This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SIR-DIFF: Sparse Image Sets Restoration with Multi-View Diffusion Model

Yucheng Mao[1,*] Boyang Wang[1,*] Nilesh Kulkarni[1,2] Jeong Joon Park[1]

[1]University of Michigan, Ann Arbor [2]Netflix

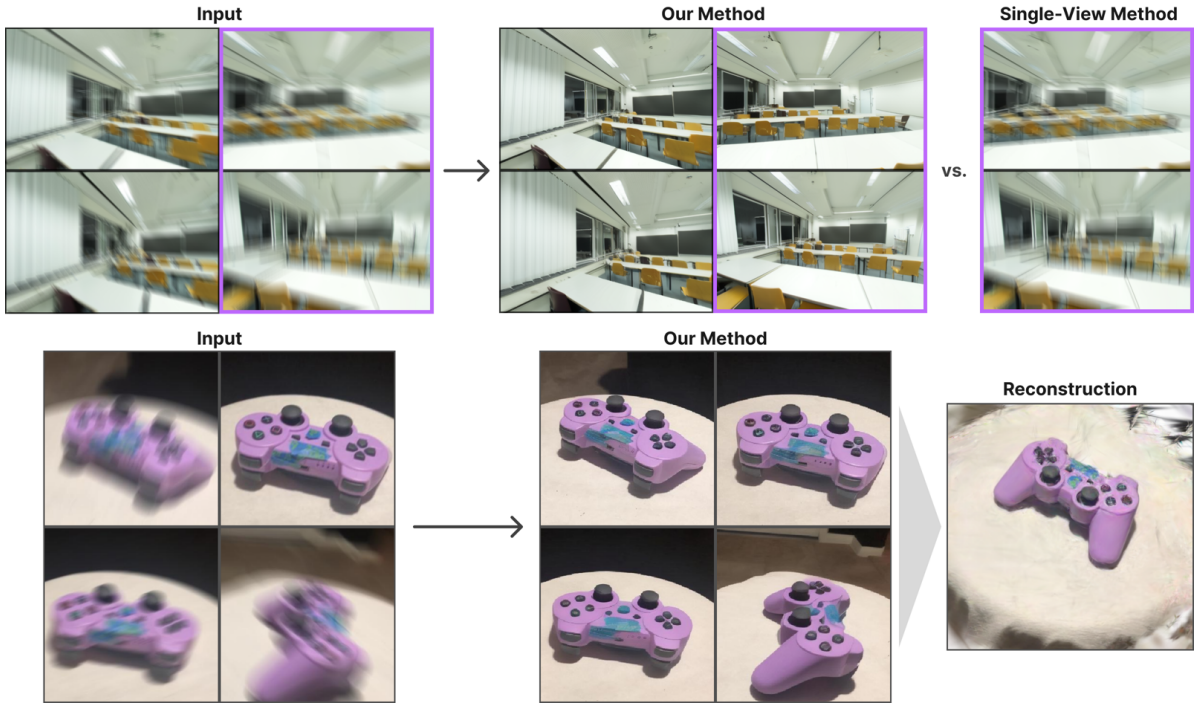


Figure 1. **Sparse View Image Restoration**. Our diffusion model takes multi-view images and jointly enhances their visual quality while maintaining 3D consistency. (Top) Four motion-blurred input images are processed by our method, resulting in sharp outputs that significantly outperform single-view restoration methods as shown in the corresponding purple boxes. (Bottom) Our method can consistently restore multi-view images (4 out of 50 shown), leading to accurate 3D reconstructions.

Abstract

The computer vision community has developed numerous techniques for digitally restoring true scene information from single-view degraded photographs, an important yet extremely ill-posed task. In this work, we tackle image restoration from a different perspective by jointly denoising multiple photographs of the same scene. Our core hypothesis is that degraded images capturing a shared scene contain complementary information that, when combined, better constrains the restoration problem. To this end, we implement a powerful multi-view diffusion model that jointly generates uncorrupted views by extracting rich information from multi-view relationships. Our experiments show that our multi-view approach outperforms existing single-view image and even video-based methods on image deblurring and super-resolution tasks. Critically, our model is trained to output 3D consistent images, making it a promising tool for applications requiring robust multi-view integration, such as 3D reconstruction or pose estimation. Project website: https://myc634.github.io/sirdiff/

1. Introduction

Image restoration approaches [8, 16, 29, 53] typically process degraded images in isolation, confronting the challenge of the original scene content. Yet in practice, we frequently capture multiple photographs of meaningful moments, yielding a natural collection of views of the same scene, albeit some degraded. This common behavior presents an untapped opportunity: leveraging multiple captures to enhance image

* Equal contribution

quality and better preserve our visual memories.

In this work, we present a novel perspective on image restoration by reformulating it as a multi-view collaborative restoration task. We leverage our key insight that multiple corrupted photographs of the same scene contain complementary information that can aid superior denoising outcomes. The concept of multi-image fusion for quality enhancement has been explored in classical super-resolution techniques where sub-pixel shifted images are combined to reconstruct fractional pixel information [19]. Our approach revisits this paradigm in a fundamentally wider context by using neural attention mechanisms to implicitly fuse overlapping information from general multi-view images to enhance even severe aberrations.

Notably, our multi-view restoration framework has particular significance for 3D computer vision applications. Many real-world tasks, from robotic SLAM to novel-view synthesis, rely fundamentally on multi-view imagery of a scene. When restoring degraded images within such multi-view sets, maintaining geometric consistency across views is crucial to preserve the underlying 3D scene assumptions. Traditional single-view restoration methods, inevitably introduce inconsistencies that can compromise these geometric constraints. In contrast, our restoration approach naturally enforces view consistency through learned neural priors.

We implement our approach, dubbed SIR-Diff (**S**parse **I**mage **R**estoration with **Diff**usion), through an innovative multi-view diffusion model architecture. Our choice of diffusion models is motivated by their natural handling of uncertainty in image restoration. We extend the UNet-based Stable Diffusion [35] architecture to process multiple views jointly by introducing three key components: (1) a unified encoder for diverse degraded inputs including low-resolution, blurred, and regular RGB images, (2) a hybrid 2D-3D convolutional structure derived from SVD (Stable Video Diffusion) [4, 5], and (3) a 3D self-attention Transformer that enables cross-view attention across all spatial and multi-view tokens.

We evaluate and show zero-shot performance on Scannet++ [61], ETH3D [38], and CO3D [32] for multi-view deblurring and super-resolution tasks and their 3D downstream applications. SIR-Diff consistently outperforms existing single-image and video restoration methods for multi-view restoration tasks in both image quality and our proposed 3D view consistency metrics. For downstream applications, we show that our multi-view restoration strategy significantly enhances 3D reconstruction quality, as evidenced by improved novel-view synthesis accuracy and feature-matching success rates. Beyond image restoration, our architecture shows promise in other multi-view tasks, notably achieving competitive depth estimation accuracy and geometric consistency compared to leading single-view methods.

2. Related Works

Image Diffusion Models. Diffusion Models (DM) show huge popularity in all domains: illustrations [9, 35, 51, 65], video creation [4, 5, 30], 3D modeling with limited input views [27, 41, 45], 4D generation [2, 3, 26], and *etc*. Most models are fine-tuned from the pre-trained weight instead of training from the beginning. One of the most widely-used pre-trained Diffusion Models is Stable Diffusion 2.1 [35], a UNet-based [36] architecture.

Multiview-Image Diffusion models. In the image-to-3D task [24], Diffusion Model is a powerful generative prior to enriching details for the novel (unseen) view synthesis [6, 17, 44, 55, 56]. Methods like MVDream [41], CAT3D [13], Zero123++ [40] jointly generate multi-views simultaneously by extending the image diffusion model (SD2.1) with a 3D self-attention layer. However, these works ignore the exploration of the convolution layer architecture in the UNet-based Diffusion Model. In this paper, we explore the methodology of applying the 3D self-attention layer to the 3D restoration task.

Image and Video Restoration. These methods aim at converting a degraded input image or video into a high-quality form. Traditional tasks include Super-Resolution [23, 53, 57, 62, 68], Deblurring [22, 25, 31], Denoise [10, 59], Compression Artifact Reduction [8, 46, 47], Inpainting [7, 58], Burst Image Deblurring [1, 14], and *etc*. These domains have achieved milestones in the 2D image and video domain [28]. However, the study on the restoration in 3D is insufficient. The restoration consistency of the texture and geometry from different views is hard to maintain when independently restoring them.

3D Reconstruction from Corrupted Inputs. 3D Gaussian Splatting (3DGS) [21, 33, 54, 63] transforms input images into explicit-represented 3D Gaussian blobs with continuous optimization to reconstruct the large 3D scene. However, these models heavily rely on the condition of the input images. Motion blurring [50, 67] or low-resolution [12, 39] inputs that lack high-frequency and sharp details may greatly deteriorate the reconstruction quality or even fail at the initialization stage. To resolve this, methods like Bad-Gaussian [39] and SuperGaussian [67] depend on dense view input to iteratively optimize the degraded artifacts. Further, under heavily deteriorated inputs, these methods cannot recover high-quality outputs when they lack a strong generative prior [39] such as a pre-trained Diffusion Model.

3. Method

In this section, we discuss our proposed approach, SIR-Diff, which can generate and restore sparse image sets from 3D scenes or objects conditioned on their degraded inputs. Specifically, we showcase our approach to two challenging tasks: motion deblurring and super-resolution. In the

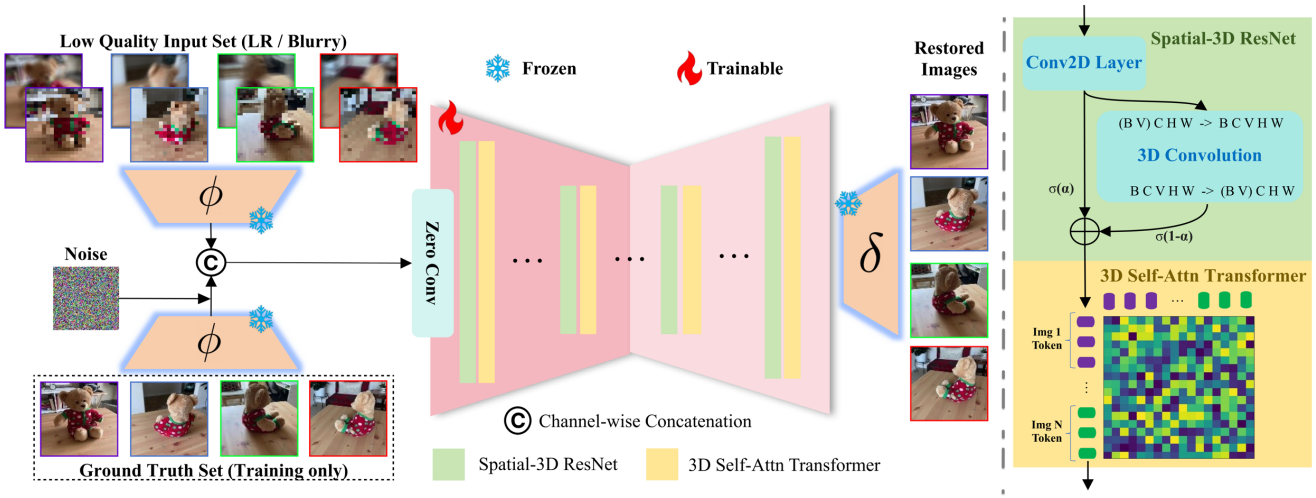


Figure 2. **Method Overview.** Our model is a UNet-based latent diffusion model that takes as input a degraded sparse image set and outputs a set of restored consistent images. The core of our approach is a diffusion model that jointly denoises image latents across multiple views by relying on our spatial-3D ResNet and 3D Self-Attn transformer. We use pre-trained encoder-decoders from SD2.1 [35] and train our denoising UNet on those encoded latents. We show view-consistent results for the task of **Deblurring** and **Super-Resolution**.

following sections, we discuss our problem setup in § 3.1, multi-view image restoration diffusion model in § 3.2, and our proposed model architecture in § 3.3.

3.1. Multi-View Generation Setup

Our goal is to restore sparse view images from degraded inputs. Typical methods to solve this problem work either on a single image [31, 35, 57] or require a coherent video [25, 68] as input. Toward the goal of sparse view image restoration, we propose a generative model that can respect the geometrical priors to perform the task of image restoration from blurry images or recovering high-resolution image data from its low-resolution counterpart. Any single-image generative model that performs independent inference on each image cannot ensure consistency over the generated image set.

To alleviate these issues, we formulate this as a generation problem to infer all the required views of the scene by modeling the joint distribution using a multi-view diffusion model. Our model takes a degraded image set as input and recovers a high-quality self-consistent image set.

3.2. Multi-View Image Restoration Diffusion Model

Our model aims to restore images when conditioned on a corrupted image set $\{I^c\}_{i=1}^N$. The model's primary objective is to restore a self-consistent high-quality image set. We do this by using a multi-view image restoration diffusion model that can jointly restore the image set. In training, we start from a dataset of non-blurry / high-resolution images $\{I^r\}_{i=1}^N$ and apply the forward diffusion process. Our input representation to the N-view diffusion model is $\mathbf{x} = \{I_1^r, \ldots, I_N^r\}$. Specifically, $\mathbf{x}^0$ corresponds to a clean signal at the start of the forward diffusion process at step 0.

Our generative model learns to reverse this noising process by predicting the added noise at every timestep of the diffusion process.

3.3. Model Architecture

We build our joint image restoration diffusion model based on a latent diffusion model [35]. Fig. 2 shows an overview of our method. The goal of our diffusion model is to denoise image latents such that a pre-trained decoder can convert them to high-quality images. When simultaneously denoising multiple image latents, it is important that latent representation remain consistent across views. We enable this by allowing the diffusion model to attend to latents from other views.

3.3.1. Latent Diffusion Models for Restoration

A Latent Diffusion Model [35] architecture, such as Stable Diffusion 2.1 (SD2.1), consists of two parts a) encoder and decoder and b) denoising UNet.

During inference, our latent denoising UNet progressively refines pure noise into a clean latent representation, conditioned on the degraded image set, through several denoising steps. Once the latent representation is clean, a decoder translates it into the restored image set.

Similarly, during training, our diffusion model is optimized to learn the denoising process of the latent representation, conditioned on the degraded input sets.

Encoder-Decoder. We initialize our encoder with the weights from an SD2.1 [35] and are frozen when we train our model. The encoder, ϕ, is used to map our images to latent vectors which are used to learn the UNet-Style diffusion model. The decoder, δ, converts the denoised latents from the diffusion model to pixels in image space.

UNet. Our latent denoising UNet, $\mathbf{s}_\theta(\cdot)$ with parameters θ, considers as input a noisy latent $\mathbf{x}^k$, encoded latents for

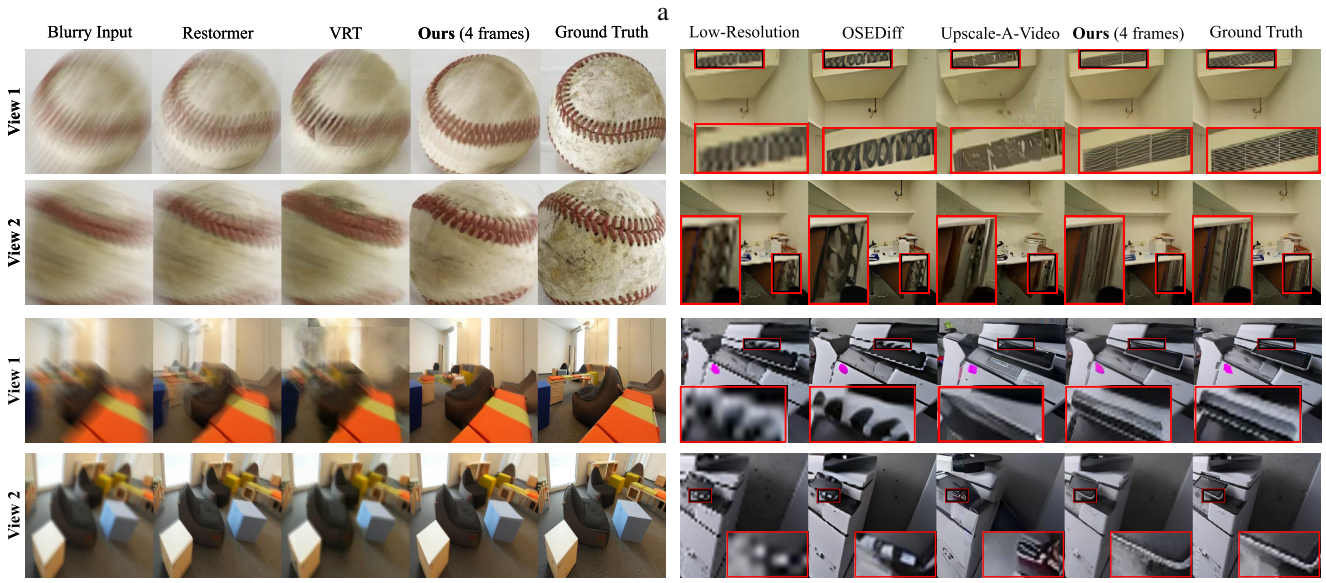


Figure 3. **Qualitative Comparisons on Motion Deblurring and Super-Resolution.** The left side is the Motion-Deblurring and the right side is the Super-Resolution results. We refer readers to supplementary for more visualizations. **Zoom in for the best view.**

degraded images, $\{I^c\}_{i=1}^N$ and the diffusion noise k to output an of estimate the noise, $\hat{\boldsymbol{\epsilon}}$, as shown in Eq. (1):

$$\hat{\boldsymbol{\epsilon}} = \mathbf{s}_\theta(\mathbf{x}^k, \phi(\{I^c\}_{i=1}^N), k). \tag{1}$$

We supervise the outputs of our UNet to match the noise added during the forward diffusion process using Eq. 2:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0,\epsilon\sim\mathcal{N}(0,I),k\sim\mathcal{U}(T)} \left\|\epsilon - \hat{\epsilon}\right\|_2^2. \tag{2}$$

3.3.2. Adapting Diffusion Model to multiple views

Starting with the image diffusion model, we expand a new dimension to represent image sets from different views. Most of the time, batch and views are merged into one dimension and split when needed. We encode the degraded image set with the same VAE encoder as the target noisy latents. Noisy latents and the condition latents are integrated by channel-wise concatenation. Then, due to channel number change, we replace the original convolution with a new zero-initialized convolution layer [65] to adapt the new number of channels input.

3.3.3. Spatial-3D Integrated UNet Architecture

Typical latent diffusion models such as Stable Diffusion [35] do not consider denoise multiple images simultaneously and often generate individual images that lead to inconsistent image generations between multi-views. Our solution to remove this inconsistency is to denoise all the views jointly by allowing the UNet to look at the intermediate latent representations from different views. As shown in the right side of Fig. 2, our architecture is composed of a Spatial-3D ResNet and a 3D self-attention Transformer.

Spatial-3D ResNet. Image diffusion model like SD2.1 only has a 2D convolution layer in the ResNet, which makes it hard to perform well in tasks that require 3D understanding. We want to enhance the 3D performance of the ResNet of the image diffusion model by introducing a 3D convolution layer to better capture geometry information as shown in Fig. 2. Training from raw will be unstable and time-consuming, but we find that the video diffusion model that is performing the image-to-video task already has a pre-trained 3D convolution layer. Understanding temporal similarity in the time dimension and understanding 3D similarity in the spatial dimension share similarity. By the experiment, we empirically find that merging two different weights into one is feasible and can lead to better performance. Hence, we initialize the 2D Convolution layer with image Diffusion Model weight (SD2.1) and the 3D convolution layer with video Diffusion weight (Stable Video Diffusion [4]). The output of each layer will be blended as follows: $O_{\text{ResNet}} = \sigma(\alpha)\times O_{2D} + \sigma(1-\alpha)\times O_{3D}$, where σ is the sigmoid function and α is usually taken as 0.5 following SVD [4]. O_{2D} is the output of 2D ResNet and O_{3D} is the output of 3D ResNet layer.

3D Self-Attention Transformer. Furthermore, we adopt the 3D multi-view consistent self-attention Transformer as shown in Fig. 2, similar to previous works [13, 41]. Our self-attention layer considers image latents from all the N views and performs self-attention. After ResNet, the model patches the latent to create p patch token per view, and in total, we have $N \times p$ tokens that are part of the attention layer. Each token calculates the attention score to all $N \times p$ tokens as shown on the right side of Fig. 2. This feature enables each token to be accessible to all conditional image sets given at the maximum level. Further, by this design, our model, in the inference stage, is scalable to a variable number of input frames with stable performance even though

Table 1. **Sparse-View Deblurring Comparisons.** We show quantitative results on three datasets and evaluate metrics as in § 4.1. VConsis. (visual consistency) is multiplied by 100. Across all methods, SIR-Diff is the most visually consistent as compared to all the baselines. On other standard metrics, we also outperform baselines. SIR-Diff (1 Frame) is a single image version of our method that does not depend on multiple views. The best is **highlighted**, and the second is underlined.

Dataset	**Scannet++** [61]			**ETH3D** [38]			**CO3D** [32]		
Method	FID ↓	LPIPS ↓	VConsis.↓	FID ↓	LPIPS ↓	VConsis.↓	FID ↓	LPIPS ↓	VConsis.↓
PromptIR [31]	81.28	0.248	7.81	150.34	0.537	10.72	166.3	0.350	12.89
Restromer [64]	49.72	0.232	6.52	73.47	0.330	9.59	144.8	0.426	11.00
VRT [25]	134.5	0.371	7.67	160.4	0.530	11.13	170.4	0.440	12.07
SIR-Diff (1 Frame)	81.58	0.247	6.45	109.6	0.350	9.55	137.9	0.357	10.36
SIR-Diff	**40.09**	**0.160**	**5.75**	**67.31**	**0.263**	**9.43**	**85.67**	**0.275**	**9.30**

we train with a handful of images per set.

However, attention communication between views leads to heavy computes. To resolve this, we only inject this module into the low-resolution layers (similar to CAT3D [13]), where we also ignore the biggest resolution layer for the UNet Encoder and Decoder. Further, we completely take away the cross-attention module across all layers instead of putting empty string inputs as placeholders like previous works [20]. We found that both of these changes help faster convergence, less training time, and better numerical results under the same number of training iterations.

4. Experiment

We evaluate SIR-Diff on the task of *sparse view* motion deblurring and super-resolution in a zero-shot setting § 4.2, and discuss our deblurring results in § 4.4 and super-resolution results in § 4.5. We discuss implementation details in § 4.2, and now we start by discussing evaluation metrics in § 4.1. Please refer to supplementary materials for the ablation study.

4.1. Metrics

We use several metrics to evaluate the performance of all the methods on the task of motion deblurring and super-resolution. Since our approach is generative in nature we borrow metrics from the literature that are good at evaluating generative models such as *Frechet Inception Distance (FID)* [15], *Learned Perceptual Image Patch Similarity (LPIPS)* [66]. In addition to these, we evaluate our super-resolution results on PSNR and SSIM similar to methods used in this literature [57, 62, 64, 68]. We also propose a new metric to evaluate view consistency between pairs of generated views.

Visual Consistency (VConsis). This metric evaluates self-consistency within the generated view set by measuring the perceptual similarity of corresponding patches between the two views. Please refer to the supplementary for additional details.

4.2. Implementation Details

We implement our latent diffusion model based on pre-trained UNet and auto-encoder from SDv2.1 as part of diffusers. We train our models on two synthetic datasets Hypersim [34] and TartanAir [52] with image resolution resized to 640×480. We randomly select the order of image sets to maintain permutation invariance in training and inference. We utilize ground truth geometry and pose data to pre-compute the overlap region between each pair of frames. We provide additional details on model training, dataset, and view selection in the supplemental.

4.3. Zero Shots Datasets For Evaluation

As all our baselines are training on different datasets, we perform a zero-shot evaluation across a diverse set of real-world datasets to keep a fair comparison. *Zero-Shot* means rather than split the dataset into training-evaluation splits, we chose some dataset for training and some other out-of-domain for evaluation. Our evaluation datasets consist of indoor and outdoor scenes along with object-centric datasets.

ScanNet++ [61]. This is a dataset of 460 indoor scanned readl 3D scenes. We evaluate all the methods on the scenes from the official test set of 50 scenes, by sampling 200 image sets with four images per set. We highlight the zero-shot performance of our method trained on synthetic indoor data transfers well to real indoor scene data.

ETH3D [38]. This dataset provides high-resolution images of indoor and outdoor scenes, with accurate depth maps captured via a LiDAR. We evaluate all the 454 samples from the ETH3D dataset.

CO3Dv2 [32]. This is an object-centric dataset containing object-level videos across 51 categories of household objects. This dataset has pose annotations from COLMAP [37] provided for all video sequences. We use the official single-sequence subsets for evaluation and randomly sample 200 image sets among them.

For all these three datasets, please refer to the supplementary material for the detail of view-selection strategy

4.4. Motion Deblurring (MD)

Our motion deblurring (MD) method targets solving arbitrary motion paths in the wild. We target at restoring original images from the set of sparse view motion blur inputs. Our paired training dataset is prepared by applying the motion-

Table 2. **Sparse-View Super-Resolution Comparisons.** VConsis. (visual consistency) is multiplied by 100. **SIR-Diff (1 Frame)** is a single image process form like other image diffusion models. Ours, in the inference stage, can adapt an arbitrary number of views. The best is **highlighted** and the second is underlined.

Dataset	**Scannet++** [61]				**ETH3D** [38]				**CO3D** [32]			
Method	PSNR ↑	SSIM ↑	LPIPS ↓	VConsis.↓	PSNR ↑	SSIM ↑	LPIPS ↓	VConsis.↓	PSNR ↑	SSIM ↑	LPIPS ↓	VConsis.↓
SAM-DiffSR [48]	24.13	0.216	0.758	6.94	19.16	0.317	0.495	8.28	23.90	0.246	0.626	10.96
OSEDiff [57]	25.22	0.808	0.154	5.51	21.93	**0.621**	0.264	8.62	24.50	**0.658**	0.277	8.86
UAV [68]	23.23	0.677	0.354	9.01	20.39	0.506	0.413	9.26	22.73	0.566	0.400	11.82
SIR-Diff (1 Frame)	25.85	0.792	0.164	5.70	21.97	0.593	0.221	7.90	24.35	0.641	0.301	9.07
SIR-Diff	**27.38**	**0.837**	**0.099**	**5.01**	**22.58**	**0.621**	**0.201**	**7.60**	**24.53**	**0.658**	**0.210**	**7.85**

blurring kernel on the clean images from the training datasets based on the codebase [1] inspired by DPS [7]. The blur kernel size is chosen from a normal distribution with mean 85px and std 12.75px.

Further, the intensity of the blur is randomly sampled in the range $[0, 1]$. The blur kernel for each image in the set is sampled independently. For evaluation, we randomly sample 4 images per view-set from the pre-computed image set and require all the methods to remove motion blurring from them.

4.4.1. Baselines

We compare against several baselines that can restore motion blurring from single-image or videos.

Restormer [64]. This is a single-image motion deblurring method that does not have the benefit of jointly deblurring across a sparse view image collection. Our method outperforms this on single image deblurring and sparse view deblurring across all metrics.

PromptIR [31]. This is all-in-one restoration method trained to deblur. We use their pre-trained model and compare the performance of their method on sparse view deblurring.

VRT [25]. This is a video-oriented restoration Transformer model trained with deblurring capability. Our method outperforms this on sparse view deblurring across all metrics. Fair comparisons on video datasets are in the supplemental.

4.4.2. Results

We evaluate the performance of baselines and our method quantitatively and qualitatively on three datasets.

Qualitative. As shown in the left side of Fig. 3, our SIR-Diff show rich details and sharp deblurring restoration quality than other methods on both object level [32] and scene level [61]. This is thanks to our architectural design information from different views can be aggregated and constructive to help the other view in attention and residual layers.

Quantitative. We assess our results using the metrics described in § 4.2, with outcomes summarized in Tab. 1. Our method outperforms all the methods on the three datasets on FID and LPIPS metrics. Restormer [64], a single-image method, lacks the ability to leverage multi-view information to enhance output quality. In contrast, our method uses 3D self-attention to achieve superior results.

We also outperform the video motion deblurring method, VRT, as they are incapable of handling sparse views well by motion estimation and require continuous and smooth video-captured views to fuse frame features. In the supplement, we provide a direct comparison with video methods quantitatively on continuous videos.

4.5. Super-Resolution (SR)

We evaluate our method on the 4x sparse-view super-resolution task. Our paired training dataset is prepared by first downsampling the high-resolution image with a scale factor of 4 and then upsampling it back to the original resolution to introduce a low-resolution effect. The resize algorithm we use is bicubic.

4.5.1. Baselines

We compare against several diffusion model baselines that handle single-image and video super-resolution.

SAM-DIFFSR [48]. This is an image diffusion model-based methodology, which targets classical super-resolution tasks like ours. Our method outperforms this on sparse view restoration on all the metrics.

OSEDiff [57]. This is a single-image diffusion model-based super-resolution model. OSEDiff reaches SOTA performance on the 2D image super-resolution.

Upscale-A-Video [68]. This is a video diffusion model-based super-resolution method. We input it with a sequential video frame (in total 4 for each scenario) but a larger motion gap compared to ordinary videos. We also provide a fair comparison between Upscale-A-Video and our method on video datasets in the supplement.

4.5.2. Results

We evaluate the performance of baselines and our method quantitatively and qualitatively on three datasets.

Qualitative. As the red rectangular box of Fig. 3 shows, for shadow artifacts and severe aliasing effects (see the second example), Upscale-A-Video [68] fails and OSEDiff [57] restoration is unfaithful and inconsistent for different views compared to ours. Moreover, our model can restore

[1] https://github.com/LeviBorodenko/motionblur

Table 3. **Gaussian Splatting Reconstruction from Blurry Images.** Under serious motion blurring, COLMAP[37] makes it hard to solve correct camera pose and SfM points, which leads to bad results on both vanilla 3DGS and BAD-GS. With the help of SIR-Diff, we can achieve restore better results and be robust to difficult motion blurring effects. **KS** refers to kernel size with normal distribution $[\mu, \sigma]$. **Inten.** refers to the blurring intensity range that is sampled uniformly. We randomly select 3 objects in CO3Dv2 [32] dataset and apply motion blur on clean images to conduct this experiment.

Difficulty	Medium			Hard		
Params	KS:[30, 10.2], Inten:[0, 0.4]			KS:[45,14.85], Inten:[0, 0.5]		
Method	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
3DGS [21]	13.47	0.610	0.607	10.42	0.421	0.801
BAD-GS [67]	10.06	0.537	0.687	8.061	0.385	0.948
Resotrmer [64]+BAD-GS [67]	25.77	0.616	0.356	23.88	0.560	0.317
SIR-Diff + BAD-NeRF [49]	23.60	0.503	0.452	22.47	0.439	0.485
SIR-Diff + BAD-GS	**26.11**	**0.661**	**0.250**	**25.33**	**0.644**	**0.277**

images with better geometry alignment as shown in the first example.

Quantitative. We use PSNR, SSIM, LPIPS [66] and our proposed visual consistency § 4.1 metrics. The result is shown in Tab. 2 to show that SIR-Diff outperforms all the existing baselines on all datasets. Compared to the image diffusion method, OSEDiff[57], our spatial-3D Resnet and 3D self-attention Transformer help each latent token to access information from all views which leads to successful multi-view restoration. Upscale-A-Video also employs a similar 3D ResNet architecture and cross-image attention mechanism; however, our model does not rely on conditioning with optical flow, which is challenging to obtain accurately in sparse views.

4.6. Down Stream Applications

We show several downstream applications of sparse multiview deblurring and super-resolution. Multi-view sparse images are used for tasks of finding **correspondences**, **pose estimation**, and **3D reconstruction**. The performance of these methods is often dependent on the input image quality. If the underlying image quality is degraded, the output of these methods is poor. Blindly restoring these images with off-the-shelf single-image methods leads to sub-optimal results because restoration from these methods lacks view consistency. We show that, with SIR-Diff, we can consistently restore image quality and help alleviate the performance gap on tasks requiring multiview sparse views.

4.6.1. Estimating Correspondences

Under severely degraded inputs it is hard for methods like LoFTR [43] to estimate correspondences. As shown in Fig. 4, (a) we show qualitatively at low-resolution LoFTR struggles to find correct correspondences. Moreover, since single-view image restoration models lack multi-view consistency this does not solve the issue. SIR-Diff can not only restore the image quality but since it is view consistent can improve the correspondences for off-the-shelf methods like LoFTR [43]. Quantitatively, as shown in Tab. 4, we also see that LoFTR can find over 380 additional correspondences as compared to the best image-based method on deblurring and on average 15 for super-resolution.

Table 4. **Feature Matching Correspondence.** We compute the correspondences using off-the-shelf LoFTR [43] on the restored images with single-view methods and SIR-Diff. Both in SR and deblurring cases, our method can enable LoFTR to find more correspondences (# of correspondences found).

	Deblurring	SR
GT Images	1979.5	1979.5
Degraded Images	640.1	1374.0
Single-View Methods [57, 64]	1262.6	1478.0
SIR-Diff	**1621.3**	**1494.2**

4.6.2. 3D Reconstruction on Degraded Images

Methods such as BAD-Gaussians [67] (BAD-GS) aim to perform 3D reconstruction from motion-blurry images, but they depend on running COLMAP [37] on blurry images to get pose initialization and sparse 3D geometry. We show in Tab. 3 that with increasing blurring effect, BAD-GS performance drops significantly while using SIR-Diff with BAD-GS leads to a much steady performance. In Fig. 5, we show a hard case of motion blurring that leads to failure of BAD-GS [67] which can be solved if we restore the image using SIR-Diff first and then do BAD-GS. We refer to supplementary for details.

4.6.3. Sparse-View 3D reconstruction

InstantSplat [11] represents a leading approach for sparse view Gaussian Splattings (GS). However, reconstructing a scene under sparse viewpoints with low-quality observations remains challenging. In Fig 6, we show an example of applying InstantSplat on an image restored using one of the baselines compared to restoration using SIR-Diff. In Tab. 5, we report the average performance of all baselines on the selected 3 scenes from **Scannet++[61]** and **CO3Dv2[32]** and build a 3DGS using InstantSplat [11]. We evaluate the GS with standard Novel View Synthesis metrics. Additional details can be found in the supplementary.

4.7. Depth Prediction for Sparse Views

We also apply our model architecture to the task of improving sparse multiview depth estimation. The only thing we edit is changing the target latent to depth image sets and the conditional latent to RGB image sets. Following Marigold [20], we use the ground truth depth label provided by the dataset for supervision signal during training.

Metrics. We evaluate the performance of methods on three metrics AbsRel and δ_1 as defined by Silberman *et al*. [42]. In the case of MDE methods, we additionally align all the predicted depth maps to find the most optimum scale and bias to make them consistent. Similar to the view consistency

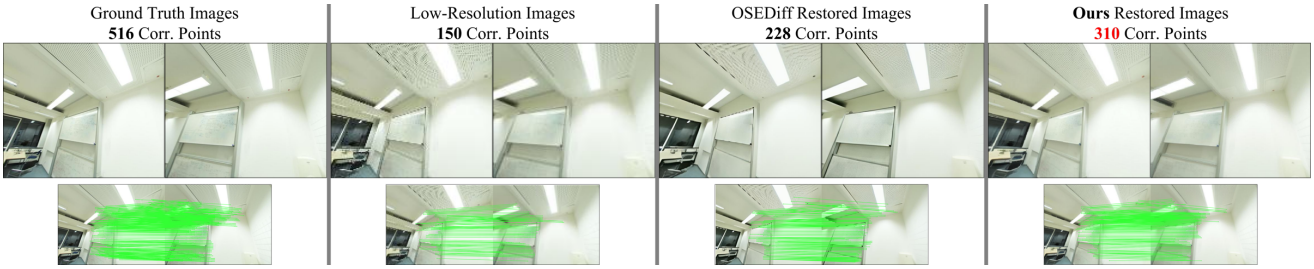


Figure 4. **Feature Matching Downstream Application.** Low-Resolution and Single-image-based SR models like OSEDiff [57] find less correspondence points than our SIR-Diff that can restore multiview images consistently. SIR-Diff recovers the highest number of correspondences (310) as compared to other baseline restoration methods.

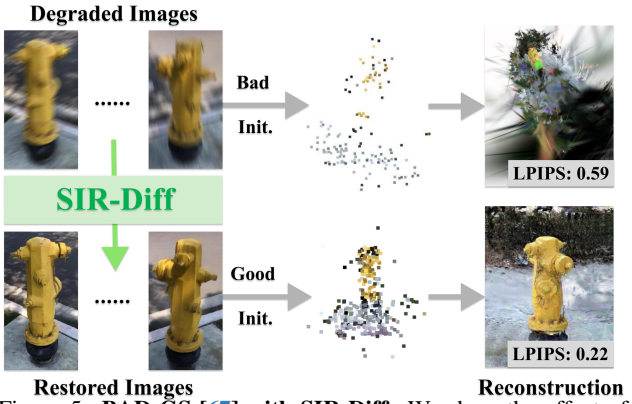


Figure 5. **BAD-GS [67] with SIR-Diff.** We show the effect of using SIR-Diff to help BAD-GS recover from catastrophic failure. BAD-GS (Top) output with a very high LPIPS↓ of 0.59 *vs*. 0.22 BAD-GS output with image-restoration using SIR-Diff (Bottom) leads to a significantly higher-quality 3D reconstruction.

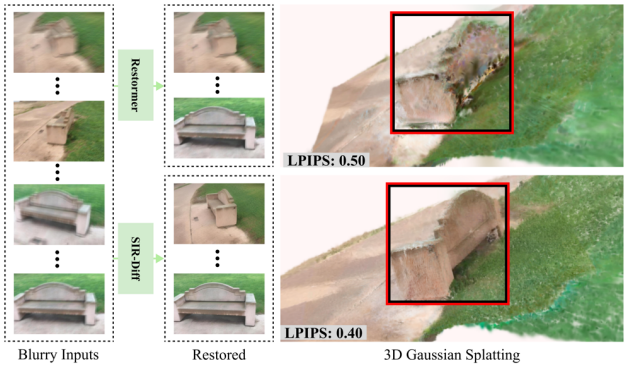


Figure 6. **InstantSplat Downstream Application.** 3D Reconstruction is done by InstantSplat [11]. SIR-Diff that can restore multiview images consistently leads to better visual and numerical reconstruction results than other single-image methods, Restormer [64].

metrics for image restoration, We evaluate **Geometric Consistency** (GConsis.) by calculating the distances between corresponding points between two different views under the camera coordinate system. Additional details on this metric are in the supplemental.

Baselines. We evaluate against several monocular depth estimation (MDE) methods such as DepthAnythingV2 [60], DepthCrafter [18], and Marigold [20].

Table 5. **Quantitative Results of Super-Resolution and Motion-Deblurring Reconstruction on InstantSplat [11].** The result is an average of 3 scenes per dataset (individual results in supp). Using SIR-Diff results in higher quality inputs to InstantSplat compared to OSEDiff [57] and Restormer [64].

	Scannet++ [61]			CO3D [32]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GT Image	31.86	0.940	0.096	23.94	0.757	0.243
Super-Resolution						
OSEDiff [57]	24.53	0.799	0.274	20.25	0.549	0.379
SIR-Diff	**26.20**	**0.827**	**0.233**	**21.78**	**0.578**	**0.357**
Motion-Deblurring						
Restormer [64]	22.02	0.789	0.309	20.14	0.579	0.465
SIR-Diff	**26.53**	**0.842**	**0.231**	**21.60**	**0.614**	**0.384**

Table 6. **Quantitative comparisons on image and video depth models.** The best is **highlighted** and the second is underlined. GConsis. refers to geometric consistency and the unit is meter.

Dataset	Scannet++ [61]			ETH3D [38]		
Method	AbsRel ↓	$\delta 1$ ↑	GConsis.↓	AbsRel ↓	$\delta 1$ ↑	GConsis.↓
DepthAnyV2 [60]	9.270	93.35	0.177	13.20	86.52	1.932
DepthCrafter [18]	21.65	89.96	0.751	**6.662**	**95.05**	1.701
Marigold [20]	11.10	89.92	0.162	7.581	94.46	0.410
SIR-Diff	**9.011**	**93.54**	**0.129**	6.663	**95.05**	**0.339**

Results. As shown in Tab. 6, our multi-view depth model is competitive with single image depth methods on the AbsRel, and δ_1 metrics are sometimes better. Meanwhile, our model outperforms these methods on the geometric consistency metric that evaluates the 3D consistency when predicting depth jointly from sparse views.

5. Conclusion & Limitations

We present SIR-Diff, a simple and effective approach for multi-view image restoration, addressing blurring and low-resolution degradation across multiple images simultaneously. SIR-Diff outperforms leading single-view methods in sparse-view restoration and enables 3D-consistent image restoration for 3D computer vision applications. Future work will explore efficient cross-image attention mechanisms, such as selective patch attention along epipolar lines, and develop explicit guarantees for 3D view consistency.

Acknowledgment

This work partially used NCSA Delta GPU at the University of Illinois through allocation CIS230313 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

[1] Miika Aittala and Frédo Durand. Burst image deblurring using permutation invariant convolutional neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 731–747, 2018. 2

[2] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024. 2

[3] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, pages 53–72. Springer, 2025. 2

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 4

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[6] 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. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023. 2

[7] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 2, 6

[8] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE international conference on computer vision*, pages 576–584, 2015. 1, 2

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2

[10] Linwei Fan, Fan Zhang, Hui Fan, and Caiming Zhang. Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art*, 2(1):7, 2019. 2

[11] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2024. 7, 8

[12] Xiang Feng, Yongbo He, Yubo Wang, Yan Yang, Wen Li, Yifei Chen, Zhenzhong Kuang, Jianping Fan, Yu Jun, et al. Srgs: Super-resolution 3d gaussian splatting. *arXiv preprint arXiv:2404.10318*, 2024. 2

[13] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 2, 4, 5

[14] Shi Guo, Xi Yang, Jianqi Ma, Gaofeng Ren, and Lei Zhang. A differentiable two-stage alignment scheme for burst image reconstruction with large shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17472–17481, 2022. 2

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[17] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7920, 2023. 2

[18] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 8

[19] Michal Kawulok, Pawel Benecki, Szymon Piechaczek, Krzysztof Hrynczenko, Daniel Kostrzewa, and Jakub Nalepa. Deep learning for multiple-image super-resolution. *IEEE Geoscience and Remote Sensing Letters*, 17(6):1062–1066, 2019. 2

[20] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 5, 7, 8

[21] 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. 2, 7

[22] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023. 2

[23] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[24] Chenghao Li, Chaoning Zhang, Joseph Cho, Atish Waghwase, Lik-Hang Lee, Francois Rameau, Yang Yang, Sung-Ho Bae, and Choong Seon Hong. Generative ai meets 3d: A survey on text-to-3d in aigc era. *arXiv preprint arXiv:2305.06131*, 2023. 2

[25] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 2024. 2, 3, 5, 6

[26] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8576–8588, 2024. 2

[27] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 2

[28] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Taming diffusion models for image restoration: A review. *arXiv preprint arXiv:2409.10353*, 2024. 2

[29] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12861–12870, 2022. 1

[30] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 2

[31] Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 5, 6

[32] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 2, 5, 6, 7, 8

[33] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 2

[34] 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. 5

[35] 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. 2, 3, 4

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[37] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 5, 7

[38] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 134–144, 2019. 2, 5, 6, 8

[39] Yuan Shen, Duygu Ceylan, Paul Guerrero, Zexiang Xu, Niloy J Mitra, Shenlong Wang, and Anna Frühstück. Supergaussian: Repurposing video models for 3d super resolution. *arXiv preprint arXiv:2406.00609*, 2024. 2

[40] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 2

[41] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 4

[42] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 7

[43] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 7

[44] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *Advances in Neural Information Processing Systems*, 36: 12349–12362, 2023. 2

[45] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2025. 2

[46] Boyang Wang, Bowen Liu, Shiyu Liu, and Fengyu Yang. Vcisr: Blind single image super-resolution with video compression synthetic data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4302–4312, 2024. 2

[47] Boyang Wang, Fengyu Yang, Xihang Yu, Chao Zhang, and Hanbin Zhao. Apisr: anime production inspired real-world anime super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25574–25584, 2024. 2

[48] Chengcheng Wang, Zhiwei Hao, Yehui Tang, Jianyuan Guo, Yujie Yang, Kai Han, and Yunhe Wang. Sam-diffsr: Structure-modulated diffusion model for image super-resolution. *arXiv preprint arXiv:2402.17133*, 2024. 6

[49] Peng Wang, Lingzhe Zhao, Ruijie Ma, and Peidong Liu. Bad-nerf: Bundle adjusted deblur neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4170–4179, 2023. 7

[50] Peng Wang, Lingzhe Zhao, Ruijie Ma, and Peidong Liu. Bad-nerf: Bundle adjusted deblur neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4170–4179, 2023. 2

[51] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2

[52] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 5

[53] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 1, 2

[54] Yuxuan Wang, Xuanyu Yi, Zike Wu, Na Zhao, Long Chen, and Hanwang Zhang. View-consistent 3d editing with gaussian splatting. In *European Conference on Computer Vision*, pages 404–420. Springer, 2025. 2

[55] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[56] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2

[57] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024. 2, 3, 5, 6, 7, 8

[58] Hanyu Xiang, Qin Zou, Muhammad Ali Nawaz, Xianfeng Huang, Fan Zhang, and Hongkai Yu. Deep learning for image inpainting: A survey. *Pattern Recognition*, 134:109046, 2023. 2

[59] Cheng Yang, Lijing Liang, and Zhixun Su. Real-world denoising via diffusion model. *arXiv preprint arXiv:2305.04457*, 2023. 2

[60] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 8

[61] 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. 2, 5, 6, 7, 8

[62] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024. 2, 5

[63] Mulin Yu, Tao Lu, Linning Xu, Lihan Jiang, Yuanbo Xiangli, and Bo Dai. Gsdf: 3dgs meets sdf for improved rendering and reconstruction. *arXiv preprint arXiv:2403.16964*, 2024. 2

[64] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 5, 6, 7, 8

[65] 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. 2, 4

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5, 7

[67] Lingzhe Zhao, Peng Wang, and Peidong Liu. Bad-gaussians: Bundle adjusted deblur gaussian splatting. *arXiv preprint arXiv:2403.11831*, 2024. 2, 7, 8

[68] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2535–2545, 2024. 2, 3, 5, 6