat{T}) \right\|_2^2, \tag{9}$$

where x is the input image and $\mathcal{R}_{\psi}(\hat{T})$ is the rendered view from viewpoint $\hat{T}$ acquired from Eq. (7). The final objective is the weighted sum of $\mathcal{L}_{rec}$ and $\mathcal{L}_{SDS}$. For above steps, the LoRA model and MVC are also employed.

4. Experiments

4.1. Experimental Setup

Datasets. We conduct experiments using two publicly available object datasets: Google Scanned Object (GSO) [10] and OmniObject3D (OO3D) [76]. We randomly select 70 instances from each dataset, synthesizing 5 camera poses and rendering their observation views. For pose estimation experiments, we sample 1000 sets from the available combinations of views per dataset. In view synthesis and reconstruction tasks, we select two views from the rendered five with the largest parallax motion around the object to minimize the overlapping between views.

Experiments and Metrics. We evaluate our proposed framework on pose estimation, novel view synthesis, and 3D reconstruction. For pose estimation, we report the relative pose error in rotation and translation, where the rotation error is the angle between estimated and ground truth poses, and the translation error is their positional difference. We adopt the standard metrics PSNR, SSIM, and LPIPS to evaluate novel view synthesis results, following Liu et al. [33], Mildenhall et al. [40]. For 3D reconstruction, we report Chamfer Distances and volumetric IoU between ground truth shapes and reconstructed ones.

4.2. Experimental Result

Pose Estimation. We compare our proposed method with RelPose++ [28], FORGE [21], RayDiffusion [86], and SfM-based HLoc [57, 58] for pose estimation given 2∼5 views of an object. We adopted the official pre-trained checkpoints for all baselines, *i.e.*, RelPose++ and RayDiffusion trained on CO3Dv2 [53], and FORGE trained on a variant of ShapeNet [4]. In the case of HLoc, we utilized SuperPoint feature [9] and LightGlue [31] for matching. We did not compare with SparsePose [61] and PF-LRM [72] since no public source code is available. Quantitative and qualitative results are depicted in Fig. 5 and Fig. 4, respectively. Figure 5 verifies the effectiveness of our proposed solution over the baselines with substantial improvements across all metrics. By leveraging Zero123 [33], *iFusion* excels at handling diverse objects thanks to its rich visual knowledge learned from Objaverse [8], which significantly differentiates our method from the baselines. In Fig. 4, we again corroborate that our solution accurately estimates camera poses even with little overlap, *e.g.*, the blue camera is on the opposite sides to the camera reference (red camera).

Although we aim to compare all methods under a training-free setting to emphasize generalization, we fine-tune RelPose++ on Objaverse as its fine-tuning achieves consistent improvement. However, even with this fine-tuning, the performance of RelPose++ remains notably lower than *iFusion*. Note that FORGE's performance is considerably lower than the officially reported results. We argue that this inconsistency may arise from differences in rendering styles, which underscores a potential limitation in FORGE's generalization ability. Similar observations regarding this issue have also been reported in PF-LRM.

Novel View Synthesis. Table 1 shows our novel view synthesis comparison against 2D-based Zero123[5], 3D-based methods, *i.e.*, FORGE and LEAP [22], and hybrid-based (2D+3D) SyncDreamer [34]. It is observed that both the 3D-based methods do not perform well under extremely few-view scenarios. Moreover, *iFusion* significantly outperforms all methods on all metrics. Figure 6 includes qualitative examples to demonstrate *iFusion*'s advantage in novel view synthesis. We observe that images generated by Zero123, although mostly visually plausible, do not faithfully represent the actual objects, especially those with

[5]By default, we use Zero123-XL for all modules that require Zero123.

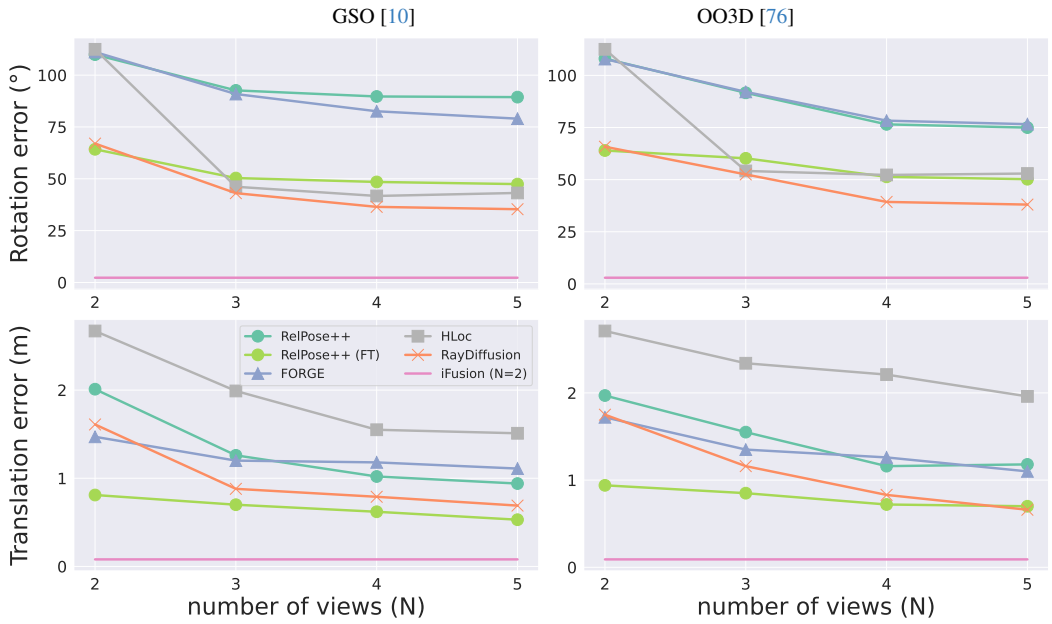

Figure 5. **Evaluation results on pose estimation.** *iFusion* achieves significant improvements in pose estimation with only 2 input views on both datasets. To ensure a comprehensive evaluation, we also assess baseline methods using more views, yet our method consistently outperforms them.

Table 1. **Novel view synthesis results.** *iFusion* performed significantly better than the original Zero123, SyncDreamer, and 3D-based methods.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓
GSO [10]	FORGE [21]	10.45	0.673	0.449
	LEAP [22]	12.51	0.751	0.312
	Zero123 [33]	15.40	0.788	0.184
	SyncDreamer [34]	15.67	0.806	0.180
	iFusion	**18.73**	**0.836**	**0.121**
OO3D [76]	FORGE	10.48	0.684	0.447
	LEAP	12.63	0.759	0.305
	Zero123	15.84	0.801	0.184
	SyncDreamer [34]	15.98	0.814	0.181
	iFusion	**19.78**	**0.851**	**0.117**

complex geometry. SyncDreamer, an enhanced iteration of Zero123, yields better outcomes but shares the same limitation. In contrast, our *iFusion* improves novel views' image fidelity by conditioning on an additional pose-free view.

3D Reconstruction. We showcase the efficacy of the *iFusion* framework in 3D reconstruction by integrating it with various existing reconstruction methods. Specifically, One-2-3-45 [32] represents image-based methods, which directly regresses SDFs from the generated multi-view images; on the other hand, Zero123-SDS [33], Magic123 [49], and DreamGaussian [67] are SDS-based approahces. For completeness, Zero123-SDS optimizes Instant-NGP [43] via Zero123-guided SDS. Magic123 combines Zero123 and SD for improved quality.[6] DreamGaussian leverages the 3D Gaussian Splatting renderer [25]. As illustrated in Tab. 2 and Fig. 7, the incorporation of *iFusion* enhances the performance of all reconstruction modules by a large margin. In addition, *iFusion* clearly outperforms other none-optimization-based methods Point-E [44] and Shape-E [23], which are trained on a large-scale private dataset. To conclude, when faithful reconstruction is desired, *iFusion* is extremely beneficial, requiring very few additional views that can be casually captured without knowing the camera poses.

4.3. Ablation Study

Sparse-view Fine-tuning. Table 3 assesses the efficacy of the proposed fine-tuning stage for object-specific novel view synthesis. This process takes approximately 30 seconds per object on a 3090 GPU. Upon examining row (a), *i.e.*, Zero123, alongside row (b), it is evident that the performance is boosted by incorporating the additional view and an accurately estimated pose. Row (c) highlights the substantial improvement from the stochastic re-sampling of multi-view conditions at each timestep, providing more robust outcomes than row (b). Moreover, the multi-view fine-tuning with LoRA in row (d) significantly enhances performance by improving the understanding of the target object. Finally, row (e) underscores the potential for achieving higher-quality synthesis by incorporating more views. All

[6]The implementations of Zero123-SDS and Magic123 are adopted from threestudio: https://github.com/threestudio-project/threestudio.

Figure 6. **Qualitative examples on novel view synthesis.** *iFusion* takes two unposed images and Zero123 [33] only conditions on the first view. We observe that *iFusion* effectively leverages the additional images without camera poses and generates more faithful images.

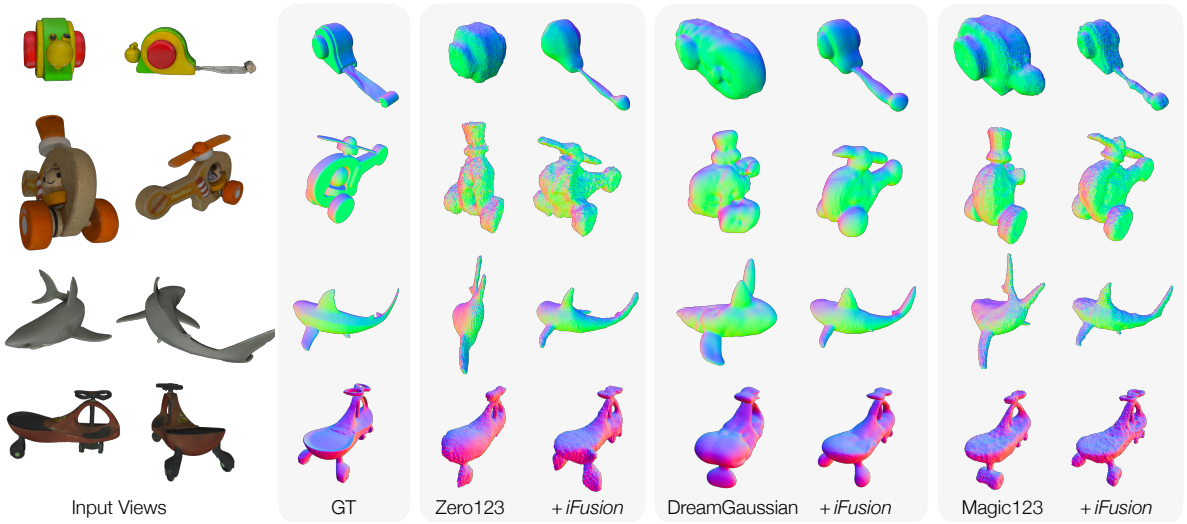

Figure 7. **Qualitative comparison of surface reconstruction.** It is clear that *iFusion* significantly enhances existing reconstruction methods including Zero123-SDS [33], DreamGaussian [67], and Magic123 [49], by adding an additional view without the camera pose.

are achieved with self-estimated camera poses.

3D Reconstruction. We validate the proposed components contributing to reconstruction in Tab. 4, using DreamGaussian as the reconstruction module on the OO3D dataset. The results in rows (a) and (b) distinctly illustrate that adding an extra view with an estimated pose and supervising with reconstruction loss significantly enhance the single-view baseline. Incorporating stochastic multi-view conditioning (MVC) further improves the performance, as evident in row (c). Finally, fine-tuning via LoRA demonstrates an additional improvement in customizing the model for faithful reconstruction of the given object.

5. Related Work

Few-shot NeRFs. Neural Radiance Fields (NeRFs) [40] have revolutionized 3D modeling with its powerful representations and high-fidelity render quality, but struggling under insufficient views. Follow-up works introduced regularizations to stabilize training [26, 45, 80], or prior models for auxiliary 3D reasoning [5, 19, 73, 82]. Nevertheless, the

Table 2. **Evaluation results on 3D Reconstruction.** We integrate *iFusion* with various state-of-the-art single-view reconstruction baselines and consistently improve their performance.

Method	GSO [10]		OO3D [76]	
	Chamfer Dist. ($\times 10^2$) ↓	Volume IoU (%) ↑	Chamfer Dist. ($\times 10^3$) ↓	Volume IoU (%) ↑
Point-E [44]	6.414	18.92	6.766	19.83
Shape-E [23]	5.839	29.00	6.086	29.02
One-2-3-45 [32]	7.173	28.77	5.424	43.75
+ *iFusion*	6.359	31.68	4.739	48.32
Zero123-SDS [33]	6.456	33.63	5.676	45.90
+ *iFusion*	4.178	39.73	3.293	56.36
DreamGaussian [67]	4.728	35.35	4.298	44.35
+ *iFusion*	3.977	42.07	2.947	57.58
Magic123 [49]	4.839	39.46	3.842	53.69
+ *iFusion*	**3.076**	**46.70**	**2.682**	**60.31**

Table 3. Ablation study of novel view synthesis on GSO [10]. Multi-view conditioning and LoRA finetuning are validated. Increased views also improve the results.

	n views	Strategy	LoRA [17]	PSNR↑	LPIPS↓
(a)	1	-	-	15.40	0.184
(b)	2	closest-view	-	16.19	0.169
(c)	2	multi-view	-	17.30	0.149
(d)	2	multi-view	✓	18.73	0.121
(e)	4	multi-view	✓	**21.32**	**0.092**

Table 4. Ablation study of 3D reconstruction on OO3D [76] using DreamGaussian [67]. MVC and LoRA are essential to the best results.

	n views	MVC	LoRA	Chamfer Dist. ↓	IoU ↑
(a)	1	-	-	4.298	44.35
(b)	2	-	-	3.427	53.04
(c)	2	✓	-	3.241	54.16
(d)	2	✓	✓	**2.947**	**57.58**

*Chamfer distance measured by $\times 10^2$ and IoU in (%)

dependency on precise camera poses remains an issue, as Lin et al. [29] showed that inaccurate poses, which often arise in pose estimation using a limited number of views, lead to degraded performance.

Diffusion for 3D Generation. Diffusion models [16, 62, 63] have emerged as the leading visual generative models. They generate visually plausible images from various input conditions [12, 13, 27, 39, 78, 79] and customize or edit existing photos with diverse controlling signals [2, 14, 50, 55, 87, 88]. Promising results have been achieved in 3D generation as well, spanning various representations such as point-clouds [36, 83, 89], voxel grids [42, 89], and tri-planes [1, 15, 60]; however, they are constrained by the limited diversity of 3D datasets, *e.g.*, ShapeNet [4]. To overcome the data scarcity, researchers utilize pre-trained 2D diffusion models [54, 56] for text-to-3D generation [6, 30, 48, 74], and further extend them for single-view reconstruction [32, 33, 38, 49, 66, 67], where the diffusion model *"dreams up"* unobserved views. However, single-view methods diverge from real-world reconstruction scenarios — the target object needs to be accurately reconstructed, not over-imagined. Although several methods propose to include additional views, accurate camera poses are still assumed [3, 24, 65, 90].

Reconstructing from Pose-free Sparse-views. To recover the unknown camera poses from sparse views, recent studies have explored either by directly regressing the pose [21, 28, 85] or through iterative refinement [61, 70, 86]. The estimated poses can then be utilized for reconstruction [21, 29, 84]. Notably, FORGE [21] combines the above two stages but lacks robustness for diverse objects and lighting. A recent follow-up, LEAP [22], utilizes DINO [46] as a prior, improving generalization but facing challenges in recovering unseen regions. PF-LRM [72], akin to our approach, achieves generalized reconstruction using knowledge acquired from Objaverse [8]. Although PF-LRM produced promising results, it differs from our training-free approach based on existing models as it requires extensive training on the Objaverse.

6. Conclusion

We propose *iFusion*, a framework that reconstructs 3D objects without requiring poses, by exploiting the rich visual knowledge in a large-scale pre-trained diffusion model. Given a few unposed images, we begin with inverting the diffusion for gradient-based pose optimization. The estimated poses, in turn, enhance the diffusion on view synthesis through multi-view fine-tuning and conditioning. Finally, by combining the estimated poses and the refined diffusion model, we demonstrate how *iFusion* achieves pose-free reconstruction. Experimental results show that our solution outperforms strong baselines on three key tasks: pose estimation, novel view synthesis, and 3D reconstruction.

References

[1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *CVPR*, 2023. 8

[2] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *CVPR*, 2023. 8

[3] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *ICCV*, 2023. 1, 8

[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5, 8

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *CVPR*, 2021. 7

[6] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *ICCV*, 2023. 8

[7] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *ECCV*, 2020. 2, 3

[8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 2, 3, 5, 8

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR-W*, 2018. 5

[10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, 2022. 5, 6, 8

[11] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*, 2014. 3

[12] Wan-Cyuan Fan, Yen-Chun Chen, Dongdong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. In *AAAI*, 2023. 8

[13] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*, 2022. 8

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2, 8

[15] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *ICML*, 2023. 8

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 8

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 2, 4, 8

[18] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. In *ICLR*, 2024. 4

[19] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, 2021. 1, 7

[20] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, 2022. 3

[21] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. In *3DV*, 2024. 1, 5, 6, 8

[22] Hanwen Jiang, Zhenyu Jiang, Yue Zhao, and Qixing Huang. Leap: Liberate sparse-view 3d modeling from camera poses. In *ICLR*, 2024. 1, 5, 6, 8

[23] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 6, 8

[24] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *CVPR*, 2023. 8

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):1–14, 2023. 2, 3, 6

[26] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *CVPR*, 2022. 7

[27] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 8

[28] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. In *3DV*, 2024. 1, 5, 8

[29] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, 2021. 8

[30] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 4, 8

[31] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 5

[32] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *NeurIPS*, 2023. 1, 4, 6, 8

[33] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 1, 2, 3, 5, 6, 7, 8

[34] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. In *ICLR*, 2024. 4, 5, 6

[35] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *ECCV*, 2022. 1

[36] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, 2021. 8

[37] Yi Ma, Stefano Soatto, Jana Košecká, and Shankar Sastry. *An invitation to 3-d vision: from images to geometric models*. Springer, 2004. 3

[38] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 reconstruction of any object from a single image. In *CVPR*, 2023. 1, 8

[39] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 8

[40] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 4, 5, 7

[41] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia*, 2022. 3

[42] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *CVPR*, 2023. 8

[43] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4):102:1–102:15, 2022. 6

[44] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 6, 8

[45] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022. 7

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 8

[47] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *CVPR*, 2020. 2, 3

[48] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 3, 4, 8

[49] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *ICLR*, 2024. 1, 4, 6, 7, 8

[50] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. Unicontrol: A unified diffusion model for controllable visual generation in the wild. In *NeurIPS*, 2023. 8

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[52] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Ben Mildenhall, Nataniel Ruiz, Shiran Zada, Kfir Aberman, Michael Rubenstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation. In *ICCV*, 2023. 1

[53] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 5

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 8

[55] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 8

[56] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 3, 8

[57] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 5

[58] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 5

[59] H. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965. 2

[60] J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *CVPR*, 2023. 8

[61] Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. SparsePose: Sparse-

view camera pose regression and refinement. In *CVPR*, 2023. 1, 5, 8

[62] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3, 8

[63] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 3, 8

[64] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 2

[65] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. In *ICCV*, 2023. 1, 8

[66] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *ICCV*, 2023. 1, 8

[67] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*, 2024. 1, 4, 6, 7, 8

[68] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 2

[69] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *CVPR*, 2023. 3

[70] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV*, 2023. 8

[71] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 4

[72] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. In *ICLR*, 2024. 5, 8

[73] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 7

[74] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023. 8

[75] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *ICLR*, 2023. 4

[76] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *CVPR*, 2023. 5, 6, 8

[77] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 2

[78] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *CVPR*, 2023. 8

[79] Binbin Yang, Yi Luo, Ziliang Chen, Guangrun Wang, Xiaodan Liang, and Liang Lin. Law-diffusion: Complex scene generation by diffusion with layouts. In *ICCV*, 2023. 8

[80] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *CVPR*, 2023. 7

[81] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting neural radiance fields for pose estimation. In *IROS*, 2021. 2, 3

[82] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 1, 7

[83] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In *NeurIPS*, 2022. 8

[84] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In *NeurIPS*, 2021. 8

[85] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *ECCV*, 2022. 1, 8

[86] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *ICLR*, 2024. 1, 5, 8

[87] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 8

[88] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*, 2023. 8

[89] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, 2021. 8

[90] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 1, 8