3D-GSR: 3D Super-Resolution with Multi-View Adapter and Gaussian Splatting

Anna-Alina Bondarets, Taras Rumezhak, Hlib Plakhtii, Alina Muliak, and Volodymyr Karpiv

SoftServe
{anbondaret, trume, hplak, almuli, vkarpi}@softserveinc.com

Abstract. The rapid advancements in 2D generative models and efficient 3D reconstruction techniques are fueling the growth of the 3D visual generative AI field. In this work, we address the challenge of 3D super-resolution by developing a consistent multi-view super-resolution model and leveraging 3D reconstruction techniques to obtain high-quality 3D models from the ones of lower quality. Our approach is versatile, generalizing to various inputs, including renderings of digital 3D assets and turn-around videos of real-world objects. Its adaptability also allows it to serve as a post-processing step for any existing 3D generative method, regardless of the underlying geometry representation.
We introduce 3D-GSR, a model that integrates 2D super-resolution with 3D reconstruction to enhance the 3D generation process. Central to our approach, we propose the Multi-View Super-Resolution (MV-SR) adapter, designed to leverage arbitrary single-image SR model to generate consistent multi-view frames of the same object. We tested our MV-SR technique with four common SR models: R-ESRGAN [32], SwinIR [15], HAT-L [6], and SD-XL based [21], showing how our adapter increases each model's accuracy in the multi-view scenario. The geometry of the asset is reconstructed using a novel 3D Gaussian Splatting [12] model. Our 3D-GSR model is adept at both geometric transformations and appearance enhancements, producing assets with intricate geometry and exceptional fidelity, excelling in capturing high-fidelity details of the high-frequency components in particular.

Keywords: Super-Resolution · 3D Gaussian Splatting · Multi-view · Novel View Synthesis · High-Resolution Novel View Synthesis

1 Introduction

The creation of 3D models through generative techniques has advanced significantly, with various approaches [22, 24, 27, 30] now enabling the generation of 3D assets, sometimes guided by text prompts [4, 7, 34, 36, 40] or reference images [16, 18, 37, 43]. These techniques leverage both image-based and 3D data for training, allowing for rapid production of diverse outputs. However, despite

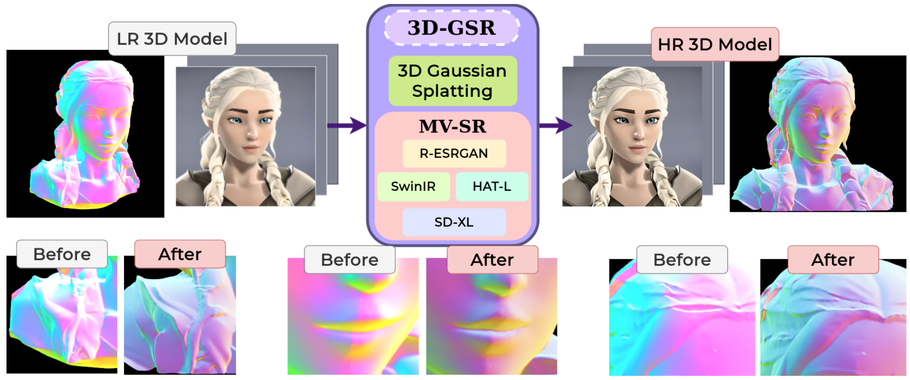

Fig. 1: Our 3D-GSR model upscales the low-resolution (LR) 3D model to a high-resolution (HR) using our multi-view adapter MV-SR for arbitrary single-image SR. Our model improves both large-scale and high-frequency details of the final 3D model.

these advancements, the 3D models generated by these methods still fall short in terms of detail and precision when compared to the latest generative models used for producing images [19, 21, 23] or videos [2, 17].

Novel View Synthesis (NVS) has long been a challenging problem in computer vision and graphics. The introduction of Neural Radiance Fields (NeRF) [20] marked a significant breakthrough, showcasing the power of implicit neural representations for generating novel views. More recently, 3D Gaussian Splatting (3D GS) [12] has emerged as a popular alternative, offering a primitive-based representation that surpasses NeRF in both rendering speed and reconstruction quality. Despite these advantages, 3DGS faces a significant limitation when performing high-resolution novel view synthesis (HRNVS) [39] from low-resolution inputs, often resulting in noticeably degraded renderings.

In this work, we tackle the problem of 3D super-resolution and HRNVS by leveraging the 3D Gaussian Splatting [12] geometry optimization model with our multi-view modification of the single image SR model.

In summary, our contributions are:

- We propose a novel MV-SR adapter, a multi-view modification of the original 2D super-resolution model to consistently upscale and refine the low-quality multi-view frames of the same object. It can be used with any single-image upscaler, and we train four common SR models in our multi-view scenario: R-ESRGAN [32], SwinIR [15], HAT-L [6], and SD-XL based [21].
- We propose our 3D super-resolution model 3D-GSR by integrating the pre-trained Multi-View SD-XL refiner together with the 3D Gaussian Splatting [12] for geometry reconstruction.
- We demonstrate the generalization of 3D-GSR to various inputs of different natures, such as real-life and digital ones. Moreover, 3D-GSR can be applied as a post-processing refinement stage for any 3D generative model using arbitrary representations of the geometry.

2 Related Works

2.1 Image Super-Resolution

The image super-resolution task aims to enhance the resolution of low-quality images, producing high-resolution counterparts while preserving details and reducing artifacts. One of the pioneering models in this domain is ESRGAN [33], which leveraged a GAN-based architecture with introduced perceptual loss to improve image quality and realistic textures. Building on ESRGAN, the improved R-ESRGAN [32] incorporated residual connections to further stabilize training and enhance performance, particularly in challenging cases like real-world images.

With the recent rapid advancements of the denoising diffusion models for image generation are also intensively explored in the super-resolution task [11, 13, 14, 21, 25, 42].

There are also a lot of transformers-based SR models [5, 6, 8, 15]. SwinIR (Swin Transformer for Image Restoration) [15] in particular utilizes a hierarchical transformer-based architecture to capture both local and global features, achieving state-of-the-art performance in image super-resolution tasks. HAT-L [6] is the most recent SOTA transformers-based upscaler incorporating a hybrid attention technique.

2.2 3D Super-Resolution

There are several important works tackling the 3D super-resolution problem. SuperGaussian [28] is a recent approach that leverages the GAN-based video super-resolution model VideoGigaGAN [38] to enhance and upscale the frame-wise images in consistent manner as a video.

There are also methods that build-up upon Neural Radiance Fields for enhancing the visual fidelity of the reconstruction [31]. The SOTA 3D upscaling model CROC [39] uses a multi-view image SR model to upscale the quality of NeRF reconstruction. Our approach is highly inspired by the image-to-3D generative Unique3D [35] model that leverages multi-view diffusion.

After the recent development of the groundbreaking 3D Gaussian Splatting (3D GS) model [12], there are several crucial studies on 3D super-resolution using the 3D GS. GaussianSR [41] leverages the SDS loss [22], initially proposed for the text-to-3D generation by DreamFusion [22]. SRGS [10] introduces a densification and texture learning to enhance the representation ability of primitives of 3D GS by leveraging the image SR.

3 Approach

Achieving generation consistency across multiple images is known to be a challenging task ([29]). The main idea of our approach is to processs the super-resolution upscaling with a single inference of the generative model to achieve consistency. For that, we process the renderings not in the frame-wise manner, but rather by composing a model sheet of them and leveraging the SR model to multi-view data, inspired by the recent multi-view SR model for NeRF in CROC [39].

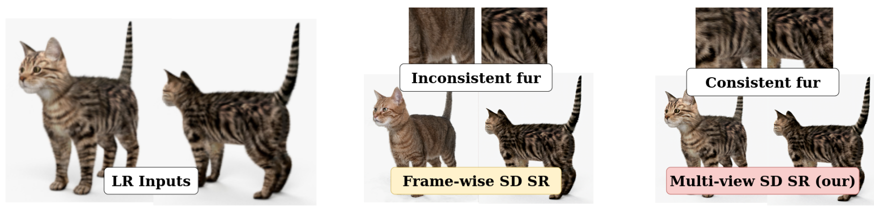

Fig. 2: Our MV-SR adapter makes the SD-based upscaling model provide consistent refinement across multiple views, while the original model creates inconsistent details.

3.1 Model sheet

The concept was first introduced in the MVDream [29] study on 3D generation using diffusion models. The authors trained a multi-view diffusion model to generate four orthogonal views arranged in a grid. Following a similar approach, we also organize orthogonal frames into a grid, which we refer to as a model sheet, as this method ensures consistent generations and upscalings.

3.2 Multi-View Super-Resolution

To leverage the capabilities of any SR upscaling model for 3D super-resolution, we adapted it to a multi-view setup, training the model on multi-view data to enhance the resolution of multiple perspectives of 3D objects simultaneously. This approach, termed MV-SR, focuses on generating consistent high-resolution outputs across various angles of the same object. Moreover, its generalization makes it possible to use any image SR model to extend it to multi-view data. More specifically, we tested four main image upscaling models in out multi-view scenario: R-ESRGAN [32], SwinIR [15], HAT-L [6], and SD-XL based [21].

We utilized the Objaverse 1.0 dataset [9], which contains over 800K 3D models. For each 3D asset, we generated a set of random renderings from 4 orthogonal angles, composing them into a single model sheet. Each model sheet, consisting of 4 frames, is being classically downscaled by x4 to create the LR and HR pairs for our MV-SR training. We trained the MV-SR as a x4 upscaler and refiner. In this setup, the image super-resolution model was trained to enhance the resolution of the model sheet, aiming to produce high-fidelity outputs that are consistent across all views.

The training process employed the MV-SR model to learn a mapping from low-resolution input model sheets to high-resolution original renderings. For a given high-resolution model sheet m_{HR} composed of 4 key frames, it is downscaled to the low-resolution image m_{LR} for the model to predict the high-resolution output from it, minimizing the reconstruction loss between the predicted and target high-resolution images. For the loss function, we use the Charbonnier loss [3]:

$$\mathcal{L} = \sqrt{\|m_{HR} - \phi(m_{LR})\|^2 + \epsilon^2}, \tag{1}$$

where ϕ is our MV-SR model, and ϵ is an empirically set constant to 10^{-3}.

To ensure consistency across multiple views, the model was trained to understand the spatial relationships between the different angles, thereby producing outputs that are coherent when viewed from any perspective. This multi-view consistency is crucial for maintaining the structural integrity of the 3D object in high resolution, avoiding artifacts that could arise from processing each view independently.

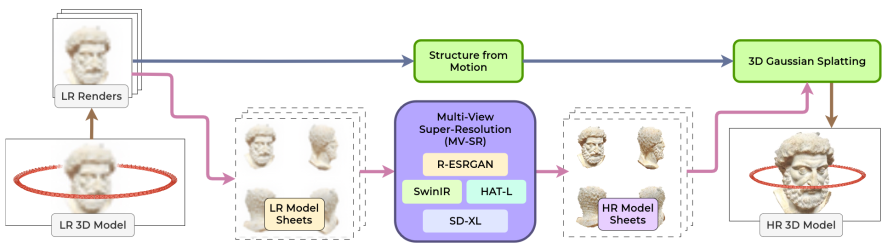

Fig. 3: Our 3D-GSR architecture. The core component is our MV-SR adapter, which leverages an arbitrary single-image SR model to multi-view data in model sheets. The SfM component creates the base point cloud from the original frames only for the best consistency, and the geometry is then refined with 3D GS on the upscaled frames.

3.3 3D-GSR

We can easily extend our Multi-View SR modification to the 3D task in our 3D-GSR pipeline by composing the model sheets from the low-resolution turnaround images of the object.

For the final 3D reconstruction, we incorporate the original implementation of 3D Gaussian Splatting [12] model. The model optimizes the initial Gaussians extracted from the original point cloud to match the updated appearance. The initially reconstructed splats are rendered into separate frames and composed into a model sheet, which is then passed to the MV-SR upscaler to produce consistent high-resolution frames with an enhanced appearance. The last iterations of the optimization process guide the 3D reconstruction using upscaled images, producing 3D assets of higher quality.

For the initialization of 3D Gaussians, an initial point cloud is used. For that, we run a Structure from Motion (SfM) algorithm from COLMAP [26] library.

Our key observation from the experiments is that it should be initialized from low-resolution originals only to create a consistent carcass for the geometry.

To train the model, we adopted the train-test split approach from Mip-NeRF 360 [1], allocating 7/8 of the input images for training and reserving the remaining 1/8 for testing. We then evaluated the rendering quality on the test set using SSIM and PSNR metrics, comparing the performance of two models—one trained on original images and the other on the upscaled versions. Additionally, we disabled the 3D Gaussian Splatting (GS) feature for optimizing view-dependent colors with spherical harmonics (SH) coefficients.

4 Experiments

4.1 3D-SR

We trained four main image upscaling models in our multi-view scenario with our MV-SR adapter: R-ESRGAN [32], SwinIR [15], HAT-L [6], and SD-XL based [21], shifting each model from the single-view scenario to the multi-view. We conducted an experiment by comparing the super-resolution upscaling and refining process of our MV models with their original versions on the model sheets. We compare them in Fig. 4, which visualizes how our proposed modification achieves consistency on different frames.

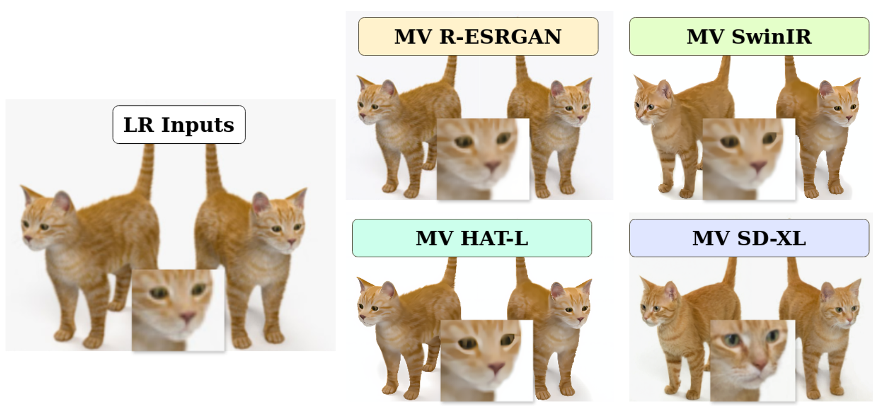

Fig. 4: We show the consistent multi-view refinements with our MV-SR adapter applied and trained for four single-image SR models: R-ESRGAN [32], SwinIR [15], HAT-L [6], and SD-XL based [21].

To evaluate the performance of our MV super-resolution models on the multi-view super-resolution task, we compared the average PSNR and SSIM scores on our dataset, as shown in Tab. 1. We assessed the similarity to the input image (ensuring the output maintains consistency with the input) and the overall quality of the super-resolved image. For comparison, we tested each of the original

Table 1: Comparison of PSNR and SSIM metrics on model sheets with the original single-image SR models R-ESRGAN [32], SwinIR [15], HAT-L [6], and SD-XL based [21], and our multi-view modifications with MV-SR.

Model	PSNR ↑	SSIM ↑
MV R-ESRGAN (our)	**32.29**	**0.922**
R-ESRGAN (model sheet)	31.87	0.917
R-ESRGAN (frame-wise)	31.89	0.917
MV SwinIR (our)	**32.33**	**0.932**
SwinIR (model sheet)	32.09	0.928
SwinIR (frame-wise)	32.21	0.926

Model	PSNR ↑	SSIM ↑
MV HAT-L (our)	**33.61**	**0.933**
HAT-L (model sheet)	33.48	0.932
HAT-L (frame-wise)	33.5	0.932
MV SD-XL (our)	**33.59**	**0.931**
SD-XL (model sheet)	33.31	0.926
SD-XL (frame-wise)	38.46	0.931

models using two different approaches: processing the whole model sheet (similar to our MV models) and processing individual frames sequentially, which is closer to the model's original training distribution. All of our multi-view models outperformed the originals in both scenarios, indicating that they provide more accurate and realistic super-resolution results, more consistent in multi-view cases than the default single-image models.

4.2 Results

We show the 3D upscaling results produced with 3D-GSR in Figs. 1, 5 and 6. Our model excels at generating high-quality 3D assets that accurately reflect the upscaling process done upon low-quality inputs. 3D-GSR is adept at both geometric transformations and appearance enhancements, producing assets with intricate geometry and exceptional fidelity.

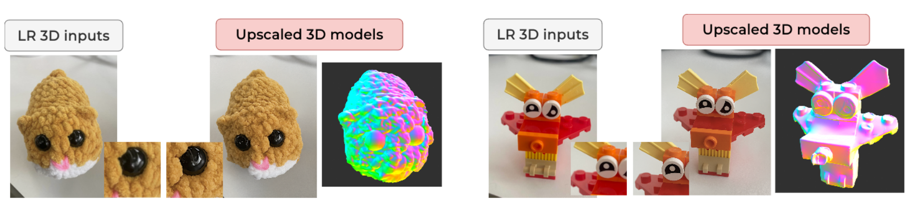

Fig. 5: Upscaling results of 3D-GSR on real-life objects.

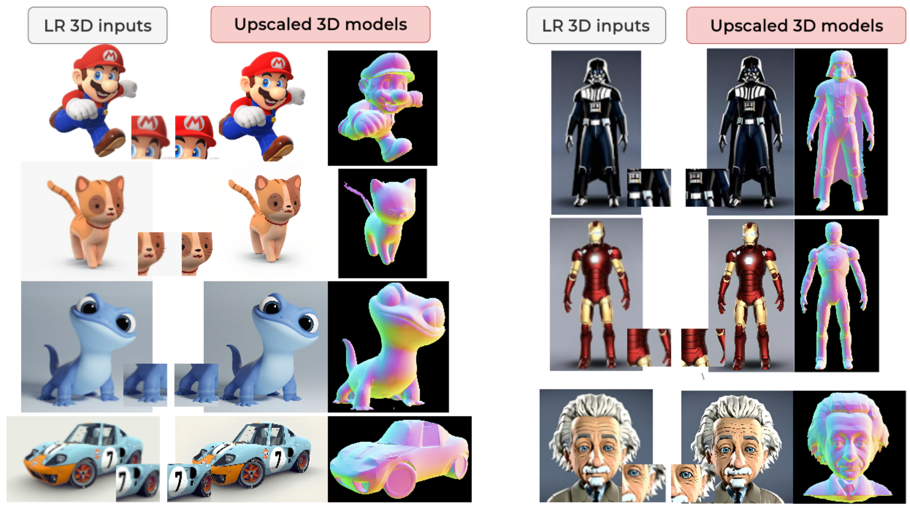

Fig. 6: Upscaling results of 3D-GSR on digital objects.

We also present examples of 3D-GSR applied to turn-around images of real-world objects in Fig. 5. We show how we can enhance the visual quality of the initially blurry object with our upscaling process.

5 Conclusions

In this work, we propose MV-SR adapter, a technique to modify arbitrary single-image 2D super-resolution model to work with multi-view data, extending its application to 3D super-resolution tasks with our 3D-GSR pipeline. We trained four common SR models in our multi-view scenario: R-ESRGAN [32], SwinIR [15], HAT-L [6], and SD-XL based [21], showing how our adapter increases each model's accuracy in the multi-view scenario.

We propose 3D-GSR, a novel approach to tackle the 3D super-resolution task by leveraging our Multi-View SR model, which allows for high-quality super-resolution of 3D assets to ensure consistent enhancement across multiple views. For geometry estimation, we utilize a pipeline of the 3D Gaussian Splatting model [12] for accurate reconstruction of 3D assets.

Our method is generalizable to both digital and real-life inputs and can be applied as a post-hoc enhancement stage for any existing 3D generative model with arbitrary geometry representation. Our model is capable of object-level refinement, while scene-level would require additional training.

Our work opens new possibilities for leveraging any 2D super-resolution models in multi-view contexts, extending beyond the single image-specific task. We propose a novel technique for 3D super-resolution, achieving high-quality upscaling and refinement of the low-quality 3D models while preserving object consistency.

References

1. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields (2022) 6
2. Chai, W., Guo, X., Wang, G., Lu, Y.: Stablevideo: Text-driven consistency-aware diffusion video editing (2023), https://arxiv.org/abs/2308.09592 2
3. Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of 1st International Conference on Image Processing. vol. 2, pp. 168–172 vol.2 (1994). https://doi.org/10.1109/ICIP.1994.413553 5
4. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation (2023) 1
5. Chen, X., Wang, X., Zhang, W., Kong, X., Qiao, Y., Zhou, J., Dong, C.: Hat: Hybrid attention transformer for image restoration (2023), https://arxiv.org/abs/2309.05239 3
6. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer (2023), https://arxiv.org/abs/2205.04437 1, 2, 3, 4, 6, 7, 8
7. Chen, Z., Wang, F., Wang, Y., Liu, H.: Text-to-3d using gaussian splatting (2024) 1
8. Conde, M.V., Choi, U.J., Burchi, M., Timofte, R.: Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration (2022) 3
9. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects (2022), https://arxiv.org/abs/2212.08051 4
10. Feng, X., He, Y., Wang, Y., Yang, Y., Li, W., Chen, Y., Kuang, Z., ding, J., Fan, J., Jun, Y.: Srgs: Super-resolution 3d gaussian splatting (2024), https://arxiv.org/abs/2404.10318 3
11. Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., Zhang, B.: Implicit diffusion models for continuous super-resolution (2023), https://arxiv.org/abs/2303.16491 3
12. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023), https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/ 1, 2, 3, 5, 8
13. Li, G., Xing, W., Zhao, L., Lan, Z., Sun, J., Zhang, Z., Zhang, Q., Lin, H., Lin, Z.: Self-reference image super-resolution via pre-trained diffusion large model and window adjustable transformer. In: Proceedings of the 31st ACM International Conference on Multimedia. p. 7981–7992. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3581783.3611866, https://doi.org/10.1145/3581783.3611866 3
14. Li, H., Yang, Y., Chang, M., Chen, S., Feng, H., Xu, Z., Li, Q., Chen, Y.: Srdiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing **479**, 47–59 (2022). https://doi.org/https://doi.org/10.1016/j.neucom.2022.01.029, https://www.sciencedirect.com/science/article/pii/S0925231222000522 3
15. Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L.V., Timofte, R.: Swinir: Image restoration using swin transformer (2021) 1, 2, 3, 4, 6, 7, 8
16. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object (2023) 1

17. Melnik, A., Ljubljanac, M., Lu, C., Yan, Q., Ren, W., Ritter, H.: Video diffusion models: A survey (2024), https://arxiv.org/abs/2405.03150 2
18. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures (2022) 1
19. Midjourney, I.: Midjourney. https://www.midjourney.com/ (2022) 2
20. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis (2020) 2
21. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023) 1, 2, 3, 4, 6, 7, 8
22. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion (2022) 1, 3
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021) 2
24. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022) 1
25. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement (2021), https://arxiv.org/abs/2104.07636 3
26. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 5
27. Shen, T., Gao, J., Yin, K., Liu, M.Y., Fidler, S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis (2021) 1
28. Shen, Y., Ceylan, D., Guerrero, P., Xu, Z., Mitra, N.J., Wang, S., Frühstück, A.: Supergaussian: Repurposing video models for 3d super resolution (2024), https://arxiv.org/abs/2406.00609 3
29. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation (2024) 3, 4
30. Sun, J., Zhang, B., Shao, R., Wang, L., Liu, W., Xie, Z., Liu, Y.: Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior (2023), https://arxiv.org/abs/2310.16818 1
31. Wang, C., Wu, X., Guo, Y.C., Zhang, S.H., Tai, Y.W., Hu, S.M.: Nerf-sr: High quality neural radiance fields using supersampling. In: Proceedings of the 30th ACM International Conference on Multimedia. ACM (Oct 2022). https://doi.org/10.1145/3503161.3547808, http://dx.doi.org/10.1145/3503161.3547808 3
32. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data (2021) 1, 2, 3, 4, 6, 7, 8
33. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C.C., Qiao, Y., Tang, X.: Esrgan: Enhanced super-resolution generative adversarial networks (2018) 3
34. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation (2023), https://arxiv.org/abs/2305.16213 1
35. Wu, K., Liu, F., Cai, Z., Yan, R., Wang, H., Hu, Y., Duan, Y., Ma, K.: Unique3d: High-quality and efficient 3d mesh generation from a single image (2024), https://arxiv.org/abs/2405.20343 3
36. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., Holynski, A.: Reconfusion: 3d reconstruction with diffusion priors (2023) 1

37. Xu, D., Yuan, Y., Mardani, M., Liu, S., Song, J., Wang, Z., Vahdat, A.: Agg: Amortized generative 3d gaussians for single image to 3d (2024) 1
38. Xu, Y., Park, T., Zhang, R., Zhou, Y., Shechtman, E., Liu, F., Huang, J.B., Liu, D.: Videogigagan: Towards detail-rich video super-resolution (2024), https://arxiv.org/abs/2404.12388 3
39. Yoon, Y., Yoon, K.J.: Cross-guided optimization of radiance fields with multi-view image super-resolution for high-resolution novel view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12428–12438 (June 2023) 2, 3, 4
40. Yu, J., Zhu, H., Jiang, L., Loy, C.C., Cai, W., Wu, W.: Painthuman: Towards high-fidelity text-to-3d human texturing via denoised score distillation (2023), https://arxiv.org/abs/2310.09458 1
41. Yu, X., Zhu, H., He, T., Chen, Z.: Gaussiansr: 3d gaussian super-resolution with 2d diffusion priors (2024), https://arxiv.org/abs/2406.10111 3
42. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting (2023), https://arxiv.org/abs/2307.12348 3
43. Zeng, B., Li, S., Feng, Y., Li, H., Gao, S., Liu, J., Li, H., Tang, X., Liu, J., Zhang, B.: Ipdreamer: Appearance-controllable 3d object generation with image prompts (2024) 1