LDM3D-VR: Latent Diffusion Model for 3D VR

Gabriela Ben Melech Stan **Diana Wofk** **Estelle Aflalo**
Shao-Yen Tseng **Zhipeng Cai** **Michael Paulitsch** **Vasudev Lal**
Intel Labs
{gabriela.ben.melech.stan, diana.wofk, estelle.aflalo,
shao-yen.tseng, zhipeng.cai, michael.paulitsch, vasudev.lal}@intel.com

Abstract

Latent diffusion models have proven to be state-of-the-art in the creation and manipulation of visual outputs. However, as far as we know, the generation of depth maps jointly with RGB is still limited. We introduce LDM3D-VR, a suite of diffusion models targeting virtual reality development that includes LDM3D-pano and LDM3D-SR. These models enable the generation of panoramic RGBD based on textual prompts and the upscaling of low-resolution inputs to high-resolution RGBD, respectively. Our models are fine-tuned from existing pretrained models on datasets containing panoramic/high-resolution RGB images, depth maps and captions. Both models are evaluated in comparison to existing related methods.

1 Introduction

Diffusion models have brought about a shift in content creation by offering accessible models that generate RGB images from text prompts or produce high-resolution RGB images from low-resolution inputs. However, the generation of depth maps jointly with RGB images, often required for virtual reality (VR) content development, still presents a challenge. Panoramas created using conventional image stitching algorithms exhibit certain drawbacks, such as artifacts and irregular shapes.

LDM3D-VR builds upon the Latent Diffusion Model for 3D (LDM3D) [39] and explores RGBD generation for panoramic views. We also create a super-resolution model based on LDM3D that jointly upscales an image alongside its corresponding depth map. To summarize, our contributions are the following: (i) we introduce LDM3D-pano, which addresses the challenge of generating panorama views jointly with their corresponding depth maps based on a text prompt (ii) we introduce LDM3D-SR that performs x4 upscaling and recovers high-resolution RGB and depth maps from low-resolution inputs (iii) we showcase this work through a demo accessible at `https://huggingface.co/spaces/Intel/ldm3d`. LDM3D-pano and LDM3d-SR are available at `https://huggingface.co/Intel`.

2 Related work

Text-to-panorama. Text-to-panorama is an important task for creating VR environments. Early approaches achieve this goal through Generative Adversarial Networks (GANs) [24, 6]. Recent advances in diffusion models have improved training stability and model generalization capacity compared to GAN, enabling text-to-panorama works [3, 40, 18]. Some of these methods [40, 3] only cover the left right rotations of panorama, i.e., without top down views. Others [18] cannot generate realistic panoramas due to the lack of training data. We propose LDM3D-pano, a novel diffusion-based approach capable of producing, from an input text prompt, a realistic RGB panorama and its corresponding panorama depth map.

Super-resolution for images. Learning-based super-resolution (SR) has been extensively studied in the past decade. Following [10], initial methods adopted convolution neural networks (CNN)

Neural Information Processing Systems (NeurIPS) 2023 Workshop on Diffusion Models.

and proposed various methods to improve reconstruction quality [17, 16, 50, 23, 49]. GANs were later introduced, which led to higher fidelity SR images [19, 42, 41]. Subsequent approaches then improved SR performance through the use of attention [49, 8, 27] and transformer-based architectures [22, 26, 43, 25, 5]. Most recently, denoising diffusion probabilistic models [14] have demonstrated proficiency in image generation [29, 33] as well as image upscaling [35, 38, 12].

Super-resolution for depth. Enhancing the resolution of a depth map is also a widely-studied problem. Naive pixel-level interpolations often yield noisy floating points at object boundaries. Learning-based approaches [44, 21, 13, 45, 51, 1] have emerged as promising alternatives.
In this work, we propose LDM3D-SR, a latent diffusion-based super-resolution model that can enhance the resolution of RGB and depth maps within the same architecture.

3 Methodology

3.1 LDM3D-pano

LDM3D-pano extends LDM3D [39] to panoramic image generation. Key changes to the architecture include adjustments to the first and last Conv2d layers of the KL-autoencoder [11], enabling it to process a 4-channel input consisting of RGB concatenated with a single-channel depth map; we denote this model as LDM3D-4c. The employed diffusion model is based on U-Net [32] operating in a 64x128x4 latent space, following [31], with the incorporation of a CLIP text encoder [28] for text conditioning through cross-attention mapping on the U-Net layers.

We adopt a two-stage fine-tuning procedure, following [29, 39]. We first fine-tune the refined version of the KL-autoencoder in LDM3D-4c, using roughly 10k samples of size 256x256 sourced from LAION-400M [37], with depth map labels produced using the DPT-BEiT-L-512 [4].

Subsequently, the U-Net backbone is fine-tuned based on Stable Diffusion (SD) v1.5 [29], employing a subset of LAION Aesthetics 6+ [36] consisting of nearly 20k tuples (captions, 512x512-sized images and depth maps produced using DPT-BEiT-L-512 [4]). We further fine-tune the U-Net on our panoramic dataset, comprised of High Dynamic Range (HDR) images–originally in 4k resolution–sourced from [47], [15]. These HDRIs are augmented into 512x1024 panoramic images utilizing the methodology from [7], producing 7828 training images and 322 validation images. Panoramic depth maps labels at 512x1024 resolution are produced using DPT-BEiT-L-512 [4]) Captions are generated using BLIP-2 [20]. Of the resulting captions, $\sim 70\%$ start with "360 view of" and $\sim 4\%$ with "panoramic view of," while the remaining captions do not feature a panorama-related mention.

3.2 LDM3D-SR

LDM3D-SR specializes in super-resolution, utilizing the KL-AE previously developed for LDM3D-4c 3.1 to now encode low-resolution (LR) images into a 64x64x4 dimensional latent space. The diffusion model used here is an adapted version of the U-Net referenced in 3.1, now modified to have an 8-channel input. This change enables conditioning on LR latent via concatenation to the high-resolution (HR) latent during training, and to noise during inference. Text conditioning is also facilitated using cross attention with a CLIP text encoder.

We finetune the U-Net in LDM3D-SR from SD-superres [2]. Training data consists of HR and LR sets with 261,045 samples each. For HR samples, we use a subset of LAION Aesthetics 6+ with tuples (captions, 512x512-sized images, and depth maps from DPT-BEiT-L-512 [4]). LR images are generated using a lightweight BSR-image-degradation method, introduced in [29] applied to the HR image. We explored three methods for generating LR depth maps: performing depth estimation on the LR depth maps (LDM3D-SR-D), utilizing the original HR depth map for LR conditioning (LDM3D-SR-O), and applying bicubic degradation to the depth map (LDM3D-SR-B).

4 Results

4.1 Panoramic RGBD generation

We evaluate text-to-pano RGBD generation using the validation set of our dataset (see 3.1).

Image evaluation. We compare LDM3D-pano to Text2light LDR [7], a model that creates a text-driven Low Dynamic Range panorama using a hierarchical approach for detail rendering, where global text-scene alignment is followed by a local sampler to facilitate patch-based panorama synthesis.

For image quality assessment, we utilize Frechet Inception Distance (FID), Inception Score (IS), and CLIP similarity; these metrics are summarized in Table 1. LDM3D-pano achieves higher FID, and comparable IS and CLIPsim compared to Text2light. LDM3D-pano's higher FID may be due to its deficiency in local awareness and the absence of training in patch-based semantic coherence; focusing on the overall, global context of the given text, potentially at the expense of finer, localized details. Nonetheless, by leveraging extensive text-to-image pretraining from [29, 39], LDM3D-pano has the capacity to generate a diverse range of images, as is reflected by its marginally higher IS and CLIPsim scores, and in visualized samples in Figure 1.

Depth evaluation. For panoramic depth evaluation, we compare LDM3D-pano to a baseline monocular panorama depth estimation model: Joint_3D60_Fres model [46]. Since diffused RGBD outputs have no ground truth depth available, we use DPT-BEiT-L-512 [4] depth as reference.

The evaluated depth maps and the reference depth are all in disparity space and thus non-metric. We primarily consider the mean absolute relative error (MARE). We fit depth estimates to the reference via least-squares over 500 randomly sampled points. This aims to rescale and reshift the depth estimates to be more closely aligned with the reference. We then compute the MARE and summarize results Table 2. As explained in Figure 2, we also report the MARE computed while excluding outlier samples where error exceeds the 90th percentile. In both cases, LDM3D-pano achieves lower MARE than the baseline panoramic depth estimation model.

Table 1: Text-to-pano image metrics at 512x1024, evaluated on 332 samples from our validation set.

Method	FID ↓	IS ↑	CLIPsim ↑
Text2light[7]	**108.30**	4.646±0.27	27.083±3.65
LDM3D-pano	118.07	**4.687±0.50**	**27.210±3.24**

Table 2: Pano depth metrics at 512x1024. Reference depth is from DPT-BEiT-L-512.

Method	MARE ↓	≤90%ile
Joint_3D60[46]	1.75±2.87	0.92±0.87
LDM3D-pano	**1.54±2.55**	**0.79±0.77**

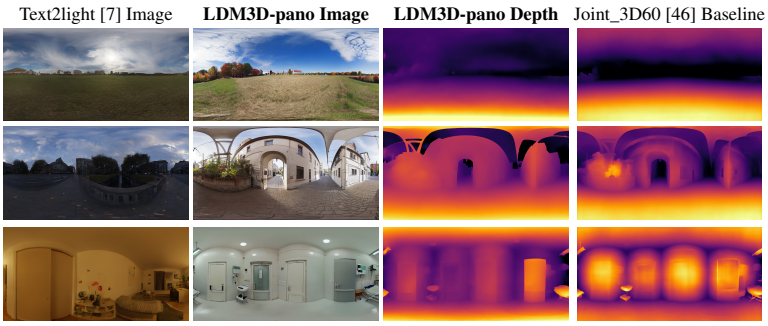

Figure 1: Qualitative comparison of text-to-panoramic RGBD generation at 512x1024. Images are compared with Text2light [7]. Depth maps are compared to Joint_3D60[46]. Captions: top–"a 360 view of a field with a few buildings in the distance," middle–"a 360 view of a city street with a bridge," bottom–"a 360 view of a hospital room."

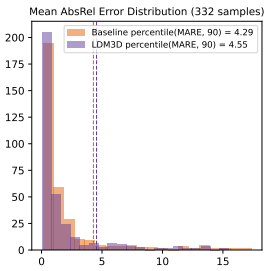

Figure 2: Error distribution across evaluated samples shows a long tail. This motivates removing outliers with error above the 90th percentile.

4.2 High-resolution RGBD generation

We evaluate HR-RGBD generation using a subset from ImageNet-Val [9] composed of 2243 samples at 512x512 resolution. The LR validation set is constructed via bicubic downscaling of HR-RGBD.

Image evaluation. In line with previous studies, we use reconstruction FID, IS, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM) to evaluate the quality of HR images; these metrics are summarized in Table 3. We compare LDM3D-SR against bicubic regression, SDx4 super-resolution based on SDv2 [29], LDMx4 [30] and SD-superres based on SDv1 [2]. The latter is the model from which we fine-tune LDM3D-SR. Our findings reveal that LDM3D-SR achieves the best FID and the second-best IS after SDx4. Conversely, bicubic regression attains the highest

scores on PSNR and SSIM. However, these two metrics tend to prefer blurriness over misaligned high-frequency details [34] and often contradict human perception [48]; this is supported by the significantly lower FID score of bicubic regression as well as by the visualization in Figure 3.

In comparing LDM3D-SR-D, LDM3D-SR-O, and LDM3D-SR-B, Table 3 also presents an ablation study on the optimal depth preprocessing method outlined in 3.2. Results reveal that employing bicubic degradation of the initial depth map is closely aligned with utilizing the original HR depth map as conditioning. Conversely, utilizing a depth map calculated from the degraded image yields inferior results, likely due to low quality depth maps.

Depth evaluation. Our depth evaluation protocol for LDM3D-SR closely follows that used for LDM3D-pano, as described in 4.1. Our baseline here is bicubic regression on depth. Table 3 reports the MARE for bicubic regression and our LDM3D-SR methods. The error in bicubically-regressed depth maps is predominantly along object boundaries that become blurred upon bicubic interpolation; since edges account for a small fraction of scene content, the MARE for bicubic regression is particularly low. Amongst the LDM3D-SR methods, -D exhibits the highest MARE while -O and -B both exhibit a lower MARE. We show additional visualizations of RGBD upscaling in Figure 4, where we observe high-resolution features in both the images and depth maps (the wings and antennae of the grasshopper, the threads of the screw). Lastly, Figure 5 does not indicate the presence of a tail in the error distribution, so we do not perform any outlier removal in this evaluation.

Table 3: x4 upscaling from 128x128 to 512x512, evaluated on 2243 samples from ImageNet-Val

Method	FID ↓	IS ↑	PSNR ↑	SSIM ↑	Depth MARE ↓
Regression, bicubic	24.686	60.135±4.16	**26.424±3.98**	**0.716±0.13**	**0.0153±0.0189**
SDx4[29]	15.865	**61.103±3.48**	24.528±3.63	0.631±0.15	N/A
LDMx4[30]	15.245	60.060±3.88	25.511±3.94	0.686±0.16	N/A
SD-superres[2]	15.254	59.789±3.53	23.878±3.28	0.642±0.15	N/A
LDM3D-SR-D	15.522	59.736±3.37	24.113±3.54	0.659±0.16	0.0753±0.0734
LDM3D-SR-O	14.793	60.260±3.53	24.498±3.59	0.665±0.16	0.0530±0.0496
LDM3D-SR-B	**14.705**	60.371±3.56	24.479±3.58	0.665±0.48	0.0537±0.0506

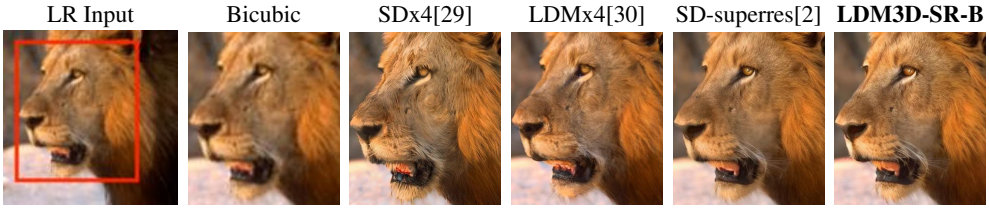

Figure 3: Qualitative comparison of x4 upscaling. Image sourced from ImageNet-Val.

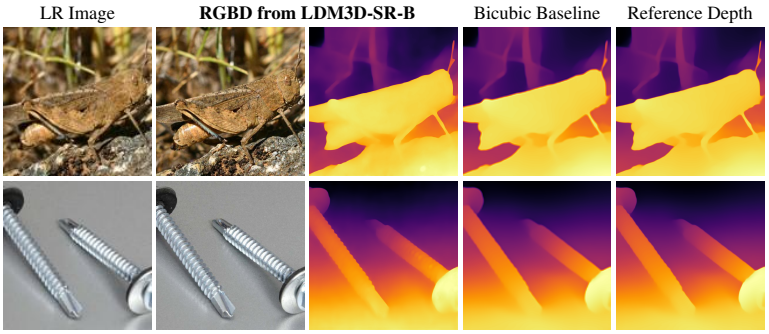

Figure 4: Additional visualization of x4 RGBD upscaling with LDM3D-SR-B. Depth maps are compared to a bicubic regression baseline against reference depth obtained using DPT-BEiT-L-512.

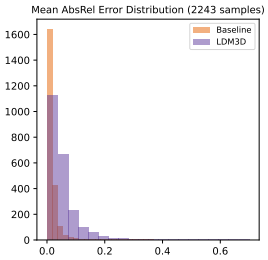

Figure 5: Since there is no long tail in the error distribution across evaluated ImageNet samples, no outlier removal is performed.

5 Conclusion

We introduce LDM3D-pano and LDM3D-SR for 3D VR applications. LDM3D-pano competes with panorama-specialized models by generating diverse high-quality panoramic images jointly with panoramic depth. LDM3D-SR focuses on RGBD upscaling, outperforming related image upscaling methods while also generating high-resolution depth maps. Future work could combine these domains to generate high-resolution panorama RGBD to further enhance immersive VR experiences.

References

[1] Ido Ariav and Israel Cohen. Fully cross-attention transformer for guided depth super-resolution. *Sensors*, 23(5):2723, 2023.

[2] Justin Pinkney at Lambda Labs. Super Resolution Lambda Labs. `https://huggingface.co/lambdalabs/stable-diffusion-super-res`. Accessed: 2023-05-30.

[3] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation, 2023.

[4] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023.

[5] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 22367–22377, June 2023.

[6] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022.

[7] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation, 2023.

[8] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proc. of the IEEE/CVF conf. on computer vision and pattern recognition*, pages 11065–11074, 2019.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

[11] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021.

[12] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10021–10030, June 2023.

[13] Jiaxin Guo, Rong Xiong, Yongsheng Ou, Lin Wang, and Chao Liu. Depth image super-resolution via two-branch network. In *Cognitive Systems and Information Processing: 6th International Conference, ICCSIP 2021, Suzhou, China, November 20–21, 2021, Revised Selected Papers 6*, pages 200–212. Springer, 2022.

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

[15] iHDRI. iHDRI.COM lighting for creatives. `https://www.ihdri.com`. Accessed: 2023-05-30.

[16] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proc. of the IEEE conf. on computer vision and pattern recognition*, pages 1646–1654, 2016.

[17] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *Proc. of the IEEE conf. on computer vision and pattern recognition*, pages 1637–1645, 2016.

[18] Blockade Lab. Blockade lab. `https://www.blockadelabs.com/`.

[19] 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 *Proc. of the IEEE conf. on computer vision and pattern recognition*, pages 4681–4690, 2017.

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[21] Tao Li, Hongwei Lin, Xiucheng Dong, and Xiaohua Zhang. Depth image super-resolution using correlation-controlled color guidance and multi-scale symmetric network. *Pattern Recognition*, 107:107513, 2020.

[22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using Swin Transformer. In *Proc. of the IEEE/CVF int. conf. on computer vision*, pages 1833–1844, 2021.

[23] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proc. of the IEEE conf. on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[24] Chieh Hubert Lin, Hsin-Ying Lee, Yen-Chi Cheng, Sergey Tulyakov, and Ming-Hsuan Yang. Infinitygan: Towards infinite-pixel image synthesis. *arXiv preprint arXiv:2104.03963*, 2021.

[25] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. Transformer for single image super-resolution. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 457–466, June 2022.

[26] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 3517–3526, 2021.

[27] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proc., Part XII 16*, pages 191–207. Springer, 2020.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

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

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Latent diffusion models (LDM) for super-resolution – compvis/ldm-super-resolution-4x-openimages. `https://huggingface.co/CompVis/ldm-super-resolution-4x-openimages`. Accessed: 2023-05-30.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[34] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021.

[35] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

[37] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[38] Shuyao Shang, Zhengyang Shan, Guangxing Liu, and Jinglin Zhang. Resdiff: Combining cnn and diffusion model for image super-resolution. *arXiv preprint arXiv:2303.08714*, 2023.

[39] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, and Vasudev Lal. Ldm3d: Latent diffusion model for 3d, 2023.

[40] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint arXiv:2307.01097*, 2023.

[41] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proc. of the IEEE/CVF int. conf. on computer vision*, pages 1905–1914, 2021.

[42] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proc. of the European conf. on computer vision (ECCV) workshops*, pages 0–0, 2018.

[43] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proc. of the IEEE/CVF conf. on computer vision and pattern recognition*, pages 17683–17693, 2022.

[44] Jun Xie, Rogerio Schmidt Feris, and Ming-Ting Sun. Edge-guided single depth image super resolution. *IEEE Trans. on Image Processing*, 25(1):428–438, 2015.

[45] Yuxiang Yang, Qi Cao, Jing Zhang, and Dacheng Tao. CODON: on orchestrating cross-domain attentions for depth super-resolution. *Int. Journal of Computer Vision*, 130(2):267–284, 2022.

[46] Ilwi Yun, Hyuk-Jae Lee, and Chae Eun Rhee. Improving 360 monocular depth estimation via non-local dense prediction transformer and joint supervised and self-supervised learning. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 36, pages 3224–3233, 2022.

[47] Greg Zaal and Rob Tuytel et al. Poly Haven. `https://polyhaven.com/`. Accessed: 2023-05-30.

[48] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

[49] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proc. of the European conf. on computer vision (ECCV)*, pages 286–301, 2018.

[50] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proc. of the IEEE conf. on computer vision and pattern recognition*, pages 2472–2481, 2018.

[51] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 5697–5707, 2022.