Generating Artwork-Style Synthetic Data for Image Super-Resolution

Anonymous CVPR submission

Paper ID 20

Abstract

Existing image super-resolution (SR) datasets predominantly rely on web-scraped natural images—photographs of real-world scenes—due to their abundance online. However, this reliance hinders SR performance in specialized domains such as artwork, which comprises artificially created visuals like posters and book covers that incorporate text. This limitation arises from the difficulty of obtaining a sufficient number of uncompressed, high-resolution artwork images from the web, as such content is scarce and often subject to copyright restrictions. To address this issue, we propose a synthetic dataset construction pipeline that leverages advanced text-to-image (T2I) diffusion models to generate high-quality artwork images. Using this pipeline, we construct Generated Artwork dataset for image Super-Resolution (GASR), a dataset specifically tailored for SR on artwork images. Although GASR-DF2K contains only ***16%*** *as many images as LSDIR, a widely used large-scale SR dataset, it consistently outperforms it on the artwork-centric benchmark Manga109. These results demonstrate the effectiveness of tailored synthetic data in bridging the domain gap and substantially improving SR performance on artwork.*

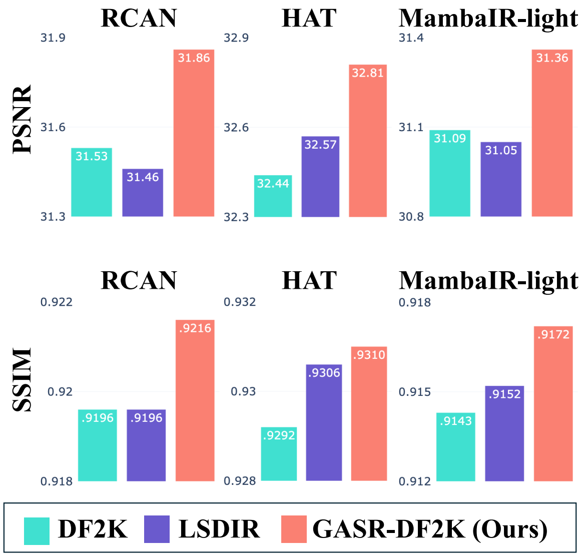


Figure 1. Performance comparison of SR models trained exclusively on conventional datasets (DF2K, LSDIR) versus our proposed dataset combining synthetic and real images (GASR-DF2K). Evaluated on Manga109, an artwork-centric benchmark, GASR-DF2K demonstrates superior SR performance, highlighting the effectiveness of incorporating domain-specific synthetic data to overcome domain biases.

1. Introduction

Image Super-Resolution (SR) is a fundamental problem in computer vision that aims to reconstruct high-resolution images from low-resolution inputs. With the rise of high-resolution displays and the growing demand for high-quality digital content, SR has extended its applications across entertainment, commercial sectors, and a wide range of domains [20–22].

With advancements in deep learning, numerous deep learning-based SR models have been proposed [4, 5, 7, 8, 24, 25]. Consequently, the importance of training datasets has also increased. In general, high-resolution images with minimal compression noise are considered suitable for SR training. To meet these requirements, efforts have been made to construct high-quality datasets. Notable examples include DIV2K [2], DF2K, which integrates DIV2K and Flickr2K [19], and more recently, LSDIR [13].

However, a key limitation of these SR dataset construction methods is their dependence on web scraping, which restricts coverage of certain domains (e.g., artworks), introducing domain biases. For instance, while DIV2K contains a diverse set of images, including people, urban and rural scenes, flora and fauna, and natural landscapes, it is noticeably lacking in specific content such as artworks. Similarly, LSDIR collects images using popular tags from the photo-sharing platform Flickr and labels from ImageNet [6]. However, due to keyword tendencies and the nature of images uploaded by Flickr users, it is predominantly composed of natural images, which are photographs of real-world scenes, resulting in a similar domain bias.

Here, “artwork” refers to visual content that integrates text and illustrations, such as book covers, posters, and digital art. Compared to natural images, artworks exhibit distinct characteristics, including consistent color distributions and well-defined line drawings. Training SR models specifically on artwork images enables domain-aware SR, which is expected to enhance edge restoration accuracy, improve fine detail reconstruction, and increase text readability.

Constructing a dataset with domain-specific images can improve SR performance within that domain. However, this approach presents several challenges. In particular, collecting a sufficient number of high-quality images for a specific domain is costly. Additionally, most images available on the web are provided in compressed formats such as JPEG, leading to a shortage of high-quality samples suitable for SR training. As a result, conventional data collection methods struggle to obtain a sufficient number of high-resolution images. This issue is even more pronounced for artwork, where copyright restrictions further complicate the acquisition of suitable training data for SR.

To address these challenges, this study proposes a novel SR dataset construction pipeline leveraging a text-to-image (T2I) diffusion model, overcoming the limitations of traditional web scraping methods. Synthetic data generated by T2I models is free from copyright restrictions and avoids common issues in web images, such as compression artifacts and low resolution, making it ideal for SR training. The main contributions of this study are as follows:

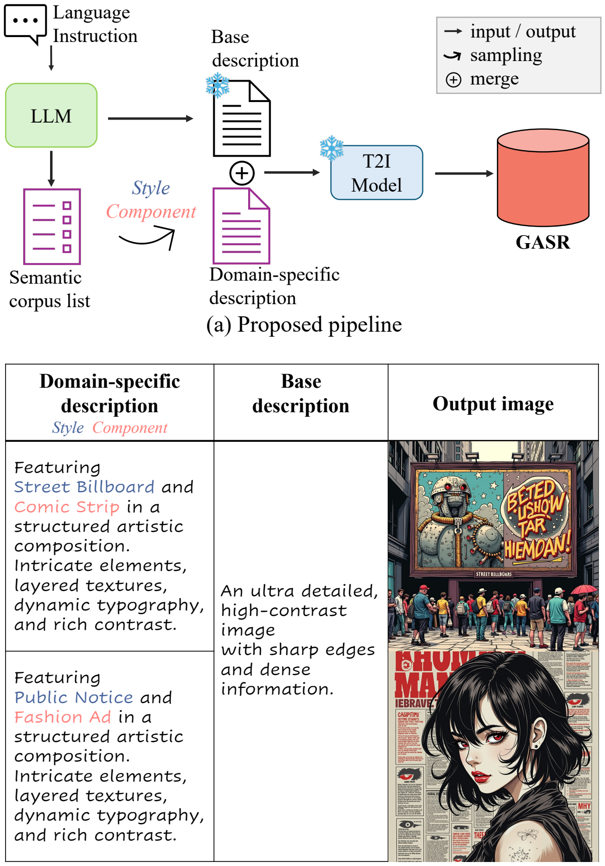
Technical Contribution. This study proposes a pipeline for constructing domain-adaptive SR datasets. Our approach leverages text prompts specifically designed to incorporate domain-specific information, along with a pre-trained generative model. This enables the low-cost generation of images from domains that were insufficiently represented in conventional datasets.

Dataset Contribution. We propose Generated Artwork dataset for image Super-Resolution (GASR), an effective dataset for artwork-oriented SR. Despite containing only 16% of the number of images in LSDIR, GASR-DF2K, a combination of GASR and DF2K, outperformed LSDIR on Manga109, a representative benchmark composed of artwork images, demonstrating its efficiency in enhancing SR performance for artwork images (see Figure 1).

Experimental Contribution. To reveal the intrinsic domain differences between SR datasets, we analyzed the semantic distribution of each dataset. Visualization showed that while conventional datasets do not sufficiently cover artwork images, GASR effectively includes a diverse range of artwork images. These findings demonstrate that our pipeline enables task-optimized, cost-effective training.

2. GASR Dataset

In this section, we introduce GASR, an SR dataset designed for artwork. The overall pipeline for image generation is illustrated in Figure 2 (a). Our pipeline leverages a large language model (LLM) to generate diverse prompts, enabling cost-effective, controllable image synthesis with a T2I model.



(a) Proposed pipeline

Domain-specific description *Style Component*	Base description	Output image
Featuring Street Billboard and Comic Strip in a structured artistic composition. Intricate elements, layered textures, dynamic typography, and rich contrast.	An ultra detailed, high-contrast image with sharp edges and dense information.	
Featuring Public Notice and Fashion Ad in a structured artistic composition. Intricate elements, layered textures, dynamic typography, and rich contrast.		

(b) Example of generated texts and images

Figure 2. Overview of the proposed data synthesis pipeline: (a) Pipeline diagram showing synthetic image generation to supplement domain-specific data, (b) Examples of generated text prompts with their corresponding synthesized images.

Figure 3. Example images from GASR. The dataset consists of art-style images generated by a T2I model.

2.1. Prompt Generation

To generate artwork images suitable for SR training, we structure prompts into two key components: base description and domain-specific description (see Figure 2(b)). Systematic incorporation of these elements enables efficient and structured prompt generation, allowing for controlled diversity in the output images.

Base Description. The base description is applied consistently to all images in the dataset, ensuring overall uniformity and embedding common characteristics suitable for SR tasks. Specifically, it is designed to enhance key SR-related attributes such as texture details, edge sharpness, and information density, enabling the generation of images optimized for SR training.

Domain-Specific Description. The domain-specific description is designed to incorporate semantic elements suitable for artwork images, enabling the generation of diverse styles. In this study, we construct a semantic corpus list using GPT-4o [1], which enumerates *Style*, defining composition and expression, and *Component*, defining subject. This ensures unique prompts while covering diverse styles. *Style* and *Component* are dynamically selected from a predefined semantic corpus list.

2.2. Image Generation

For image generation, we utilize FLUX.1-dev [12], as it is capable of generating high-resolution images (1024×1024). The model takes the prompts generated in Section 2.1 as input to synthesize images. As shown in Figure 3, the resulting GASR dataset consists of artwork images and is designed to capture diverse visual characteristics of artwork.

2.3. Complementary Real-Image Datasets

The GASR dataset provides diverse and high-quality generated artwork images for SR model training. However, an inherent domain gap exists between generated and real images [9], which may limit the generalization performance of SR models. To address this issue, we incorporate real images from the DF2K dataset as complementary data. DF2K is a widely used SR dataset consisting of high-resolution images with faithfully preserved pixel-level details, and it helps bridge the domain gap by providing realistic image content for training.

2.4. Dataset Analysis

Pixel Quality. We analyze the resolution, compression noise, and structural diversity of existing SR datasets to assess their suitability for training (see Table 1). In terms of resolution, the proposed GASR dataset is comparable to LSDIR, ensuring sufficient detail for SR training.

To evaluate compression noise, we use the blockiness measure [3], which quantifies the intensity of JPEG compression artifacts. Higher values indicate more severe compression artifacts. The proposed GASR dataset achieves a blockiness value comparable to other SR datasets, indicating that the generated images are not affected by compression noise.

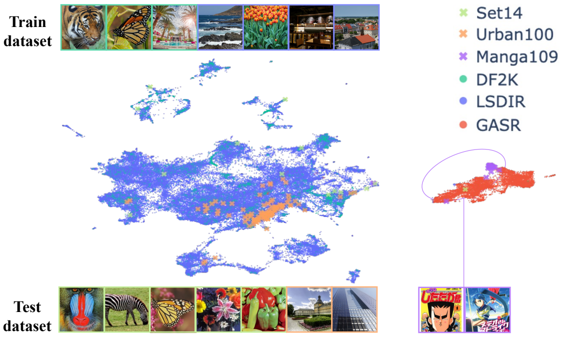


Figure 4. UMAP [15] visualization of embeddings obtained from the CLIP image encoder [17].

Table 1. Dataset statistics. #Images: Number of images, #Pixels: Average number of pixels per image, Blockiness: Median blockiness value, #Segments: Average number of segments per image.

Dataset	#Images	#Pixels	Blockiness	#Segments
DIV2K [2]	800	2.8M	0.47	104
DF2K [2, 19]	3,450	2.8M	0.47	103
LSDIR [13]	84,991	1.1M	0.82	92
GASR (Ours)	10,304	1.0M	0.87	156

For structural diversity, we use the average number of segments per image [16], computed using SAM [11]. A higher segment count reflects more diverse content and less uninformative low-frequency region during training. GASR is designed to reduce such redundancy and achieves a higher segment count than other SR datasets.

Overall, GASR provides high resolution, minimal compression noise, and rich structural diversity, all of which contribute to more efficient training and improved SR performance.

Semantics. To evaluate the suitability of GASR for SR training dataset, we analyze its semantic characteristics. Figure 4 visualizes the distribution of datasets. Datasets primarily composed of natural images, such as DF2K and LSDIR, exhibit similar distributions, whereas Manga109 shows a significantly different distribution. The proposed GASR dataset demonstrates a distribution close to that of Manga109, confirming its high semantic consistency with artwork-based datasets.

3. Experiments

3.1. Experimental Setting

Models. We use three models: the CNN-based RCAN [24], the Transformer-based HAT [5], and the Mamba-based

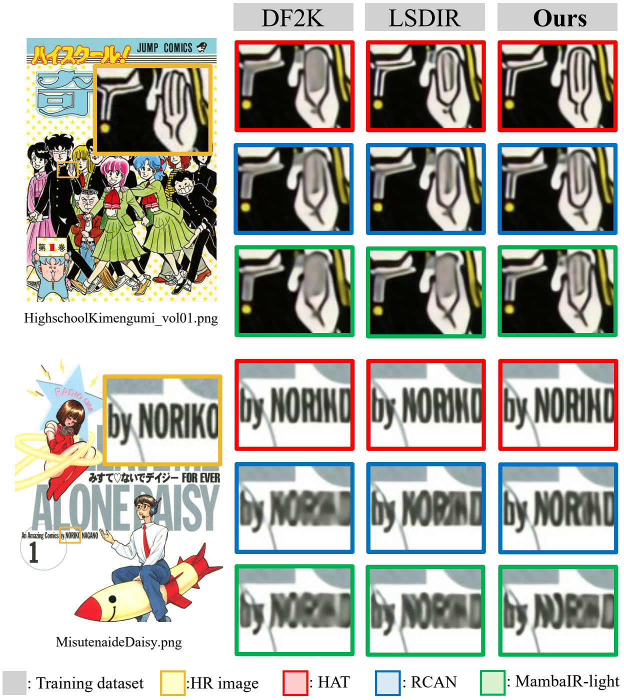


Figure 5. Visual comparison of SR models trained on DF2K, LSDIR, and GASR-DF2K (Ours).

MambaIR-light [8].

Training Datasets. We compare DF2K [2, 19] and LSDIR [13] with GASR-DF2K, which combines GASR and DF2K.

Evaluation Datasets. We use Set14 [23], Urban100 [10], and Manga109 [14].

Evaluation Metrics. PSNR and SSIM are calculated on the Y (luminance) channel of the YCbCr color space.

Implementation Details. We primarily follow the training settings of the original SR model papers [5, 8, 24]. For RCAN, considering the scale of the training data, we train for 500k iterations. The scale factor is set to ×4, and HR-LR pairs are generated using bicubic downsampling. Data augmentation includes random rotations of 90°, 180°, and 270°, along with horizontal flipping.

3.2. Experimental Results

Main Results. To evaluate the effectiveness of the proposed GASR-DF2K dataset for artwork-oriented SR, we compare it with existing datasets DF2K and LSDIR. As shown in Table 2, GASR-DF2K achieves the best performance on the artwork benchmark Manga109 across all models, despite containing fewer images than LSDIR. Moreover, as shown in Figure 5, models trained on GASR-DF2K show significant improvements in edge sharpness and text clarity compared to those trained on other datasets.

Effect of Real Dataset Usage. To demonstrate the effectiveness of integrating real images, we compare models trained with and without real data. As shown in Table 3, GASR-DF2K, which combines GASR and DF2K, outperforms DF2K alone, highlighting the benefits of integrating generated images. In contrast, GASR, trained solely on generated images, performs worse than DF2K, underscoring the importance of incorporating real images. These results suggest that appropriately integrating real and generated images is crucial to improve SR performance.

Table 2. Performance comparison of SR models trained on DF2K, LSDIR and GASR-DF2K (Ours).

Model (Params)	Dataset	#Images	Urban100 PSNR	 SSIM	Manga109 PSNR	 SSIM
MambaIR-light [8] (924k)	DF2K	3,450	26.53	0.7987	31.09	0.9143
	LSDIR	84,991	**26.66**	**0.8017**	31.05	0.9152
	GASR-DF2K	13,754	26.59	0.7993	**31.36**	**0.9172**
RCAN [24] (15.5M)	DF2K	3,450	26.89	0.8093	31.53	0.9196
	LSDIR	84,991	**27.08**	**0.8133**	31.46	0.9196
	GASR-DF2K	13,754	26.97	0.8101	**31.86**	**0.9216**
HAT [5] (20.7M)	DF2K	3,450	27.93	0.8365	32.44	0.9292
	LSDIR	84,991	**28.45**	**0.8469**	32.57	0.9306
	GASR-DF2K	13,754	28.05	0.8378	**32.81**	**0.9310**

Table 3. Effect of real data (MambaIR-light)

Dataset	Set14 PSNR	 SSIM	Urban100 PSNR	 SSIM	Manga109 PSNR	 SSIM
GASR	28.59	0.7833	26.31	0.7938	30.64	0.9107
DF2K	28.76	0.7853	26.53	0.7987	31.09	0.9143
GASR-DF2K	**28.79**	**0.7858**	**26.59**	**0.7993**	**31.36**	**0.9172**

Table 4. Comparison of T2I models (MambaIR-light, GASR-DF2K).

Model	Set14 PSNR	 SSIM	Urban100 PSNR	 SSIM	Manga109 PSNR	 SSIM
SD-2.1 [18]	28.73	0.7834	26.42	0.7926	31.04	0.9117
FLUX.1-dev [12]	**28.79**	**0.7858**	**26.59**	**0.7993**	**31.36**	**0.9172**

Effect of T2I Model Versions. Table 4 compares FLUX.1-dev with Stable Diffusion 2.1 (SD-2.1) [18] for generating SR training data. FLUX.1-dev consistently outperforms SD 2.1 across all benchmarks. This improvement may stem from architectural differences: while SD 2.1 uses a CNN-based U-Net, FLUX.1-dev adopts a Transformer-based U-Net, potentially enabling better consistency and finer detail in generated images. As generative models evolve, further improvements in SR training data quality are expected.

4. Conclusion

In this study, we proposed a dataset construction pipeline leveraging domain-specific generated images and introduced an efficient approach for SR training. Our results show that high performance can be achieved with significantly fewer images than conventional large-scale datasets, highlighting a new and effective direction for SR dataset construction.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 1, 3, 4

[3] Dinesh Bhardwaj and Vinod Pankajakshan. A jpeg blocking artifact detector for image forensics. *Signal Processing: Image Communication*, 68:155–161, 2018. 3

[4] Hanting Chen et al. Pre-trained image processing transformer. In *CVPR*, 2021. 1

[5] Xiangyu Chen et al. Activating more pixels in image super-resolution transformer. In *CVPR*, 2023. 1, 3, 4

[6] Jia Deng and ohters. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[7] Chao Dong et al. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 1

[8] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *ECCV*, 2024. 1, 4

[9] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20555–20565, 2023. 3

[10] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 4

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, 2023. 3

[12] Black Forest Labs. `https://github.com/black-forest-labs/flux`, 2024. 3, 4

[13] Yawei Li et al. Lsdir: A large scale dataset for image restoration. In *CVPRW*, 2023. 1, 3, 4

[14] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76:21811–21838, 2017. 4

[15] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 3

[16] Go Ohtani et al. Rethinking image super-resolution from training data perspectives. In *ECCV*, 2024. 3

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[19] Radu Timofte et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 1, 3, 4

[20] Boyang Wang, Fengyu Yang, Xihang Yu, Chao Zhang, and Hanbin Zhao. Apisr: anime production inspired real-world anime super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25574–25584, 2024. 1

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

[22] Shizhuo Xu, Vibekananda Dutta, Xin He, and Takafumi Matsumaru. A transformer-based model for super-resolution of anime image. *Sensors*, 22(21):8126, 2022. 1

[23] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Proc. 7th Int. Conf. Curves Surf.*, 2010. 4

[24] Yulun Zhang et al. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 1, 3, 4

[25] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *ICCV*, 2023. 1