IE-SRGS: An Internal-External Knowledge Fusion Framework for High-Fidelity 3D Gaussian Splatting Super-Resolution

Xiang Feng[1,2*], Tieshi Zhong[1*], Shuo Chang[1], Weiliu Wang[1], Chengkai Wang[3], Yifei Chen[3], Tongyu Hu[1], Yuhe Wang[1], Zhenzhong Kuang[1†], Xuefei Yin[4], Yanming Zhu[4]

[1]School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, 310018, China
[2]ShanghaiTech University, Shanghai, 201210, China
[3]Hangzhou Dianzi University, Hangzhou, 310018, China
[4]School of Information Communication and Technology, Griffith University, QLD, 4215, Australia

Abstract

Reconstructing high-resolution (HR) 3D Gaussian Splatting (3DGS) models from low-resolution (LR) inputs remains challenging due to the lack of fine-grained textures and geometry. Existing methods typically rely on pre-trained 2D super-resolution (2DSR) models to enhance textures, but suffer from 3D Gaussian ambiguity arising from cross-view inconsistencies and domain gaps inherent in 2DSR models. We propose IE-SRGS, a novel 3DGS SR paradigm that addresses this issue by jointly leveraging the complementary strengths of external 2DSR priors and internal 3DGS features. Specifically, we use 2DSR and depth estimation models to generate HR images and depth maps as external knowledge, and employ multi-scale 3DGS models to produce cross-view consistent, domain-adaptive counterparts as internal knowledge. A mask-guided fusion strategy is introduced to integrate these two sources and synergistically exploit their complementary strengths, effectively guiding the 3D Gaussian optimization toward high-fidelity reconstruction. Extensive experiments on both synthetic and real-world benchmarks show that IE-SRGS consistently outperforms state-of-the-art methods in both quantitative accuracy and visual fidelity.

Introduction

3D scene reconstruction plays a crucial role in various applications. A central task is novel view synthesis (NVS) to generate photorealistic images from unseen viewpoints. 3D Gaussian Splatting (3DGS) has recently emerged as a effective solution for real-time, high-fidelity rendering by explicitly representing scenes as a collection of 3D Gaussians (Kerbl et al. 2023). Despite its advantages in rendering quality and speed, 3DGS struggles to reconstruct accurate scenes from low-resolution (LR) inputs due to the lack of fine-grained textures and geometric details. Moreover, acquiring, storing, and transmitting high-resolution (HR) multi-view data is often costly or infeasible in practical scenarios, motivating the need for effective super-resolution (SR) methods that enable HR 3DGS reconstruction from LR observations.

Several recent methods (Yoon and Yoon 2023; Ko et al. 2025; Xie et al. 2024; Feng et al. 2024b; Shen et al. 2024) have explored 3DGS super-resolution, typically employing pre-trained 2D super-resolution (2DSR) models such as single-image super-resolution (SISR) or video super-resolution (VSR) to upsample LR views and generate pseudo-HR supervision for training 3DGS-based models. However, directly employing 2DSR models has two fundamental issues: (1) cross-view inconsistency, as 2D models process each view independently without enforcing multi-view consistency; and (2) domain gap, due to the discrepancy between the 2D training data and the target 3D novel scenes. Together, these issues lead to ambiguity during 3D Gaussian optimization. While prior works have attempted to address these issues through alignment, regularization, or post-processing techniques, substantial limitations remain.

In this work, we propose IE-SRGS, a novel 3DGS SR paradigm that mitigates ambiguity by jointly leveraging the complementary strength of external 2DSR priors and internal 3DGS features. Pre-trained 2DSR models offer strong HR detail priors but lack cross-view consistency and adaptability to novel 3D scenes. In contrast, multi-scale 3DGS models naturally enforce cross-view consistency and adapt to scene geometry but struggle to recover fine-grained textures from LR inputs (Yan et al. 2024; Yu et al. 2024b). IE-SRGS bridges this gap through three key components: (1) applying pre-trained 2DSR and depth estimation models to generate HR images and depth maps as external knowledge; (2) employing a multi-scale 3DGS model to produce cross-view consistent, domain-adaptative counterparts as internal knowledge; and (3) introducing a mask-guided fusion strategy to effectively integrate both sources and jointly guide the 3D Gaussain optimization. By systematically fusing external and internal knowledge, IE-SRGS effectively suppresses visual artifacts, restores fine details, and enables high-quality HR 3DGS from only LR inputs. Figure 1 illustrates our paradigm shift through internal-external knowledge fusion. Extensive experiments demonstrate that IE-SRGS consistently outperforms state-of-the-art (SOTA) methods.

Our contributions are summarized as follows:

- A novel 3DGS SR paradigm, IE-SRGS, that addresses the ambiguity issue by integrating complementary in-

*Equal contribution
†Corresponding author: Zhenzhong Kuang

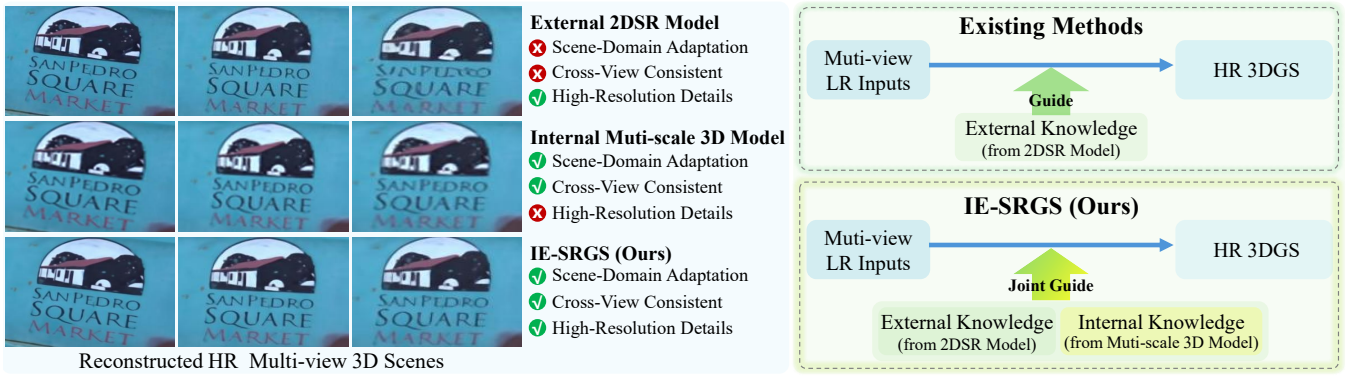

Figure 1: Comparison between existing methods and IE-SRGS. Left: External 2D Super-Resolution (2DSR) models provide detail but lack consistency; internal 3D models offer consistency but lack details. IE-SRGS achieves consistent and detailed outputs. Right: Existing methods rely solely on external priors. IE-SRGS jointly integrates external and internal knowledge, enabling high-quality high-resolution (HR) 3D Gaussian Splatting (3DGS) from low-resolution (LR) multi-view inputs.

ternal (multi-scale 3DGS-based) and external (2DSR-based) knowledge.

- Added geometric priors from the depth estimation model for optimization of Gaussians in 3DGS SR.
- A mask-guided strategy that effectively integrate internal and external guidance to supervise 3DGS optimization.
- Extensive experimental validation on both synthetic and real-world benchmarks, demonstrating that IE-SRGS consistently outperforms SOTA methods.

Related Work

Novel View Synthesis

NVS aims to generate photorealistic images from unseen viewpoints using multi-view inputs (Gortler et al. 2023). NeRF-based methods (Mildenhall et al. 2020; Xie et al. 2023; Reiser et al. 2021; Yariv et al. 2023; Huang et al. 2022; Chen et al. 2024) achieve high visual quality, but suffer from slow rendering due to costly ray marching. Grid-based methods (Müller et al. 2022; Sun, Sun, and Chen 2022; Shao et al. 2023; Chen et al. 2022; Barron et al. 2023; Liu et al. 2020) partially improve efficiency via structured discretization. Recent 3DGS methods (Kerbl et al. 2023; Zhang et al. 2024b; Yang et al. 2024b; Shi et al. 2025; Zhang et al. 2024a; Cheng et al. 2024; Qiao et al. 2025) have achieved real-time, high-fidelity NVS, but heavily rely on HR inputs to recover fine textures and geometry. However, acquiring HR multi-view images is costly or infeasible. To address this, methods like multi-scale reconstruction (Yan et al. 2024) and Mip-Splatting (Yu et al. 2024b) have been proposed to improve detail handling, but still struggle to recover high-frequency content without external priors.

Single-Image and Video Super-Resolution

SISR and VSR are widely used to generate external HR priors for 3DGS SR. SISR enhances spatial details from LR images using CNNs (Lim et al. 2017; Zhang et al. 2018b; Ding et al. 2025), Transformers (Liang et al. 2021), or GAN/diffusion models (Ledig et al. 2017; Wang et al. 2018; Saharia et al. 2022). VSR extends this by using temporal information across frames through motion-aware architectures (Chan et al. 2021, 2022; Liang et al. 2024a). While effective in 2D tasks, these models face challenges when applied to multi-view 3D reconstruction. They process each view independently, introducing cross-view inconsistencies, and often suffer from performance degradation due to domain gaps between training data and target 3D scenes. These limitations motivate our framework, which explicitly addresses cross-view consistency and domain adaptation for 3DGS SR.

3D Super-Resolution

High-quality 3D reconstruction from LR inputs remains challenging for both NeRF- and 3DGS-based methods. Earlier NeRF-based works (Wang et al. 2022; Feng et al. 2024a; Yoon and Yoon 2023; Lee, Li, and Lee 2024) leverage supersampling or 2DSR priors to improve texture quality, but often suffer from inefficiency or limited adaptability. Recent 3DGS-based methods (Shen et al. 2024; Ko et al. 2025; Xie et al. 2024; Feng et al. 2024b; Yu et al. 2024a) enable real-time rendering without extensive scene-specific fine-tuning, such as SRGS (Feng et al. 2024b) and GaussianSR (Yu et al. 2024a) methods. However, they rely mainly on external priors, largely ignoring internal 3D consistency and scene-specific adaptation. To address these limitations, we propose IE-SRGS, which integrates external high-frequency priors with internal 3D consistency from multi-scale 3DGS representations, and explicitly incorporates geometric depth priors to further enhance reconstruction quality.

Preliminaries

3D Gaussian Splatting. 3DGS (Kerbl et al. 2023) represents a 3D scene using a set of anisotropic 3D Gaussians $\mathcal{G} = \{\mathbf{g}_1, \mathbf{g}_2, ..., \mathbf{g}_N\}$, where each Gaussian is parameterized by a 3D position μ, covariance matrix Σ, color c, and opacity α. Each Gaussian defines a density distribution:

$$\mathbf{g}^{3D}(x) = \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right), \quad (1)$$

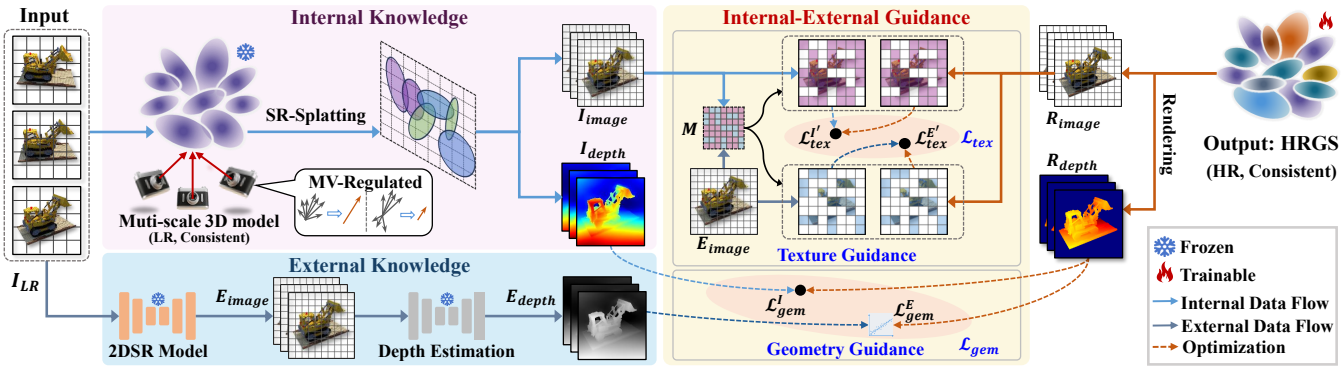

Figure 2: Overall framework of the proposed IE-SRGS. Given LR multi-view images, IE-SRGS extracts external guidance from pre-trained 2DSR models and internal guidance from a multi-scale 3D model. A mask-guided internal-external guidance integration provides structured, adaptive supervision for final HR 3DGS optimization.

where x denotes a 3D location. These Gaussians are projected into screen space and rasterized using splatted alpha blending to render novel views. In this work, we aim to enhance HR 3DGS reconstruction from LR inputs by improving the texture and geometric quality of Gaussian primitives via internal–external knowledge fusion.

Mip-Splatting. Mip-Splatting (Yu et al. 2024b) enhances 3DGS rendering stability by applying a 3D smoothing operation that suppresses aliasing and high-frequency noise under magnification. Specifically, each 3D Gaussian is convolved (i.e., $\otimes$) with a low-pass filter $\mathbf{g}_{\text{low}}$ before projection:

$$\mathbf{g}_{\text{reg}}^{\text{3D}}(x) = (\mathbf{g}^{\text{3D}} \otimes \mathbf{g}_{\text{low}})(x), \tag{2}$$

which yields a regularized Gaussian with enlarged covariance, effectively controlling the sampling rate and smoothing sharp artifacts. This operation preserves geometric consistency across views and improves robustness to domain variations. In our framework, we leverage this multi-scale rendering consistency to mitigate cross-view ambiguity.

Method

Overview. The overall framework of IE-SRGS is shown in Figure 2. Given a set of LR multi-view images, we first apply a pre-trained 2DSR model and a depth estimator to generate HR images and depth maps, forming the external guidance rich in texture details. Meanwhile, a multi-scale 3DGS model is constructed directly from LR inputs and optimized via multi-view regularization to ensure cross-view consistency and scene adaptation. We then apply a SR splatting to generate internal, upscaled images and depth maps, which serve as the internal guidance. Finally, we introduce a mask-guided fusion strategy to integrate internal and external knowledge, leveraging their complementary strengths in texture and geometry to guide HR 3DGS optimization.

External Knowledge for HR Detail Restoration

Reconstructing high-quality HR 3DGS from LR inputs is challenging due to lacking HR fine textures and accurate geometry. To address this, we propose to leverage external knowledge from pre-trained 2DSR and depth estimation models, which provide strong priors learned from large-scale datasets. Specifically, we use SwinIR (Liang et al. 2021) to generate super-resolved images with rich texture details, and Depth Anything V2 (Yang et al. 2024a) to estimate depth maps from these images, offering geometric cues. The resulting super-resolved image and depth map are denoted as E_{image} and E_{depth}, respectively, and serve as external guidance for optimizing the 3D Gaussians.

To inject external knowledge into the 3DGS model, we supervise the optimization of its Gaussian using both the texture and geometric information provided by E_{image} and E_{depth}. For texture guidance, we compare the rendered image R_{image}, obtained by projecting the current Gaussian onto the target view during optimization, with the external reference E_{image} using a weighted sum of L_1 loss and D-SSIM loss $\mathcal{L}_{\text{ds}}$ (Kerbl et al. 2023):

$$\mathcal{L}_{\text{tex}}^{E} = (1-\lambda)\mathcal{L}_1(E_{\text{image}}, R_{\text{image}}) + \lambda\mathcal{L}_{\text{ds}}(E_{\text{image}}, R_{\text{image}}), \tag{3}$$

where λ balances pixel-wise accuracy and structural similarity. For geometric supervision, we align the rendered depth R_{depth}, obtained via splatted rasterization during Gaussian optimization, with E_{depth} using a relaxed relative loss (Xiong et al. 2023) based on Pearson correlation:

$$\mathcal{L}_{\text{gem}}^{E} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\text{Cov}(R_{\text{depth}}^{i}, E_{\text{depth}}^{i})}{\sqrt{\text{Var}(R_{\text{depth}}^{i})\,\text{Var}(E_{\text{depth}}^{i})}}\right), \tag{4}$$

where N is the number of pixels, and $\text{Cov}(\cdot)$ and $\text{Var}(\cdot)$ denote covariance and variance, respectively. Together, these losses guide 3D Gaussians to capture fine textures and accurate geometry from external knowledge.

Internal Knowledge for Ambiguity Correction

While external knowledge offers rich texture and geometry priors for Gaussian optimization, it suffers from: (1) cross-view inconsistency, as 2D models process each view independently without enforcing 3D coherence; and (2) domain gaps between training data and target scenes. These limitations introduce ambiguity in Gaussian optimization. To address this, we introduce internal knowledge from a multi-scale 3DSR model that inherently enforces cross-view consistency and adaptability. We adopt Mip-Splatting (Yu et al.

2024b) as the internal backbone for its anti-aliasing capability and scale-consistent representation.

To extract internal guidance, we first train a multi-scale 3DGS model using LR multi-view inputs. To further enhance cross-view consistency, we adopt Multi-View Regulation (MV-Regulation) (Du, Wang, and Yu 2024), which supervises multiple views jointly during optimization. This strategy reduces overfitting to individual views and promotes geometric coherence across viewpoints. We then apply SR-Splatting to generate HR internal references. Specifically, 3D Gaussians are projected onto a 2D screen-space plane and upsampled using a predefined scaling factor. The upsampled splats are rasterized to produce internally scaled images and depth maps, denoted as I_{image} and I_{depth}, which serve as internal guidance for Gaussian optimization.

To correct the ambiguities introduced by external supervision, we supervise the rendered image R_{image} and depth map R_{depth} using the internal references I_{image} and I_{depth} through dedicated internal texture and geometry losses:

$$\mathcal{L}^I_{\text{tex}} = (1-\lambda)\mathcal{L}_1(I_{\text{image}}, R_{\text{image}}) + \lambda\mathcal{L}_{\text{ds}}(I_{\text{image}}, R_{\text{image}}), \quad (5)$$

$$\mathcal{L}^I_{\text{gem}} = \mathcal{L}_1(I_{\text{depth}}, R_{\text{depth}}). \quad (6)$$

These losses guide the 3D Gaussians toward consistent and scene-adaptive reconstructions, complementing the HR details provided by external knowledge.

Internal–External Fusion for HRGS Optimization

To fully exploit the complementary strengths of internal and external guidance, we propose integrating them through a unified supervision framework. This fusion enables accurate texture restoration and robust geometric reconstruction, and is structured into two components: Texture Guidance Integration and Geometric Guidance Integration.

Texture Guidance Integration. Considering that the inconsistencies and local artifacts introduced by 2DSR models are typically localized, and internal guidance ensures cross-view consistency but lacks HR details, we propose a mask-guided integration strategy for texture supervision. This can effectively address inconsistencies while preserving the high-quality texture details. Specifically, an uncertainty map $D(p)$ is computed at each pixel p to measure the discrepancy between I_{image} and E_{image}:

$$D(p) = \frac{|I_{\text{image}}(p) - E_{\text{image}}(p)|}{I_{\text{image}}(p) + \epsilon}, \quad (7)$$

where ϵ is a small constant (10^{-6}) for numerical stability. Then, a binary mask $M(p)$ is obtained with a threshold T:

$$M(p) = \begin{cases} 1, & D(p) \geq T \\ 0, & D(p) < T \end{cases}. \quad (8)$$

This mask guides the texture supervision process: regions with high discrepancy are trained using internal references, while the the remaining areas follow external guidance. The final texture loss is defined as:

$$\mathcal{L}_{\text{tex}} = \mathcal{L}^{I'}_{\text{tex}} + \mathcal{L}^{E'}_{\text{tex}}, \quad (9)$$

where $\mathcal{L}^{I'}_{\text{tex}} = \mathcal{L}^{I}_{\text{tex}} \odot M(p)$ and $\mathcal{L}^{E'}_{\text{tex}} = \mathcal{L}^{E}_{\text{tex}} \odot (1 - M(p))$ and $\odot$ denotes element-wise multiplication.

Geometric Guidance Integration. As geometric structures are relatively coarse and less sensitive to local variance, we directly combine internal and external geometric losses via a weighted sum by balancing factors λ_i and λ_e:

$$\mathcal{L}_{\text{gem}} = \lambda_i\mathcal{L}^I_{\text{gem}} + \lambda_e\mathcal{L}^E_{\text{gem}}, \quad (10)$$

The final loss for HR 3DGS optimization is defined as:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{tex}} + \mathcal{L}_{\text{gem}}. \quad (11)$$

Through joint internal-external guidance at both texture and geometry levels, IE-SRGS effectively resolves ambiguities, preserves HR details, and achieves high-fidelity 3D reconstruction from LR inputs.

Experiments

Datasets and Evaluation Metrics

We evaluate IE-SRGS on 21 scenes from four public datasets to demonstrate its robustness and generalization. Specifically, we use 9 real-world indoor and outdoor scenes from Mip-NeRF 360 (Barron et al. 2022), 2 from Deep Blending (Hedman et al. 2018), 2 from Tanks&Temples (Knapitsch et al. 2017), and 8 synthetic scenes from NeRF Synthetic (Mildenhall et al. 2020). To construct LR inputs, training views are downsampled using bicubic interpolation, by a factor of 8 for Mip-NeRF 360 (for $4\times$ 3D super-resolution) and by a factor of 4 for the other datasets, following SRGS (Feng et al. 2024b) for fair comparison. We assess reconstruction quality using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) (Wang, Simoncelli, and Bovik 2003), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018a).

Implementation Details

IE-SRGS is built upon the open-source Mip-Splatting codebase, with modification to the Gaussian rasterization module for depth map rendering. For internal model training, we use 3 randomly sampled views in MV-Regulation and optimize the multi-scale 3DGS model for 30,000 iterations using the same hyperparameters as Mip-Splatting (Yu et al. 2024b).

Methods	NeRF Synthetic		
	PSNR↑	SSIM↑	LPIPS↓
3DGS (Kerbl et al. 2023)	21.77	0.867	0.104
SwinIR-3DGS (Liang et al. 2021)	30.38	0.945	0.059
NeRF-SR (Wang et al. 2022)	28.90	0.927	0.099
DiSR-NeRF (Lee, Li, and Lee 2024)	26.00	0.890	0.123
FastSR-NeRF (Lin et al. 2024)	30.47	0.944	0.075
CROC (Yoon and Yoon 2023)	30.71	0.945	0.067
Mip-splatting (Yu et al. 2024b)	24.59	0.909	0.101
GaussianSR (Yu et al. 2024a)	28.37	0.924	0.087
SuperGaussian (Shen et al. 2024)	28.44	0.923	0.067
SRGS (Feng et al. 2024b)	30.83	0.948	0.056
Ours	**30.97**	**0.952**	**0.054**
Upper Bound	33.37	0.969	0.032

Note: Upper Bound refers to the results obtained by training HR 3DGS directly with HR inputs.

Table 1: Quantitative comparison of $4\times$ 3D super-resolution results on NeRF Synthetic dataset.

Methods	Mip-NeRF360			Deep Blending			Tanks&Temples		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS (Kerbl et al. 2023)	20.72	0.617	0.396	27.02	0.851	0.304	19.62	0.715	0.337
SwinIR-3DGS (Liang et al. 2021)	26.68	0.762	0.299	29.29	0.892	0.279	23.32	0.805	0.281
Mip-splatting (Yu et al. 2024b)	26.43	0.754	0.304	28.93	0.885	0.283	23.04	0.790	0.293
GaussianSR (Yu et al. 2024a)	25.60	0.663	0.368	28.28	0.873	0.307	–	–	–
SRGS (Feng et al. 2024b)	26.88	0.767	0.286	29.49	0.896	0.275	23.41	0.807	0.278
Squence Matters (Ko et al. 2025)	27.02	0.774	0.279	–	–	–	23.43	0.808	**0.274**
Ours	**27.15**	**0.779**	**0.278**	**29.63**	**0.899**	**0.271**	**23.52**	**0.810**	**0.274**
Upper Bound	27.23	0.797	0.254	29.73	0.905	0.243	23.51	0.828	0.242

Note: Upper Bound refers to the results obtained by training HR 3DGS directly with HR inputs.

Table 2: Quantitative comparison of 4× 3D super-resolution results on three real-world datasets Mip-NeRF360, Deep Blending, and Tanks&Temples evaluated by PSNR, SSIM, and LPIPS.

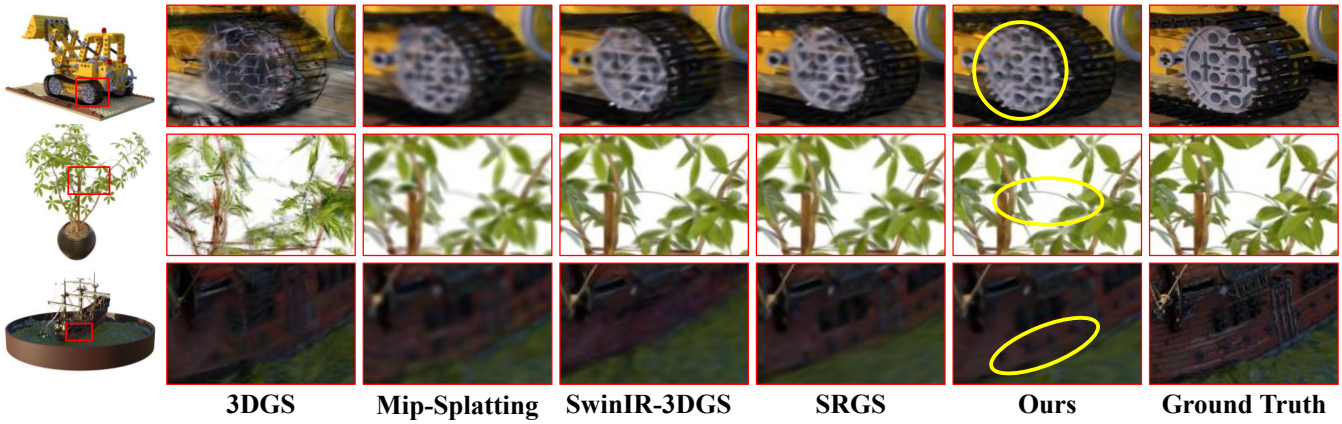

Figure 3: Qualitative comparisons of 4× 3D super-resolution on the NeRF Synthetic dataset. IE-SRGS achieves sharper textures and higher fidelity compared to SOTA methods 3DGS, Mip-Splatting, SwinIR-3DGS, and SRGS.

We set $\lambda_i = 0.001$ and $\lambda_e = 0.0001$ for internal and external supervision and use a discrepancy threshold T of 0.9 for real-world scenes and 0.6 for synthetic scenes. All experiments are run on a single NVIDIA RTX 4090 GPU.

Performance Comparison

We compare IE-SRGS with a range of SOTA 3D SR methods, including NeRF-SR (Wang et al. 2022), CROC (Yoon and Yoon 2023), DiSR-NeRF (Lee, Li, and Lee 2024), FastSR-NeRF (Lin et al. 2024), GaussianSR (Yu et al. 2024a), SuperGaussian (Shen et al. 2024), SRGS (Feng et al. 2024b), and Sequence Matters (Ko et al. 2025). SRGS is the most relevant baseline and Sequence Matters reports the best performance. For methods without public code (CROC, FastSR-NeRF, GaussianSR, SuperGaussian), we cite results from their papers. For real-world datasets lacking their results, we include comparisons with 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023), Mip-Splatting (Yu et al. 2024b), and a baseline variant SwinIR-3DGS.

Quantitative Comparison. Table 1 presents results on the NeRF Synthetic dataset for 4× 3D SR, a most common benchmark for SOTA methods. As shown, IE-SRGS outperforms all SOTA methods across all metrics, achieving the best performance. Notably, IE-SRGS improves upon its backbone Mip-Splatting by 25.9% in PSNR, 4.73% in SSIM, and 46.5% in LPIPS, clearly demonstrating the effectiveness of internal-external knowledge fusion. Moreover, its performance closely approaches the Upper Bound trained on HR inputs, highlighting its strong capability to recover accurate textures and geometries from LR inputs.

Table 2 reports results on three real-world datasets: Mip-NeRF360, Deep Blending, and Tanks&Temples. While many SOTA methods avoid evaluation on real-world scenes due to increased complexity and domain shifts, we evaluate across all datasets to thoroughly assess the robustness of our method. As shown, IE-SRGS again achieves the best performance across all metrics and datasets, closely approaching the Upper Bound. These results highlight its strong generalization under real-world LR constraints and validate our core claim that joint internal–external guidance enables consistent, high-quality 3D SR across diverse domains.

Qualitative Comparison. Figure 3 and Figure 4 show qualitative results for 4× 3D SR on synthetic and real-world datasets, respectively. As observed, standard 3DGS exhibits severe blurring due to the absence of HR details. Mip-Splatting improves consistency but still lacks fine textures, revealing the limitations of internal knowledge alone. Methods using external knowledge, such as SwinIR-3DGS and SRGS, recover partial details but introduce noticeable artifacts and distortions, especially in regions with complex structures. In contrast, IE-SRGS consistently generates sharper textures, more accurate geometry, and visually co-

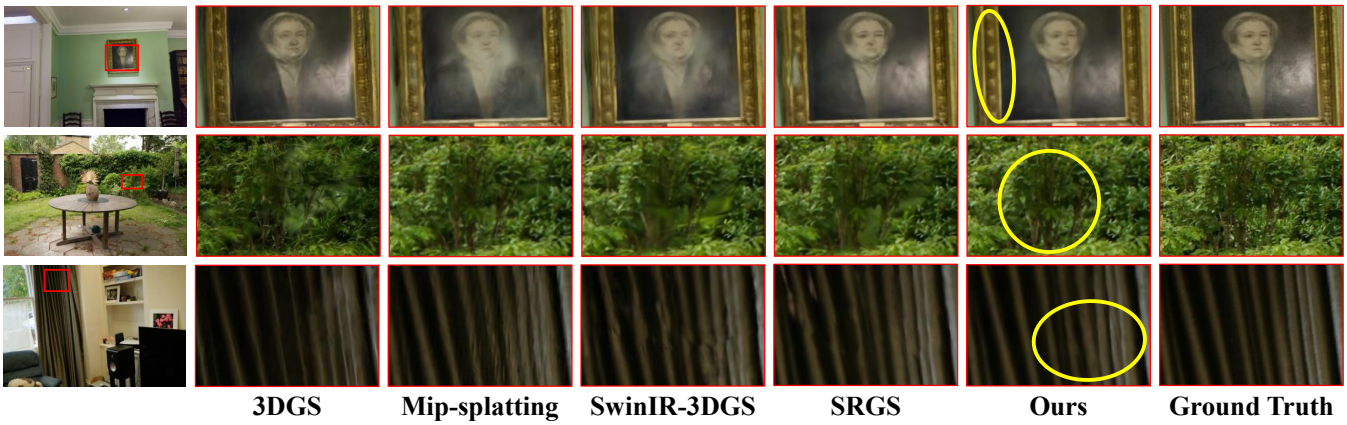

Figure 4: Qualitative comparisons of 4× 3D super-resolution on real-world datasets. IE-SRGS reconstructs finer textures and preserves more structural details compared to SOTA methods 3DGS, Mip-Splatting, SwinIR-3DGS, and SRGS.

Methods	MipNeRF-360 PSNR↑	SSIM↑	LPIPS↓
Mip-Splatting (Baseline)	26.43	0.754	0.304
+ MV-Regulation	26.68	0.757	0.297
Mip-Splatting (Baseline)	26.43	0.754	0.304
+ External Texture Guidance (E_{image})	26.69	0.762	0.300
+ External Geometric Guidance (E_{depth})	26.72	0.763	0.299
+ Internal Texture Guidance (I_{image})	27.00	0.775	0.283
+ Internal Geometric Guidance (I_{depth})	27.05	0.775	0.282
+ Mask-Guided Texture Integration	**27.15**	**0.779**	**0.278**

Table 3: Ablation study on component effectiveness of IE-SRGS. MV-Regulation is applied to Mip-Splatting to form a multi-view consistency baseline, while guidance components are progressively added without MV-Regulation.

	Backbones	MipNeRF-360 PSNR↑	SSIM↑	LPIPS↓
External	PSRT (Shi et al. 2022)	25.19	0.697	0.353
	Ours (PSRT)	**25.67**	**0.725**	**0.310**
	SwinIR (Liang et al. 2021)	25.23	0.699	0.332
	Ours (SwinIR)	**25.73**	**0.729**	**0.306**
Internal	A-Splatting (Liang et al. 2024b)	24.14	0.629	0.411
	Ours (A-Splatting)	**25.48**	**0.716**	**0.324**
	Mip-Splatting (Yu et al. 2024b)	25.04	0.687	0.349
	Ours (Mip-Splatting)	**25.73**	**0.729**	**0.306**

Table 4: Performance comparison across different backbones with and without our joint internal-external guidance.

herent results across all datasets, validating the effectiveness of internal–external fusion.

Ablation Study

Component Contribution Analysis. To evaluate the effectiveness of each component in IE-SRGS, we conduct ablation studies on the Mip-NeRF360 dataset. We first enhance Mip-Splatting with MV-Regulation to serve as a reference baseline for evaluating the impact of enforcing multi-view consistency. Then, we progressively introduce external texture guidance (E_{image}), external geometric guidance (E_{depth}), internal texture guidance (I_{image}), internal geometric guidance (I_{depth}), and finally, mask-guided texture integration.

Table 3 summarizes the quantitative results. Each component yields incremental improvements across all metrics, confirming the effectiveness of each design. Notably, the full internal-external joint guidance framework outperforms both the original and MV-Regulation-enhanced baselines, indicating that the joint guidance not only provides stronger multi-view consistency constraints but also enhances texture fidelity. Figure 5 presents qualitative comparisons. The Mip-Splatting baseline suffers from blurred textures and missing details. MV-Regulation improves consistency but still lacks high-frequency recovery. Introducing external guidance sharpens textures but introduces local artifacts (e.g., in the grass regions). Adding internal guidance reduces these artifacts but retains some blurring. With mask-guided fusion, internal and external cues are effectively combined, producing sharp, artifact-free textures and accurate geometry, as observed in both rendered images and depth maps.

Backbone Generalization Analysis. To assess the generalizability of our joint internal-external guidance, we apply it to alternative backbone models. Specifically, for external guidance, SwinIR is replaced with PSRT (Shi et al. 2022); for internal guidance, Mip-Splatting is replaced with Analytic-Splatting (Liang et al. 2024b). We conduct experiments on two representative scenes, bicycle and stump, from the MipNeRF-360 dataset. As shown in Table 4, our joint guidance consistently improves performance across different backbones on all metrics, confirming its strong robustness and backbone-agnostic generalization.

Hyperparameter Sensitivity. To assess the effect of the threshold T in mask-guided texture integration, we conduct a sensitivity analysis on two representative scenes from the Mip-NeRF360 dataset. The T balances internal and external guidance: $T = 0$ relies solely on internal supervision, while $T = 1$ heavily trusts external guidance in ambiguous regions. As shown in Figure 6, PSNR gradually improves as T increases from 0 to 0.9, indicating that moderate external

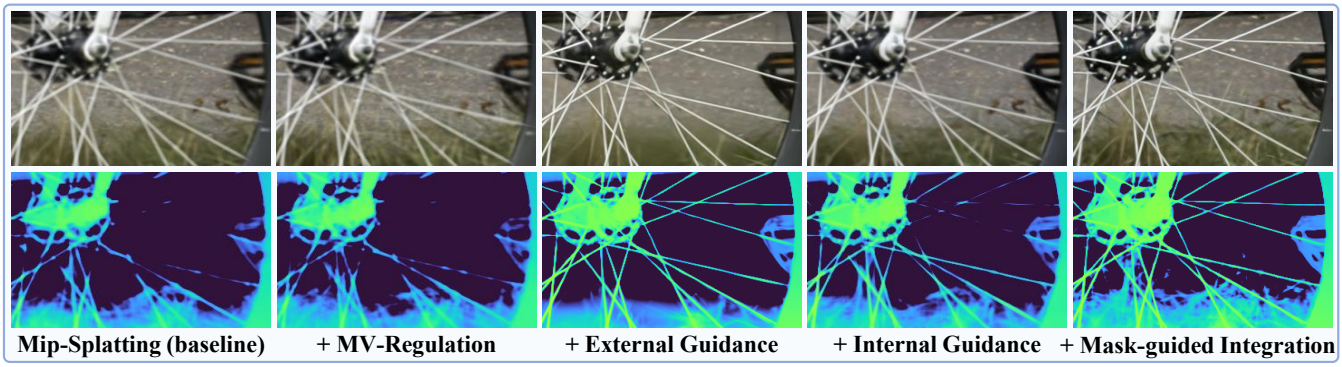

Figure 5: Qualitative results of the component effectiveness analysis. MV-Regulation is first added to Mip-Splatting to form a consistency baseline, while external, internal, and joint guidance components are progressively added without MV-Regulation.

Datasets	Methods	External Train	Internal Train	Depth Estimation	HRGS Train	Total Train Time	Inference (FPS)
NeRF Synthetic	SRGS	45.0s	-	-	13min11s	13mins 56s	191
	IE-SRGS	45.0s	6min40s	23.9s	11mins42s	19mins30s	**260**
MipNeRF-360	SRGS	2min39s	-	-	43min12s	45mins 51s	92
	IE-SRGS	2min39s	10min30s	48.3s	40mins20s	54mins17s	**119**

Table 5: Time breakdown and inference speed comparison of SRGS vs. IE-SRGS on NeRF Synthetic and MipNeRF-360.

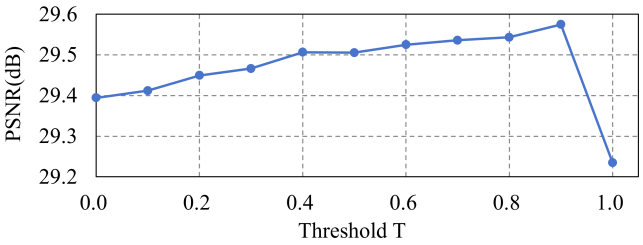

Figure 6: Threshold T sensitivity in mask-guided texture integration on two scenes from the Mip-NeRF360 dataset.

Methods	MipNeRF-360		
	PSNR↑	SSIM↑	LPIPS↓
Mip-Splatting (Yu et al. 2024b)	25.02	0.728	0.417
SRGS (Feng et al. 2024b)	25.27	0.741	0.405
Ours	**25.64**	**0.755**	**0.386**

Table 6: Quantitative results on 8× super-resolution on the bicycle and stump scenes from MipNeRF-360 dataset.

guidance enhances texture quality. The performance drop at $T = 1$ suggests that excessive reliance on external priors introduces artifacts. These results confirm that IE-SRGS is robust to threshold selection over a wide range and further validate the effectiveness of integrating internal knowledge.

Efficiency Analysis. To assess the efficiency of IE-SRGS, we analyze the runtime of each component and compare it with the SOTA method SRGS (Feng et al. 2024b). As shown in Table 5, although IE-SRGS includes additional modules such as internal training and depth estimation, the total training time increases only slightly, by 7-8 minutes, demonstrating the efficiency of our design. Importantly, IE-SRGS achieves significantly faster inference, with 260 FPS on NeRF Synthetic and 119 FPS on MipNeRF-360, compared to 191 FPS and 92 FPS for SRGS. This gain is due to faster convergence enabled by our joint internal-external guidance and mask-guided integration. These results show that IE-SRGS achieves high-quality reconstruction with minimal added cost and superior runtime efficiency.

Scaling Robustness Analysis. To assess the scalability of IE-SRGS, we perform an additional 8× SR experiment on the MipNeRF-360 dataset, focusing on two representative scenes: bicycle and stump. Existing SOTA methods rarely report results under such extreme scaling, our experiment offers a more rigorous evaluation of model robustness. Table 6 presents the quantitative results. IE-SRGS consistently outperforms both the baseline and SOTA method across all metrics, without requiring scene-specific fine-tuning. These results further highlight the strength of our internal–external guidance framework in preserving reconstruction quality even under large magnification factors.

Conclusion

We proposed IE-SRGS, a novel framework for 3D SR that integrates external and internal knowledge to optimize 3DGS. By combining HR detail priors with cross-view consistency and scene adaptation, IE-SRGS achieves high-fidelity 3D reconstruction from LR inputs. Extensive experiments demonstrate that IE-SRGS consistently outperforms SOTA methods and closely approaches the performance of HR upper bounds. This work lays the foundation for future research on unified frameworks for 3D low-level tasks and more effective internal-external knowledge integration.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62372147, U21B2040) and Zhejiang Provincial Natural Science Foundation of China (Grant No. LQK26F020005, LDT23F02025F02).

References

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2023. Zip-NeRF: Anti-aliased grid-based neural radiance fields. In *IEEE/CVF International Conference on Computer Vision*, 19697–19705.

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. Basicvsr: The search for essential components in video super-resolution and beyond. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4947–4956.

Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5972–5981.

Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; and Su, H. 2022. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, 333–350.

Chen, S.; Zhang, Y.; Xu, Y.; and Zou, B. 2024. Structure-aware neural radiance fields without posed camera. *Pattern Recognition*, 151: 110419.

Cheng, K.; Long, X.; Yang, K.; Yao, Y.; Yin, W.; Ma, Y.; Wang, W.; and Chen, X. 2024. Gaussianpro: 3D gaussian splatting with progressive propagation. In *International Conference on Machine Learning*.

Ding, J.; Zhao, Y.; Pei, L.; Shan, Y.; Du, Y.; and Li, W. 2025. Modal-invariant progressive representation for multimodal image registration. *Information Fusion*, 117: 102903.

Du, X.; Wang, Y.; and Yu, X. 2024. MVGS: Multi-view-regulated Gaussian Splatting for Novel View Synthesis. arXiv:2410.02103.

Feng, X.; He, Y.; Wang, Y.; Wang, C.; Kuang, Z.; Ding, J.; Qin, F.; Yu, J.; and Fan, J. 2024a. ZS-SRT: An efficient zero-shot super-resolution training method for Neural Radiance Fields. *Neurocomputing*, 590: 127714.

Feng, X.; He, Y.; Wang, Y.; Yang, Y.; Li, W.; Chen, Y.; Kuang, Z.; Fan, J.; Jun, Y.; et al. 2024b. Srgs: Super-resolution 3D gaussian splatting. *arXiv preprint arXiv:2404.10318*.

Gortler, S. J.; Grzeszczuk, R.; Szeliski, R.; and Cohen, M. F. 2023. The lumigraph. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 453–464.

Hedman, P.; Philip, J.; Price, T.; Frahm, J.-M.; Drettakis, G.; and Brostow, G. 2018. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics*, 37(6): 1–15.

Huang, Y.-H.; He, Y.; Yuan, Y.-J.; Lai, Y.-K.; and Gao, L. 2022. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18342–18352.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4): 1–13.

Ko, H.-k.; Park, D.; Park, Y.; Lee, B.; Han, J.; and Park, E. 2025. Sequence Matters: Harnessing Video Models in 3D Super-Resolution. In *AAAI Conference on Artificial Intelligence*, 4356–4364.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4681–4690.

Lee, J. L.; Li, C.; and Lee, G. H. 2024. DiSR-NeRF: Diffusion-Guided View-Consistent Super-Resolution NeRF. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20561–20570.

Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; and Van Gool, L. 2024a. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 33: 2171–2182.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *IEEE/CVF International Conference on Computer Vision Workshops*, 1833–1844.

Liang, Z.; Zhang, Q.; Hu, W.; Zhu, L.; Feng, Y.; and Jia, K. 2024b. Analytic-splatting: Anti-aliased 3d gaussian splatting via analytic integration. In *European Conference on Computer Vision*, 281–297.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Lee, K. M. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 136–144.

Lin, C.-Y.; Fu, Q.; Merth, T.; Yang, K.; and Ranjan, A. 2024. Fastsr-nerf: Improving nerf efficiency on consumer devices with a simple super-resolution pipeline. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 6036–6045.

Liu, L.; Gu, J.; Zaw Lin, K.; Chua, T.-S.; and Theobalt, C. 2020. Neural sparse voxel fields. In *Advances in Neural Information Processing Systems*, volume 33, 15651–15663.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 405–421.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4): 1–15.

Qiao, Y.; Shao, M.; Meng, L.; and Xu, K. 2025. RestorGS: Depth-aware Gaussian Splatting for Efficient 3D Scene Restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11177–11186.

Reiser, C.; Peng, S.; Liao, Y.; and Geiger, A. 2021. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *IEEE/CVF International Conference on Computer Vision*, 14335–14345.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4713–4726.

Shao, R.; Zheng, Z.; Tu, H.; Liu, B.; Zhang, H.; and Liu, Y. 2023. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16632–16642.

Shen, Y.; Ceylan, D.; Guerrero, P.; Xu, Z.; Mitra, N.; Wang, S.; and Fr"uhst"uck, A. 2024. SuperGaussian: Repurposing Video Models for 3D Super Resolution. In *European Conference on Computer Vision*, 215–233.

Shi, C.; Yang, C.; Hu, X.; Yang, Y.; Ding, J.; and Tan, M. 2025. MMGS: Multi-model synergistic Gaussian splatting for sparse view synthesis. *Image and Vision Computing*, 105512.

Shi, S.; Gu, J.; Xie, L.; Wang, X.; Yang, Y.; and Dong, C. 2022. Rethinking alignment in video super-resolution transformers. In *Advances in Neural Information Processing Systems*, volume 35, 36081–36093.

Sun, C.; Sun, M.; and Chen, H.-T. 2022. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5459–5469.

Wang, C.; Wu, X.; Guo, Y.-C.; Zhang, S.-H.; Tai, Y.-W.; and Hu, S.-M. 2022. Nerf-sr: High quality neural radiance fields using supersampling. In *ACM International Conference on Multimedia*, 6445–6454.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Loy, C. C. 2018. ESRGAN: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers*, volume 2, 1398–1402.

Xie, S.; Wang, Z.; Zhu, Y.; and Pan, C. 2024. SuperGS: Super-Resolution 3D Gaussian Splatting via Latent Feature Field and Gradient-guided Splitting. *arXiv preprint arXiv:2410.02571*.

Xie, X.; Gherardi, R.; Pan, Z.; and Huang, S. 2023. HollowNeRF: Pruning hashgrid-based NeRFs with trainable collision mitigation. In *IEEE/CVF International Conference on Computer Vision*, 3480–3490.

Xiong, H.; Muttukuru, S.; Upadhyay, R.; Chari, P.; and Kadambi, A. 2023. SparseGS: Real-Time 360° Sparse View Synthesis using Gaussian Splatting. *arXiv:2312.00206*.

Yan, Z.; Low, W. F.; Chen, Y.; and Lee, G. H. 2024. Multi-scale 3D gaussian splatting for anti-aliased rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20923–20931.

Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth anything v2. In *Advances in Neural Information Processing Systems*, 21875–21911.

Yang, Z.; Gao, X.; Zhou, W.; Jiao, S.; Zhang, Y.; and Jin, X. 2024b. Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20331–20341.

Yariv, L.; Hedman, P.; Reiser, C.; Verbin, D.; Srinivasan, P. P.; Szeliski, R.; Barron, J. T.; and Mildenhall, B. 2023. Bakedsdf: Meshing neural sdfs for real-time view synthesis. In *ACM SIGGRAPH*, 1–9.

Yoon, Y.; and Yoon, K.-J. 2023. Cross-guided optimization of radiance fields with multi-view image super-resolution for high-resolution novel view synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12428–12438.

Yu, X.; Zhu, H.; He, T.; and Chen, Z. 2024a. GaussianSR: 3D Gaussian Super-Resolution with 2D Diffusion Priors. *arXiv preprint arXiv:2406.10111*.

Yu, Z.; Chen, A.; Huang, B.; Sattler, T.; and Geiger, A. 2024b. Mip-splatting: Alias-free 3d gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19447–19456.

Zhang, J.; Li, J.; Yu, X.; Huang, L.; Gu, L.; Zheng, J.; and Bai, X. 2024a. CoR-GS: sparse-view 3D Gaussian splatting via co-regularization. In *European Conference on Computer Vision*, 335–352.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *European Conference on Computer Vision*, 286–301.

Zhang, Z.; Hu, W.; Lao, Y.; He, T.; and Zhao, H. 2024b. Pixel-GS: Density Control with Pixel-aware Gradient for 3D Gaussian Splatting. In *European Conference on Computer Vision*, 326–342.