

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

Anonymous 3DV submission

Paper ID 188

1. Outline

We provide code, animations, and this document to complement our main paper. The following is its outline:

1. In Sec. 2, we delve into the implementation details, covering aspects such as hyper-parameters, runtime, and dataset collection. For demonstration purposes, we provide the code in the `code` directory.
2. In Sec. 3, we provide an ablation study on camera pose estimation to verify the efficacy of the proposed canonical poses and timestep annealing.
3. In Sec. 4, we showcase additional qualitative results to validate the effectiveness of our method visually. Dynamic visualizations of example reconstructions are available in the `video` directory.
4. In Sec. 5, we present the experimental results obtained from evaluating our method on the GSO dataset [2] released by FORGE [4].
5. In Sec. 6, we discuss the limitations of our work and potential ways to improve them.

2. Implementation Details

2.1. Hyper-parameters

We employ Adam [5] as the optimizer for both pose estimation and sparse-view fine-tuning. In the case of pose estimation, we optimize the initial poses for 100 steps with an initial learning rate of 0.1. The learning rate is dynamically reduced if the L2 loss stops decreasing, handled by the *ReduceLROnPlateau* scheduler from PyTorch [11]. Specifically, we set the reduction factor to 0.6 and the patience to 10. Afterward, in the sparse-view fine-tuning stage, the model is fine-tuned for 30 steps, with the learning rate annealed from 10^{-3} to 10^{-4} and the rank of the injected LoRA [3] parameter set to 12. Note that this process is performed individually for each object, taking approximately 30 seconds on a single Nvidia 3090 GPU with a batch size of 16. For 3D reconstruction, we follow the default hyper-parameters of each reconstruction method, *i.e.*, One2345 [7], Zero123-SDS [8], Magic123 [12], and DreamGaussian [13], when

combining with *iFusion*. Please refer to their official implementations for details.

For demonstration purposes, we provide the code, which includes several example objects for reviewers to test our method with. Additional details and instructions can be found in the `README.md` file located in the `code` directory.

2.2. Dataset Collection

We use Pyrender¹ to render images for evaluation. Following Liu et al. [8], the transformation is defined using the spherical coordinate system with θ , ϕ , and r representing the elevation angle, azimuth angle, and distance towards the center, respectively. In practice, we sample camera viewpoints on the unit sphere with $\theta \in [\pi/4, 3\pi/4]$, $\phi \in [0, 2\pi]$ and r is uniformly sampled in the interval of $[1.2, 2.0]$. The field of view of the perspective camera is set to 49.1° . All images are rendered in the resolution of 512×512 with transparent background.

3. Ablation on Pose Estimation

We provide additional ablation to validate whether the use of more poses for initialization, namely T_0 in Eq. (7), leads to more accurate camera pose estimation, and it is confirmed in Tab. 1. The reported computation time is measured on a single Nvidia 3090 GPU, and the recall is assessed based on the rotation error. According to Tab. 1, we employed $n = 4$ initial poses for a better trade-off between speed and accuracy for all experiments unless otherwise specified. Additionally, we observed that linearly annealing the timestep t lead to significantly more accurate pose estimation, as demonstrated in Tab. 2.

4. Qualitative Results

To further corroborate the effectiveness of our proposed pose estimation strategy described in Sec. 3.1, we present additional qualitative visualization in Fig. 1. These results support our assumption that the acquired understanding of

¹<https://github.com/mmatl/pyrender>

Table 1. Ablation study of the **number of initial poses** for pose estimation on GSO [2].

	n poses	Recall \uparrow			Time (s) \downarrow
		5°	10°	20°	
(a)	1	33.07	36.21	38.36	22.30
(b)	2	60.57	69.14	73.07	38.51
(c)	4	74.79	84.29	88.57	70.59
(d)	8	78.21	88.93	92.43	133.73

Table 2. Ablation study of **timestep annealing** for pose estimation on GSO [2].

	n poses	t annealing	Recall \uparrow		
			5°	10°	20°
(a)	4	-	48.61	56.67	61.39
(b)	4	✓	74.79	84.29	88.57

diverse objects in Zero123 [8] can be leveraged for other tasks, such as pose estimation.

In Fig. 2, we visualize the generated images obtained through the proposed multi-view fine-tuning and conditioning described in Sec. 3.2. The results demonstrate that *iFusion* is capable of leveraging additional views, ensuring high-fidelity generation. Finally, we showcase a more comprehensive comparison of reconstructed objects in Fig. 3. These results complement Fig. 7 of our main manuscript.

In addition, animations of the reconstructed objects are provided in the `video` directory, showing the outcomes with and without the application of *iFusion* to Magic123 [12]. The input views, *i.e.*, the reference view and the query view, are also provided for comparison. Consistently, incorporating an additional view using our method yields the best results.

5. Evaluation on FORGE’s dataset

In this section, we offer quantitative results obtained using the GSO dataset [2] sourced from FORGE [4], complementing the pose estimation experiment outlined in Sec. 4.2. The OO3D dataset [14] used by the authors remains unavailable to the public. It is important to note that GSO and OO3D are object datasets that do not provide rendered images, potentially resulting in slight variations in rendering style between our dataset and theirs. Additionally, due to the absence of a common test split within the community, the selection of objects for evaluation may differ as well. In this experiment, we strictly followed the author’s settings and conducted the experiments using the official checkpoint. Similar to Fig. 5 of the main paper, we present the results of *iFusion* utilizing two views, in contrast to

Table 3. **Evaluation results on GSO [2] from FORGE [4]**. Even with only two input views, *iFusion* achieves accurate pose estimation, surpassing the performance of FORGE with five views.

Dataset	Method	Rot. error \downarrow	Trans. error \downarrow
GSO [2]	FORGE [4]	14.90	0.37
	<i>iFusion</i>	3.16	0.11

FORGE’s official employment of five views. As depicted in Tab. 3, this experiment confirms the effectiveness of the proposed method.

6. Limitations and Future Works

While our methods deliver highly accurate camera poses, our pose estimation run time is higher than feed-forward-based methods, *e.g.*, RelPose++ [6]. This is attributed to the optimization nature of our approach, which involves back-propagation for updating the poses. Moreover, when we fine-tune Zero123 [8] on estimated poses and additional input views, it is worth noting that Zero123, originally adapted from the 2D-based Stable Diffusion (SD), lacks 3D awareness. This structural limitation prevents it from generating multi-views with consistency. However, our framework holds potential for integration with other diffusion-based novel view synthesizers [9, 10] that enforce multi-view consistency by incorporating 3D-aware modules onto SD.

References

- [1] 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. 3
- [2] 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. 1, 2
- [3] 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. 1
- [4] Hanwen Jiang, Zhenyu Jiang, Kristen Grauman, and Yuke Zhu. Few-view object reconstruction with unknown categories and camera poses. In *3DV*, 2024. 1, 2
- [5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1
- [6] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. In *3DV*, 2024. 2
- [7] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single



Figure 1. **More qualitative results on pose estimation.** The predicted poses (thin) and their corresponding ground truth (bold), are plotted in the same color, while the **reference views** are plotted in red. Our method achieves accurate pose estimation for diverse objects, benefiting from the extensive knowledge acquired from Objaverse [1].

- image to 3d mesh in 45 seconds without per-shape optimization. In *NeurIPS*, 2023. 1
- [8] 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, 6
- [9] 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. 2
- [10] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 2
- [11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 1
- [12] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skokhodov, 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, 2, 6
- [13] Jiayang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*, 2024. 1, 6
- [14] 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. 2

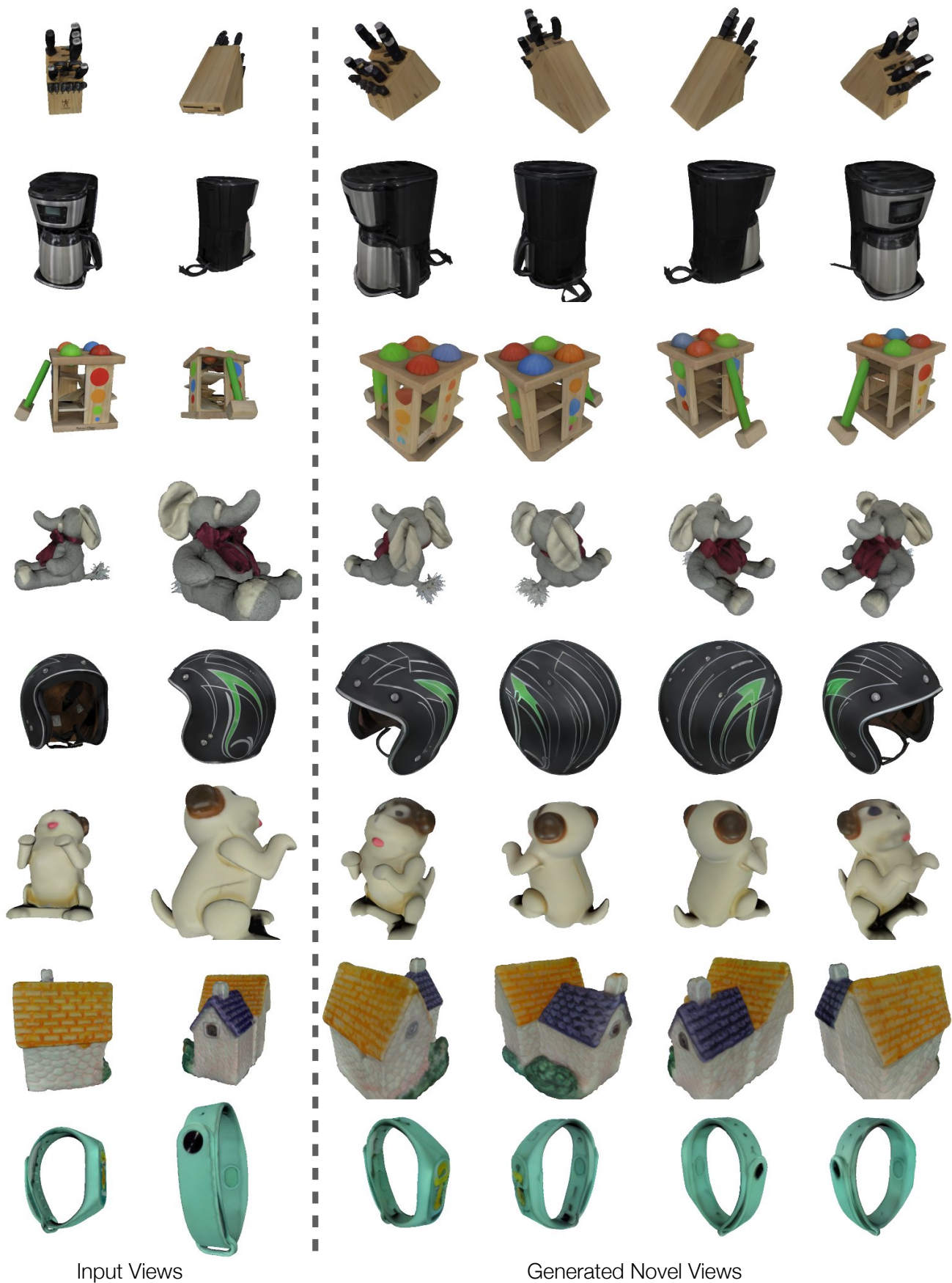


Figure 2. **More qualitative results on novel view synthesis.** By employing the proposed multi-view fine-tuning and conditioning, *iFusion* can efficiently incorporate additional views into the model, resulting in enhanced generation fidelity.

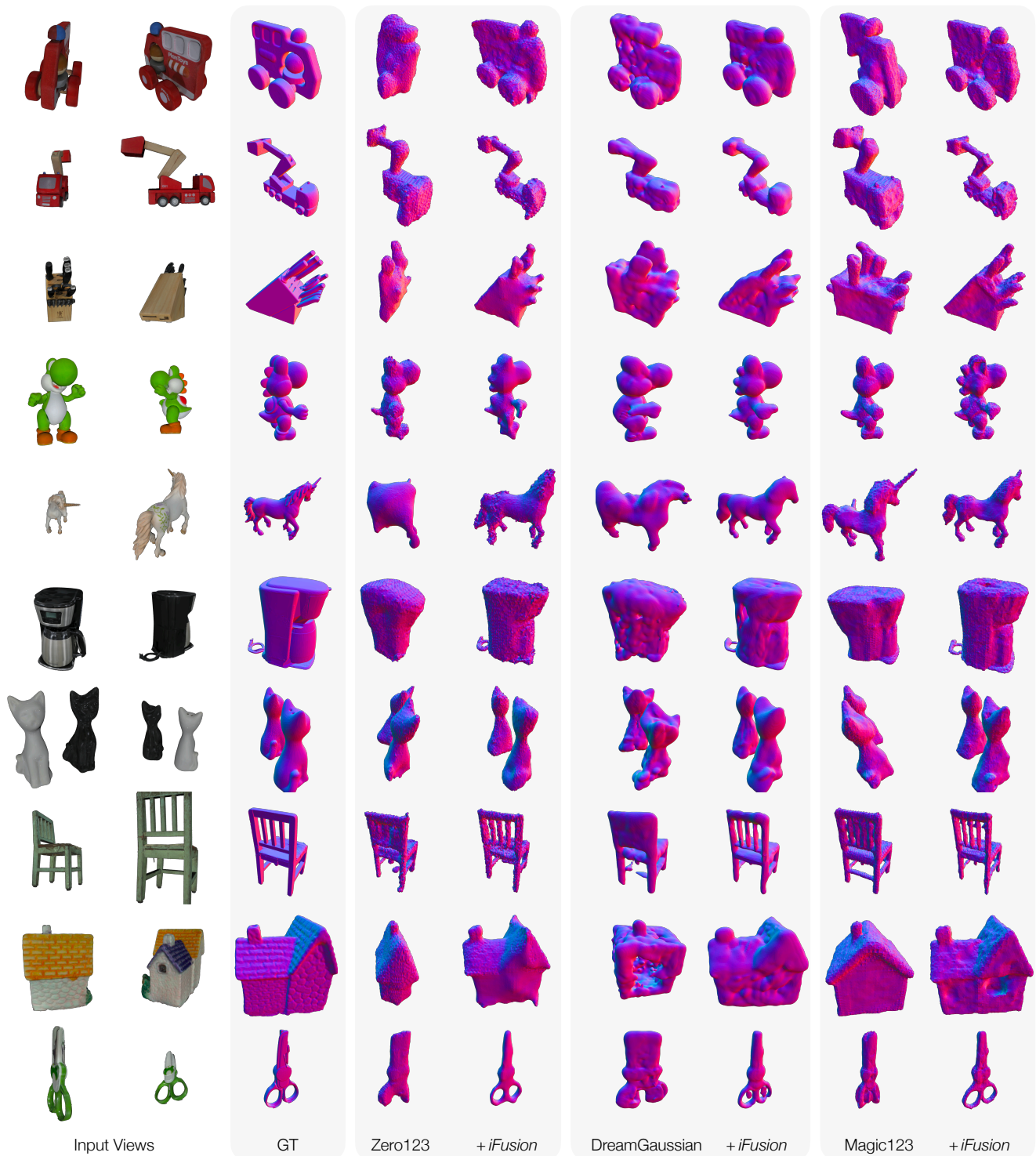


Figure 3. **More qualitative comparisons on surface reconstruction.** We integrate *iFusion* with Zero123-SDS [8], DreamGaussian [13], and Magic123 [12] to perform pose-free reconstruction given sparse views. The results indicate that our method operates as an effective add-on, consistently enhancing existing single-view reconstruction methods.