GSFix3D: Diffusion-Guided Repair of Novel Views in Gaussian Splatting

Jiaxin Wei[1] Stefan Leutenegger[2] Simon Schaefer[1]
[1]Technical University of Munich [2]ETH Zurich

Abstract

*Recent developments in 3D Gaussian Splatting have significantly enhanced novel view synthesis, yet generating high-quality renderings from extreme novel viewpoints or partially observed regions remains challenging. Meanwhile, diffusion models exhibit strong generative capabilities, but their lack of awareness of specific scene information hinders accurate 3D reconstruction tasks. To address these limitations, we introduce **GSFix3D**, a novel framework that improves the visual fidelity in under-constrained regions by distilling prior knowledge from diffusion models into 3D representations, while preserving consistency with observed scene details. At its core is **GSFixer**, a latent diffusion model obtained via our customized fine-tuning protocol that can leverage both mesh and 3D Gaussians to adapt pretrained generative models to a variety of environments and artifact types from different reconstruction methods, enabling robust novel view repair for unseen camera poses. Moreover, we propose a random mask augmentation strategy that empowers GSFixer to plausibly inpaint missing regions. Experiments on challenging benchmarks demonstrate that our GSFix3D and GSFixer achieve state-of-the-art performance, requiring only minimal scene-specific fine-tuning on captured data. Real-world test further confirms its resilience to potential pose errors. Our code and data are publicly available: https://github.com/GSFix3D/GSFix3D.*

1. Introduction

3D Gaussian Splatting (3DGS) [9] has recently emerged as an efficient and expressive explicit representation that models scenes using a set of 3D Gaussian primitives and enables photorealistic rendering through differentiable rasterization. Compared to previous Neural Radiance Fields (NeRF) [16] approaches, it achieves faster convergence and significantly higher rendering speeds. However, a key limitation persists in those optimization-based representations as they heavily rely on meticulously curated and densely sampled input views to achieve high visual fidelity near the training camera poses. In regions with sparse observations or from viewpoints that deviate substantially from the training data, 3DGS struggles to infer plausible geometry and appearance, often producing artifacts such as incomplete surfaces, unnatural geometry, or visible holes that severely degrade image quality. Moreover, obtaining sufficient coverage and accurate measurements often requires labor-intensive data collection, costly high-end 3D scanners, and skilled operators, which largely limits the accessibility of such methods for casual users with only mobile devices.

In parallel, text-to-image generative models based on latent-space denoising diffusion, such as Stable Diffusion [25], have shown the remarkable ability to synthesize diverse and photorealistic images. Trained on large-scale, captioned images from the internet, those models effectively gain a deep understanding of 2D visual concepts. To obtain greater control over diffusion model outputs, a variety of techniques, such as ControlNet [40], T2I-Adapters [17], and LoRA [5], have been proposed. Though powerful, these methods are primarily designed for image generation rather than repairing, and thus often lack input-output consistency, making them unsuitable for direct integration into 3D reconstruction pipelines where spatial and visual fidelity are critical.

To combine the strengths of diffusion models with existing 3D reconstructions, we introduce a novel view repair framework, GSFix3D, tailored for 3D Gaussian Splatting. Our method renders novel view images from initial reconstructions and refines them using a scene-adapted latent diffusion model by removing rendering artifacts and completing missing content. These enhanced images are then treated as pseudo-inputs and lifted back into 3D space to improve the underlying reconstruction. The key to our pipeline is a dedicated fine-tuning strategy that enables the pretrained diffusion model to internalize scene-specific priors, model artifact patterns, and develop inpainting capabilities using our proposed random mask augmentation. In contrast to DIFIX [35], which relies on large-scale curated real image pairs for training yet still lacks inpainting capabilities and struggles with unseen artifacts, our method requires only a one-time pretraining on two small synthetic datasets [1, 24] to obtain a general base model. This base model can then be efficiently fine-tuned on the same cap-

tured data used for initial reconstruction, enabling adaptation to diverse scenes. The resulting module, GSFixer, acts as a plug-and-play image enhancer, transforming imperfect renderings into high-quality, photorealistic images. Our main contributions are as follows:

- We propose GSFix3D, a new pipeline for repairing novel views in 3DGS reconstructions that leverages the diffusion model, GSFixer, to enhance under-constrained regions. We exploit the complementary properties between 3DGS and traditional mesh representations to further boost repairing performance.
- We introduce a customized fine-tuning protocol for pretrained diffusion models tailored to the novel view repair task. This protocol efficiently adapts the model to diverse scenes and reconstruction pipelines and enables it to internalize scene-specific priors, learn artifact patterns, and develop strong inpainting capabilities through our proposed random mask augmentation.
- Experiments on challenging benchmarks demonstrate state-of-the-art performance under extreme novel viewpoints, with only a few hours of fine-tuning on the same captured data used for reconstruction using a single consumer GPU. Additional tests on self-collected real-world data further validate its robustness to pose inaccuracies. We will release the real-world data and selected extreme novel views from the Replica dataset [29].

2. Related Work

2.1. 3D Reconstruction and Mapping

Traditional dense reconstruction methods, such as KinectFusion [18], fuse per-frame depth maps into a volumetric grid. Follow-up work improves scalability by using efficient data structures like octrees [28, 31] and voxel hashing [20, 21]. Though the reconstructed geometry suffices for robotics tasks such as navigation, it often lacks realism in visualization. Recently, NeRF [16] represents scenes as implicit neural functions. Several NeRF-based SLAM systems combine tracking and mapping within this framework [30, 43]. Despite producing high-quality renderings, NeRF methods are computationally expensive and struggle with real-time applications. 3DGS [9] addresses these limitations by representing scenes with explicit, differentiable Gaussian primitives, enabling faster rendering and optimization. This has led to several 3DGS-based SLAM systems: GS-SLAM [37] uses opacity thresholds to drive adaptive Gaussian insertion, SplaTAM [8] employs a densification mask based on rendered silhouettes and depth, while MonoGS [15] relies on monocular depth estimates with variable uncertainty. To improve efficiency, RTG-SLAM [23] categorizes Gaussians as either opaque or transparent and updates only unstable ones, whereas GSFusion [33] integrates Truncated Signed Distance Field (TSDF) [2] and 3DGS in a hybrid framework and employs a quadtree-based image segmentation strategy to reduce redundant splats. Despite these advances, challenges persist in handling under-constrained areas and achieving artifact-free reconstruction. We build our approach on 3DGS reconstructions due to their real-time performance, photorealistic rendering, and full differentiability, which make them particularly suitable for downstream repair tasks.

2.2. Novel View Repair

Although dense-view reconstruction has become increasingly reliable, novel view rendering remains susceptible to artifacts, especially in under-constrained regions. Prior work has largely focused on sparse-view settings, where such degradation is more obvious. [13] introduces a deceptive diffusion model that refines novel views rendered from few-view reconstructions and uses an uncertainty measure to improve consistency. RI3D [22] uses two separate diffusion models for repairing visible regions and inpainting missing areas, whereas ours integrates these tasks into a single model. To improve temporal coherence, several methods leverage video diffusion models. 3DGS-Enhancer [14] is the first to train a video diffusion model on a large-scale dataset created with pairs of low and high-quality images. GenFusion [36] fine-tunes a video diffusion model on artifact-prone RGB-D videos using a masking strategy that simulates common view-dependent artifacts for content-aware outpainting, while [42] uses training-free scene-grounding guidance to steer the video diffusion model toward temporally consistent synthesis. Despite promising results, these methods rely heavily on customized preprocessing steps to bootstrap initial reconstructions and carefully curated datasets to train diffusion models effectively.

In this paper, we focus on novel view repair for reconstructions where artifacts still persist despite extensive coverage. SGD [39] introduces a tailored diffusion pipeline for autonomous driving scenarios, using adjacent frames as conditioning inputs and leveraging LiDAR point cloud to train a ControlNet for explicit depth control. DIFIX [35] takes a step toward general view repair by training a single-step diffusion model on a large curated dataset of real noisy–clean image pairs, created via handcrafted corruption strategies. However, its performance drops when exposed to unseen artifacts and it struggles with inpainting. In contrast, our GSFixer is obtained through a lightweight fine-tuning protocol. With minimal pretraining on synthetic data and fine-tuning on captured reconstruction data, GSFixer achieves robust artifact removal, adapts to diverse pipelines and scenes, and exhibits strong inpainting capabilities, all within a single model that runs efficiently on consumer hardware.

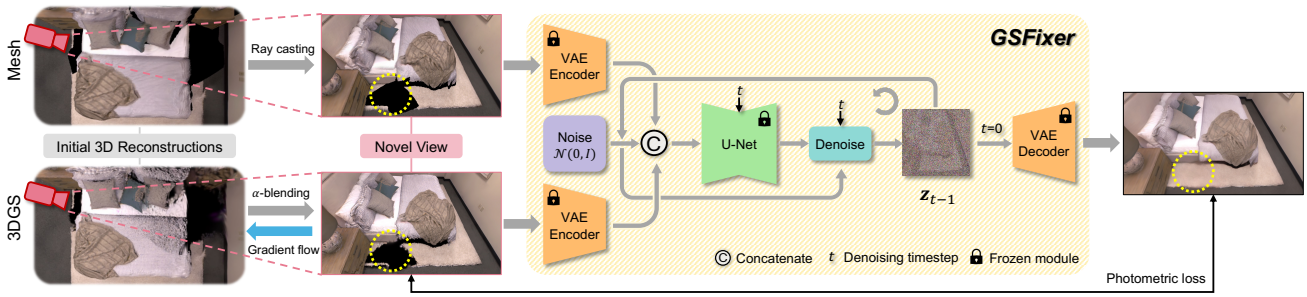

Figure 1. System overview of the proposed GSFix3D framework for novel view repair. Given initial 3D reconstructions in the form of mesh and 3DGS, we render novel views and use them as conditional inputs to GSFixer. Through a reverse diffusion process, GSFixer generates repaired images with artifacts removed and missing regions inpainted. These outputs are then distilled back into 3D by optimizing the 3DGS representation using photometric loss.

3. Method

Our goal is to enhance the photorealism of novel views in reconstructed 3DGS scenes, especially for viewpoints distant from the original camera trajectories and suffer from limited observations. We present a customized fine-tuning protocol to adapt a pretrained diffusion model for artifact removal and view inpainting (Sec. 3.1). We then describe our inference scheme (Sec. 3.2), and how the fine-tuned model integrates into the full pipeline to improve the visual quality of novel views (Sec. 3.3). An overview of our method is illustrated in Fig. 1.

3.1. Fine-Tuning Protocol

Given a reconstructed 3DGS scene, we formulate the image repair task as a conditional generation problem and fine-tune a pretrained latent diffusion model, i.e. Stable Diffusion v2 [25], to learn the conditional distribution $p(I^{gt}|I^c)$ where $I^{gt} \in \mathbb{R}^{H\times W\times 3}$ denotes the ground truth RGB image and $I^c \in \mathbb{R}^{H\times W\times 3}$ is the condition image rendered from the imperfect reconstruction.

In our approach, we further extend the conditioning input to two rendered images: one from the 3D Gaussian Splatting representation (I^{gs}) and another from a mesh representation (I^{mesh}). Thus, the actual conditional distribution becomes $p(I^{gt}|I^{mesh}, I^{gs})$. This dual-conditioning strategy is motivated by the complementary strengths of 3DGS and traditional mesh-based reconstructions. 3DGS, as an optimization-based method, tends to suffer in regions with sparse observations, often leading to visible artifacts such as holes or incomplete geometry. Mesh reconstructions, though usually less photorealistic at lower resolutions, offer more coherent geometry and stronger spatial priors in under-constrained areas. By jointly leveraging both representations, we aim to provide the diffusion model with richer appearance cues for image refinement. To ensure that the mesh input remains geometrically consistent yet independent from the 3DGS optimization process, we obtain the mesh and the 3DGS map simultaneously using GSFusion [33], an online RGB-D mapping system. This avoids directly extracting the mesh from the 3DGS representation, as done in prior works [3, 6], which could introduce correlated artifacts. The overall fine-tuning protocol is presented in Fig. 2. We conduct an ablation study in Sec. 4.4 comparing the performance of using both inputs versus 3DGS alone, validating the effectiveness of our design choice.

3.1.1 Network Architecture

Diffusion models [19, 25, 27] are a class of generative frameworks that generate data by learning to invert a progressively noised process. We use a frozen Variational Autoencoder (VAE) [10] to encode all images into a latent space, enabling diffusion-based learning in a more compact domain. For a given image I, its latent code is obtained via the encoder $\mathcal{E}$: $\mathbf{z} = \mathcal{E}(I)$. This results in a latent triplet $(\mathbf{z}^{mesh}, \mathbf{z}^{gs}, \mathbf{z}^{gt})$. To train the denoising model, we follow the standard Denoising Diffusion Probabilistic Models (DDPM) [4] formulation and incrementally add standard Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ to the clean ground-truth latent $\mathbf{z}_0 := \mathbf{z}^{gt}$ over T discrete timesteps, producing a sequence $\{\mathbf{z}_t\}_{t=1}^{T}$. The noisy latent at timestep t is then given by:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

where $\bar{\alpha}_t$ denotes the cumulative product of noise schedule coefficients [4, 26]. Following [7], we repurpose the U-Net backbone from the pretrained diffusion model into a conditional denoiser for image repair. We concatenate the latent codes along the feature dimension to form the input $\bar{\mathbf{z}}_t = \text{concat}(\mathbf{z}^{mesh}, \mathbf{z}^{gs}, \mathbf{z}_t)$. To accommodate the increased channel count, we expand the first layer of the U-Net by duplicating the original weight tensor and dividing its values by three. This design choice maintains the original weight distribution and prevents excessive activation scaling, allowing us to preserve the initialization behavior of the pretrained model while enabling conditional

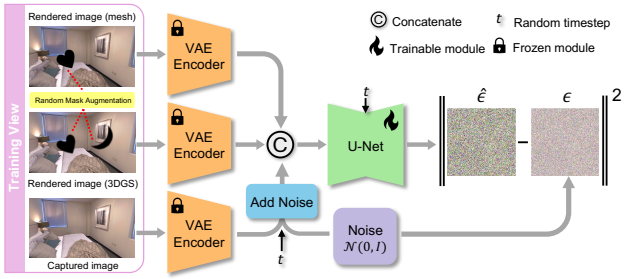

Figure 2. Illustration of the customized fine-tuning protocol for adapting a pretrained diffusion model into GSFixer, enabling it to handle diverse artifact types and missing regions.

inputs. The conditional U-Net ϵ_θ is then trained to predict the added noise by minimizing a standard DDPM objective:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0,\epsilon\sim\mathcal{N}(0,\mathbf{I}),t\sim\mathcal{U}[1,T]}\left[\|\epsilon - \hat{\epsilon}\|^2\right], \tag{2}$$

where $\hat{\epsilon} = \epsilon_\theta(\bar{\mathbf{z}}_t, t)$ is the predicted noise.

3.1.2 Data Augmentation

To construct our training set, we render each captured view using both the mesh and the 3DGS map, resulting in paired triplets $(I^{mesh}, I^{gs}, I^{gt})$, where I^{gt} is the original captured RGB image, I^{gs} is the image rendered from 3DGS via α-blending, and I^{mesh} is obtained via ray-casting on the mesh. This process requires no additional data beyond the original captured RGB images, their corresponding camera poses, and the reconstructed maps.

Direct fine-tuning on these triplets can already help the diffusion model adapt to the scene and learn to remove specific artifacts in the 3DGS rendering. However, one major challenge remains: the model's ability to inpaint missing regions, which usually appear as black holes in novel views due to under-constrained geometry or occlusions. Since all training images are rendered from the original captured viewpoints, they are mostly complete in appearance and fail to expose the model to such corner cases.

To explicitly train the model to handle incomplete renderings, we introduce a masking-based data augmentation scheme. For each training triplet, we randomly select a semantic mask from a set of annotated real-image masks [32]. The key intuition is to leverage the diverse mask shapes derived from real-world object semantics, which not only enhances the realism of the masked regions but also eliminates the need for manually designing complex rules to simulate missing areas caused by various factors such as occlusion or under-constrained observations. This mask is applied in two distinct ways: (1) the same mask is overlaid on both I^{mesh} and I^{gs}, simulating occlusions that might occur in novel views; and (2) an additional, independent mask is applied solely to I^{gs} to simulate the common degradation of 3DGS renderings in regions with limited observations. To better approximate the soft boundaries in 3DGS renderings, we further apply a small amount of Gaussian blur to the mask used on I^{gs}. We evaluate the impact of this augmentation strategy in Sec. 4.4, where we compare models trained with and without random masks and show its importance in improving the inpainting ability for novel views.

3.2. Inference with GSFixer

At inference time, we freeze the fine-tuned U-Net parameters and apply the model to novel views, as illustrated in Fig. 1. We begin by encoding the conditional inputs, i.e., the rendered images from novel viewpoints, into the latent space using the frozen VAE encoder. The latent for the target image to be generated, $\mathbf{z}_t$, is initialized as standard Gaussian noise. We then concatenate these latent codes in the same order used during fine-tuning to form the diffusion model input: $\bar{\mathbf{z}}_t = \text{concat}(\mathbf{z}^{mesh}, \mathbf{z}^{gs}, \mathbf{z}_t)$. To generate the fixed image, we iteratively denoise $\mathbf{z}_t$ using the deterministic Denoising Diffusion Implicit Model (DDIM) [26] schedule to perform efficient non-Markovian sampling. The update at each timestep is as follows:

$$\mathbf{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{z}}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(\bar{\mathbf{z}}_t, t), \tag{3}$$

where the clean latent $\hat{\mathbf{z}}_0$ is estimated as:

$$\hat{\mathbf{z}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\bar{\mathbf{z}}_t, t)\right), \tag{4}$$

derived directly from the forward diffusion formulation in Eq. (1). After completing the denoising process, the final fixed image is obtained by decoding the predicted clean latent using the VAE decoder $\mathcal{D}$: $\hat{I}^{fixed} = \mathcal{D}(\hat{\mathbf{z}}_0)$.

3.3. GSFix3D: Diffusion-Guided Novel View Repair

The final stage of our GSFix3D framework lifts the output of the diffusion model, i.e., GSFixer, back into the 3D representation. Thanks to the full differentiability of 3DGS, we can continue optimizing the parameters of the initial 3DGS reconstruction by minimizing a photometric loss between the fixed image $\hat{I}^{fixed}$ and the rendered image I^{gs}:

$$\mathcal{L}_{\text{pho}} = (1-\lambda)\|\hat{I}^{fixed} - I^{gs}\|_1 + \lambda\mathcal{L}_{\text{SSIM}}(\hat{I}^{fixed}, I^{gs}),$$

where λ is a weighting factor, and $\mathcal{L}_{\text{SSIM}}$ denotes the Structural Similarity loss. We also enable adaptive density control during optimization, following [9], to fill in previously empty or under-populated regions.

To reduce inconsistencies in the repaired images and improve global coherence, we further append the repaired views and their corresponding poses to the original captured datasets and then optimize over this augmented dataset for several iterations. Note that we use a sparse set of keyframes recorded during the initial reconstruction phase instead of the full dataset to avoid redundant and time-consuming optimization.

Method	ScanNet++			Replica		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SplaTAM	23.03	0.791	0.311	23.82	0.833	0.267
SplaTAM + DIFIX	23.06	0.789	0.220	22.97	0.790	0.262
SplaTAM + DIFIX-ref	22.79	0.799	0.203	22.97	0.830	0.217
SplaTAM + GSFixer	**25.11**	**0.831**	**0.188**	**25.67**	**0.839**	**0.215**
RTG-SLAM	19.54	0.777	0.341	25.00	**0.860**	0.247
RTG-SLAM + DIFIX	19.43	0.762	0.245	24.02	0.811	0.214
RTG-SLAM + DIFIX-ref	19.29	0.769	0.223	23.89	0.834	**0.193**
RTG-SLAM + GSFixer	**24.80**	**0.824**	**0.204**	**26.27**	0.843	0.228
GSFusion (gs)	24.58	**0.838**	0.308	22.10	0.844	0.296
GSFusion (gs) + DIFIX	24.34	0.818	0.193	21.81	0.772	0.273
GSFusion (gs) + DIFIX-ref	23.83	0.822	0.184	21.91	0.821	0.224
GSFusion (gs) + GSFixer	24.79	0.833	0.196	23.87	0.830	0.251
GSFusion (mesh+gs) + GSFixer	**25.30**	0.837	**0.183**	**25.98**	**0.845**	**0.219**

Table 1. Comparisons of diffusion-based repair methods on the ScanNet++ and Replica datasets. The best result is highlighted in **bold**, and the second-best is underlined. The text inside ( ) indicates the format of the reconstruction used.

4. Experiments

4.1. Experimental Setup

Evaluation Datasets and Metrics. We compare different methods on two challenging benchmark datasets: ScanNet++[38] and Replica[29]. ScanNet++ is a real-world indoor dataset containing high-quality RGB-D data. Each scene includes two separate camera trajectories for training and evaluation, respectively. Following [33], we select four scenes from ScanNet++: 8b5caf3398, 39f36da05b, b20a261fdf, and f34d532901. The Replica dataset consists of photorealistic synthetic indoor scenes with accurate RGB-D imagery. We use consistent trajectories from [43] for reconstruction and fine-tuning. To enable the quantitative assessment of novel views and to evaluate inpainting capabilities, we manually render ground truth novel views from extreme viewpoints with large unobserved regions (see Sec. 6.1 in the supplementary). We use three common metrics to measure rendering quality and fidelity: PSNR, SSIM, and LPIPS. All reported results are averaged over scenes within each dataset.

Baselines. We compare our GSFixer against two variants from [35]: DIFIX and DIFIX-ref. DIFIX is a single-step image diffusion model trained on 80k noisy-clean image pairs curated from real-world datasets. DIFIX-ref extends this setup by incorporating an additional reference view as input, introducing multi-view constraints to enhance performance. In addition to GSFusion, we also include two recent Gaussian SLAM methods, SplaTAM[8] and RTG-SLAM[23], as alternative sources of 3D reconstructions, each exhibiting distinct artifact patterns due to differences in initialization and optimization strategies. We apply the above image repair models to novel view renderings produced by each of these reconstruction methods.

Implementation Details. We adopt Stable Diffusion v2 as our base latent diffusion model, disabling text prompt and applying the fine-tuning protocol described in Sec. 3.1. During training, we use the DDPM noise scheduler with 1000 diffusion steps. For inference, we follow the DDIM scheduler with only 4 steps for accelerated sampling. Considering the difficulty of collecting large-scale real-world training pairs for this task, we first pretrain the modified U-Net (see Sec. 3.1.1) for 6k iterations on two synthetic datasets: Hypersim[24] for indoor scenes and Virtual KITTI [1] for outdoor street environments. We use a batch size of 2 and accumulate gradients over 16 steps to stabilize training with the Adam optimizer. The learning rate is set to 3×10^{-5}. We acquire the geometrically aligned mesh and 3DGS map by running GSFusion and fine-tune the pretrained model separately for each scene. For real scenes from ScanNet++, we fine-tune for 800 iterations. For synthetic scenes from Replica, we fine-tune for 400 iterations. The fine-tuning process typically takes 4 hours for ScanNet++ and 2 hours for Replica. As for 3DGS optimization in GSFix3D, we perform 20 iterations for each repaired image and 50 iterations over the augmented dataset. All experiments are conducted on a single NVIDIA RTX 4500 Ada GPU with 24GB VRAM.

4.2. Results

Table 1 reports quantitative results on ScanNet++ and Replica. For SplaTAM and RTG-SLAM, which output only 3DGS maps, we fine-tune GSFixer exclusively on rendered images from their reconstructions. Despite this constraint, GSFixer consistently outperforms DIFIX and DIFIX-ref across all metrics on ScanNet++, with over 5 dB PSNR gain in the *RTG-SLAM+GSFixer* setting. Qualitative results in Fig. 3 show that, in the RTG-SLAM example, DIFIX and DIFIX-ref leave a large black hole where a window is missing, while GSFixer fills it with plausible content. In the SplaTAM example, baselines leave colorful floaters, whereas GSFixer learns their patterns and removes them.

For GSFusion, which provides both a mesh and a 3DGS map, we introduce a dual-input setting, *GSFu-*

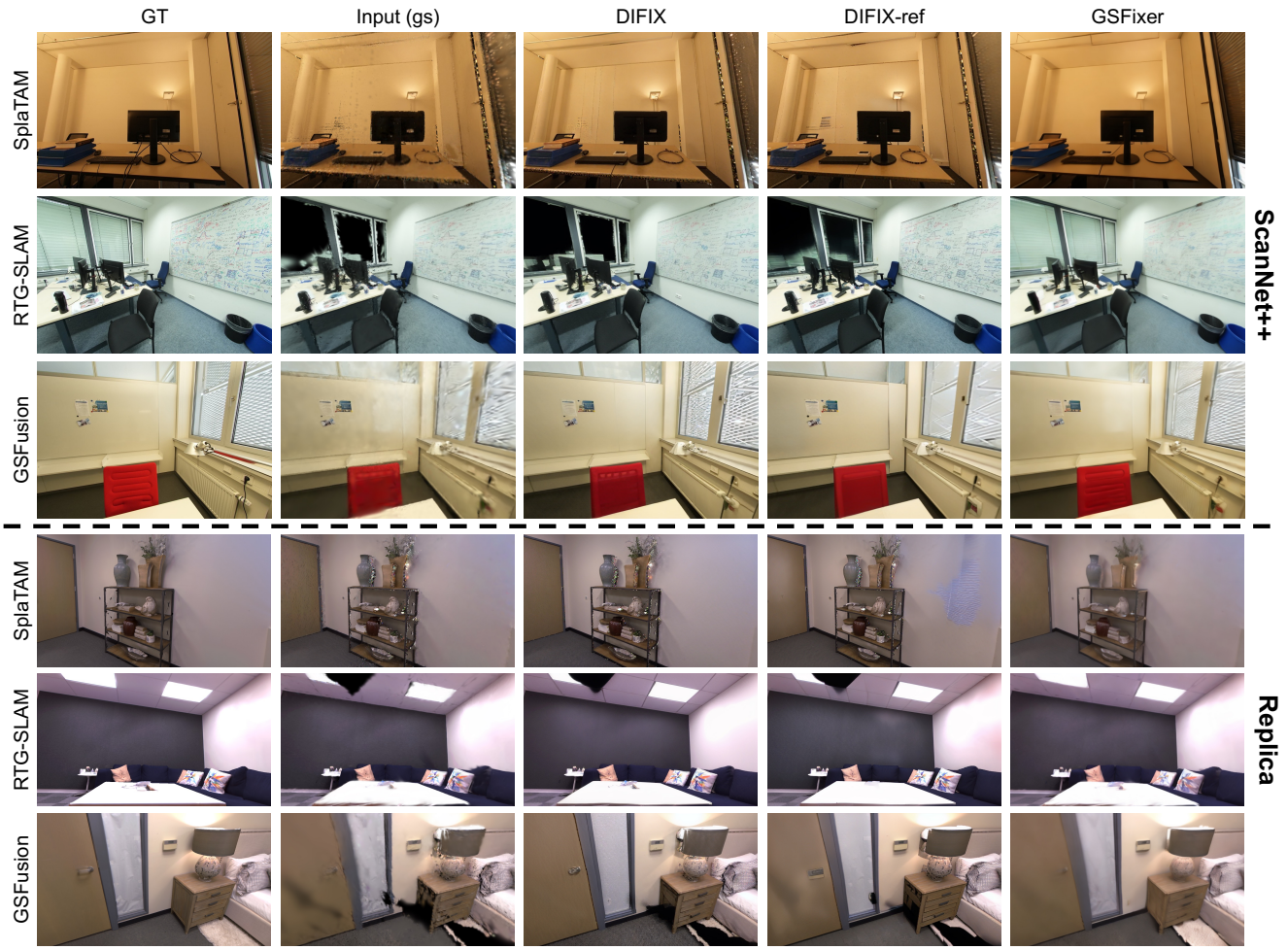

Figure 3. Qualitative comparisons of diffusion-based repair methods on the ScanNet++ and Replica datasets. All examples use only 3DGS reconstructions as the input source. Our GSFixer effectively removes artifacts and fills in large holes, where both DIFIX and DIFIX-ref fail to produce satisfactory results.

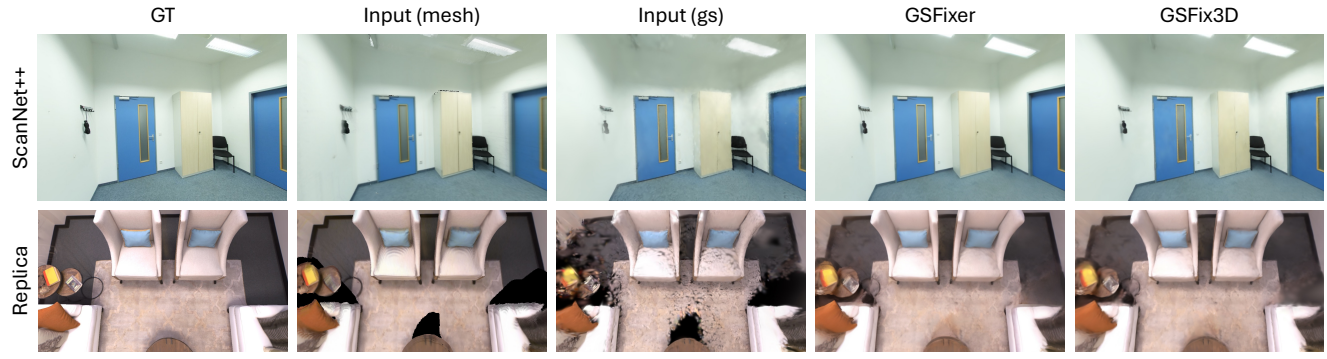

Figure 4. Qualitative comparison between GSFixer and GSFix3D on the ScanNet++ and Replica datasets. Both mesh and 3DGS reconstructions from GSFusion are used as input sources. The 2D visual improvements from GSFixer are effectively distilled into the 3D space by GSFix3D.

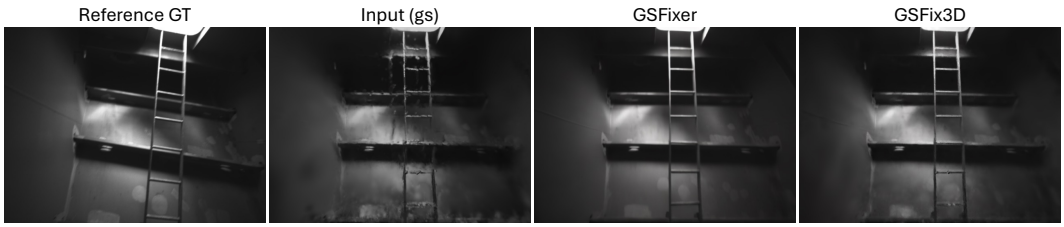

Figure 5. Novel view repair on self-collected ship data. Our method is robust to pose errors, effectively removing shadow-like floaters.

sion(mesh+gs)+GSFixer, that further boosts performance over the single-input variant. We analyze this effect in detail in Sec. 4.4. On the more challenging Replica dataset, where we evaluate on manually selected extreme novel viewpoints with large unobserved regions, GSFixer again outperforms baselines in PSNR and remains competitive in SSIM. The strong inpainting ability of GSFixer is visually evident on the Replica dataset in Fig. 3. Interestingly, DIFIX and DIFIX-ref achieve lower LPIPS scores in some cases, which we attribute to their sharp visual details. This is likely due to their training on 80k noisy-clean image pairs curated from real-world datasets (though the dataset is not publicly available), whereas GSFixer is only pretrained on two synthetic datasets and fine-tuned on a limited amount of clean captured data. We explore additional comparisons in Sec. 7.1.

The overall performance of our GSFix3D framework is reported in Tab. 2. Compared to the direct outputs from GSFixer, lifting the repaired images back into the 3D representation leads to improved perceptual quality thanks to multi-view constraints, as evidenced by higher PSNR and SSIM scores. However, due to the optimization characteristics of the 3DGS representation, the final renderings tend to be less smooth than the 2D generative results, which accounts for the slightly higher LPIPS values. Qualitative examples are provided in Fig. 4. We further apply the full GSFix3D framework to SplaTAM and RTG-SLAM reconstructions in Sec. 7.3 of the supplementary material.

4.3. Real-World Evaluation in the Wild

We collect a stereo sequence inside a ship structure using an Intel RealSense D455 camera. We compute depth maps for the left camera using FoundationStereo [34] for improved quality, and estimate camera poses with OKVIS2 [12]. Since no ground truth is available, the estimated poses may contain errors. Those post-processed data are then fed into GSFusion to obtain an initial 3DGS reconstruction. We fine-tune a GSFixer model using 3DGS renderings as input. Fig. 5 shows a novel view example where shadow-like floaters appear near the ladder due to inaccurate poses. Our method effectively removes these artifacts in 2D and distills the correction back into the 3D representation, demonstrating robustness to common pose errors in real-world data collection, particularly in uncontrolled settings without high-end equipment or precise calibration. Additional real-world results are provided in Sec. 7.4 of the supplementary, including a test on an outdoor scene [41] using a LiDAR-Inertial-Camera Gaussian Splatting SLAM system [11], which further demonstrates the practical adaptability of our method.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓
ScanNet++	GSFusion (gs)	24.58	0.838	0.308
	GSFusion (mesh+gs) + GSFixer	25.30	0.837	**0.183**
	GSFusion (mesh+gs) + GSFix3D	**25.63**	**0.845**	0.238
Replica	GSFusion (gs)	22.10	0.844	0.296
	GSFusion (mesh+gs) + GSFixer	25.98	0.845	**0.219**
	GSFusion (mesh+gs) + GSFix3D	**26.49**	**0.864**	0.252

Table 2. Comparisons of GSFixer and GSFix3D on the ScanNet++ and Replica datasets.

4.4. Ablation Studies

Image Conditions. To analyze the impact of different input image conditions, we evaluate GSFixer under three input configurations: mesh-only, 3DGS-only, and dual-input (mesh+3DGS), with results in Tab. 3. On the synthetic Replica, which provides highly accurate measurements, mesh-based renderings tend to be of higher quality than their 3DGS counterparts from novel viewpoints. As a result, the *GSFusion(mesh)+GSFixer* setting achieves better rendering performance than *GSFusion(gs)+GSFixer*. In contrast, on the real-world ScanNet++ dataset, 3DGS reconstructions outperform mesh renderings due to noisy depth, making *GSFusion(gs)+GSFixer* the better choice. When both images are used together as input, we observe complementary advantages: the dual-input setup leads to improved performance on ScanNet++ and a modest gain on Replica.

Qualitative results in Fig. 6 further highlight this benefit. For example, in a ScanNet++ scene, the mesh-rendered image suffers from geometric inaccuracies along the table edge, while the 3DGS-rendered image shows visual gaps on the table surface. When both are used to condition GSFixer, these issues are effectively mitigated. Similarly, in a Replica scene, the mesh-rendered image exhibits blurry textures on the pillow, and the 3DGS-rendered image contains visible holes on the floor. Combining both inputs allows GSFixer to resolve these artifacts by leveraging strengths

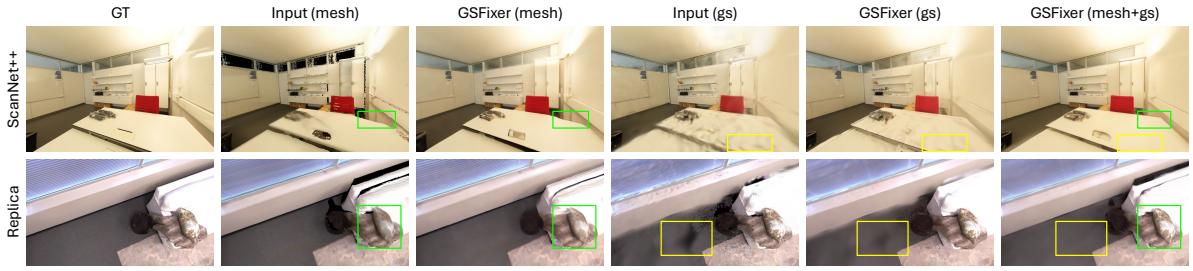

Figure 6. Qualitative ablation of input image conditions on the ScanNet++ and Replica datasets. We compare GSFixer results using three types of inputs rendered from GSFusion: mesh-only, 3DGS-only, and dual-input. The artifacts (highlighted by green and yellow boxes) present in the single-input settings are effectively mitigated with the dual-input configuration.

Figure 7. Qualitative ablation of random mask augmentation on the Replica dataset. We compare GSFixer results fine-tuned with and without our proposed augmentation strategy. The differences in inpainting quality highlight the improved ability to fill large missing regions when augmentation is used.

from each source. Additional experiments on SplaTAM and RTG-SLAM are presented in Sec. 7.2 of the supplementary.

Random Mask Augmentation. To validate the effectiveness of our proposed data augmentation strategy in improving inpainting capability for novel view repair, we conduct an ablation study by disabling the random mask augmentation during fine-tuning on the Replica dataset. We choose Replica for this evaluation due to its challenging novel views with extensive unobserved regions and visible holes. As shown in Tab. 4, GSFixer fine-tuned with random mask augmentation consistently outperforms the variant without augmentation across all metrics. It is also evident in Fig. 7. The 3DGS-rendered image contains a large missing region on the whiteboard. Without random mask augmentation, GSFixer struggles to inpaint the hole even when given an additional mesh-rendered image as a condition. In contrast, our full model with augmentation successfully fills in the missing region with coherent and realistic textures, demonstrating its generalization to real occlusions.

5. Conclusion

GSFix3D raises the bar for novel view repair in 3DGS reconstructions, requiring no massive real data curation or costly pertaining, only minimal fine-tuning on a small set of captured views. By coupling this efficient fine-tuning protocol with a dual-input design that fuses mesh and 3DGS cues, and empowering it with random mask augmentation as the key to strong inpainting performance, the resulting diffusion model, GSFixer, removes artifacts, fills missing regions with plausible detail, and adapts seamlessly to different scenes and reconstruction pipelines. Across diverse and challenging benchmarks, our method consistently outperforms prior diffusion-based approaches, validating its effectiveness, adaptability, and robustness even under pose inaccuracies, underscoring its practicality for a wide range of 3D reconstruction scenarios.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓
ScanNet++	GSFusion (mesh)	17.87	0.750	0.358
	GSFusion (mesh) + GSFixer	24.64	0.823	0.198
	GSFusion (gs)	24.58	**0.838**	0.308
	GSFusion (gs) + GSFixer	24.79	0.833	0.196
	GSFusion (mesh+gs) + GSFixer	**25.30**	0.837	**0.183**
Replica	GSFusion (mesh)	23.20	**0.849**	0.217
	GSFusion (mesh) + GSFixer	**26.61**	0.846	**0.200**
	GSFusion (gs)	22.10	0.844	0.296
	GSFusion (gs) + GSFixer	23.87	0.830	0.251
	GSFusion (mesh+gs) + GSFixer	25.98	0.845	0.219

Table 3. Ablation of image conditions on ScanNet++ and Replica.

GSFusion (mesh+gs)	PSNR↑	SSIM↑	LPIPS↓
+ GSFixer (w/o mask)	23.54	0.830	0.231
+ GSFixer (w mask)	**25.98**	**0.845**	**0.219**

Table 4. Ablation of random mask augmentation on Replica.

Acknowledgement. We gratefully acknowledge support from the EU project AUTOASSESS (Grant 101120732) and TUM MIRMI Seed-Fund project IMMERGENCY. We also thank Jaehyung Jung and Sebastián Barbas Laina for their help with ship data collection and processing, and Helen Oleynikova for her valuable feedback on the manuscript.

References

[1] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 1, 5

[2] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 2

[3] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5354–5363, 2024. 3

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[6] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 3

[7] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *arXiv preprint arXiv:2505.09358*, 2025. 3

[8] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21357–21366, 2024. 2, 5

[9] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2, 4

[10] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 3

[11] Xiaolei Lang, Laijian Li, Chenming Wu, Chen Zhao, Lina Liu, Yong Liu, Jiajun Lv, and Xingxing Zuo. Gaussian-lic: Real-time photo-realistic slam with gaussian splatting and lidar-inertial-camera fusion. *arXiv preprint arXiv:2404.06926*, 2024. 7, 3, 5

[12] Stefan Leutenegger. Okvis2: Realtime scalable visual-inertial slam with loop closure. *arXiv preprint arXiv:2202.09199*, 2022. 7, 1

[13] Xinhang Liu, Jiaben Chen, Shiu-Hong Kao, Yu-Wing Tai, and Chi-Keung Tang. Deceptive-nerf/3dgs: Diffusion-generated pseudo-observations for high-quality sparse-view reconstruction. In *European Conference on Computer Vision*, pages 337–355. Springer, 2024. 2

[14] Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *Advances in Neural Information Processing Systems*, 37:133305–133327, 2024. 2

[15] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18039–18048, 2024. 2

[16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[17] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 1

[18] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011. 2

[19] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[20] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32 (6):1–11, 2013. 2

[21] Helen Oleynikova, Zachary Taylor, Marius Fehr, Roland Siegwart, and Juan Nieto. Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1366–1373. IEEE, 2017. 2

[22] Avinash Paliwal, Xilong Zhou, Wei Ye, Jinhui Xiong, Rakesh Ranjan, and Nima Khademi Kalantari. Ri3d: Few-shot gaussian splatting with repair and inpainting diffusion priors. 2025. 2

[23] Zhexi Peng, Tianjia Shao, Yong Liu, Jingke Zhou, Yin Yang, Jingdong Wang, and Kun Zhou. Rtg-slam: Real-time 3d reconstruction at scale using gaussian splatting. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2, 5

[24] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021. 1, 5

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 4

[27] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3

[28] Frank Steinbrucker, Christian Kerl, and Daniel Cremers. Large-scale multi-resolution surface reconstruction from rgb-d sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3264–3271, 2013. 2

[29] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2, 5, 1

[30] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6229–6238, 2021. 2

[31] Emanuele Vespa, Nikolay Nikolov, Marius Grimm, Luigi Nardi, Paul HJ Kelly, and Stefan Leutenegger. Efficient octree-based volumetric slam supporting signed-distance and occupancy mapping. *IEEE Robotics and Automation Letters*, 3(2):1144–1151, 2018. 2

[32] Navve Wasserman, Noam Rotstein, Roy Ganz, and Ron Kimmel. Paint by inpaint: Learning to add image objects by removing them first. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18313–18324, 2025. 4

[33] Jiaxin Wei and Stefan Leutenegger. Gsfusion: Online rgb-d mapping where gaussian splatting meets tsdf fusion. *IEEE Robotics and Automation Letters*, 2024. 2, 3, 5, 1

[34] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5249–5260, 2025. 7, 1

[35] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26024–26035, 2025. 1, 2, 5

[36] Sibo Wu, Congrong Xu, Binbin Huang, Andreas Geiger, and Anpei Chen. Genfusion: Closing the loop between reconstruction and generation via videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6078–6088, 2025. 2

[37] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19595–19604, 2024. 2

[38] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 5

[39] Zhongrui Yu, Haoran Wang, Jinze Yang, Hanzhang Wang, Jiale Cao, Zhong Ji, and Mingming Sun. Sgd: Street view synthesis with gaussian splatting and diffusion prior. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3812–3822. IEEE, 2025. 2

[40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 1

[41] Chunran Zheng, Qingyan Zhu, Wei Xu, Xiyuan Liu, Qizhi Guo, and Fu Zhang. Fast-livo: Fast and tightly-coupled sparse-direct lidar-inertial-visual odometry. In *2022 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4003–4009. IEEE, 2022. 7, 3, 5

[42] Yingji Zhong, Zhihao Li, Dave Zhenyu Chen, Lanqing Hong, and Dan Xu. Taming video diffusion prior with scene-grounding guidance for 3d gaussian splatting from sparse inputs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6133–6143, 2025. 2

[43] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12786–12796, 2022. 2, 5, 1