Arbitrary-Scale Image Generation and Up-sampling using Latent Diffusion Model and Neural Implicit Decoder

Authors: Jinseok Kim, Tae-Kyun Kin, KAIST,
Imperial College London, AI Lab, LG Electronics

ABSTRACT

Super-resolution (SR) and image generation are essential tasks in computer vision. Current methods often produce fixed-scale images with artifacts, limited diversity, and inconsistent scaling, requiring significant memory and time for larger images. This paper introduces a method for generating images at any scale using a pre-trained auto-encoder, a latent diffusion model, and an implicit neural decoder. This approach operates in latent space for efficiency and outputs scale-consistent images using MLPs. Joint denoising and alignment losses improve image quality. Experiments show that our method outperforms existing models in quality, diversity, scale-consistency, speed, and memory use

.Keywords: Super Resolution, Implicit Neural Decoder, Diffusion Models, Image Generation

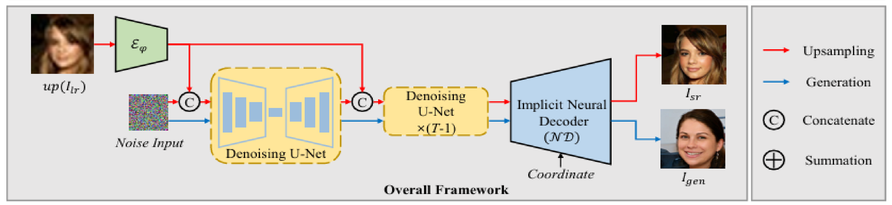
RESULTS

RESULTS FOR LSUN BEDROOM DATASET

Dataset	Resolution	Method	FID	Precision	Recall	Self-SSIM
Bedroom	128	MSPIE	11.39	66.45	26.97	1.00
		Ours	7.20	59.69	20.63	1.00
		Scale Party	10.15	62.50	20.63	1
	160	MSPIE	16.45	63.84	23.09	0.10
		Ours	7.43	58.52	32.12	0.96
		Scale Party	9.85	64.14	22.02	0.92
	192	MSPIE	12.65	58.10	25.93	0.10
		Ours	7.73	59.57	27.98	0.95
		Scale Party	9.91	64.77	21.10	0.89

MODEL ARCHITECHTURE

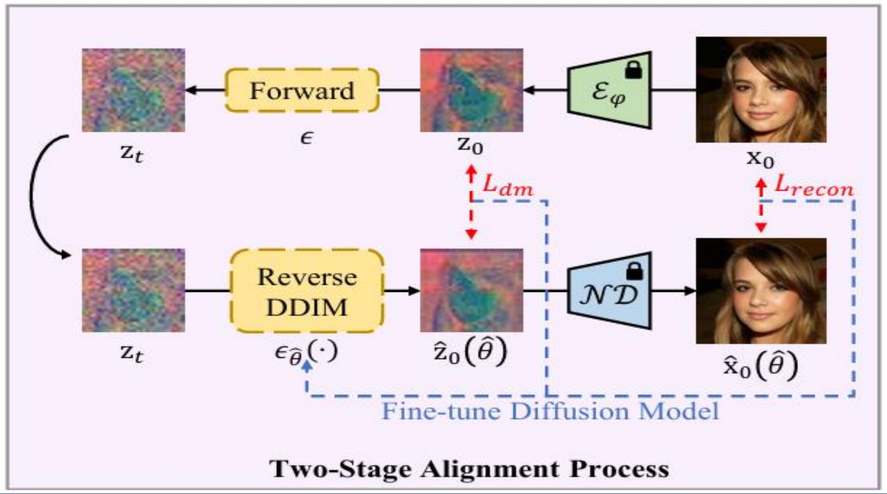
We propose an architecture combining a Latent Diffusion Model (LDM) and Local Implicit Image Function (LIIF) decoder for arbitrary-scale SR and image generation. The pre-trained auto-encoder (with a fixed encoder/decoder) is combined with an MLP-based decoder to map to any scale. The diffusion process in latent space reduces learning complexity and enhances efficiency. Image losses are backpropagated through the decoder to the 0-th and any t-th diffusion steps, aligning image and latent spaces. This approach outperforms other architectures in quality, diversity, and efficiency across benchmarks and tasks..



Overall Framework

TWO-STAGE ALIGNMENT PROCESS

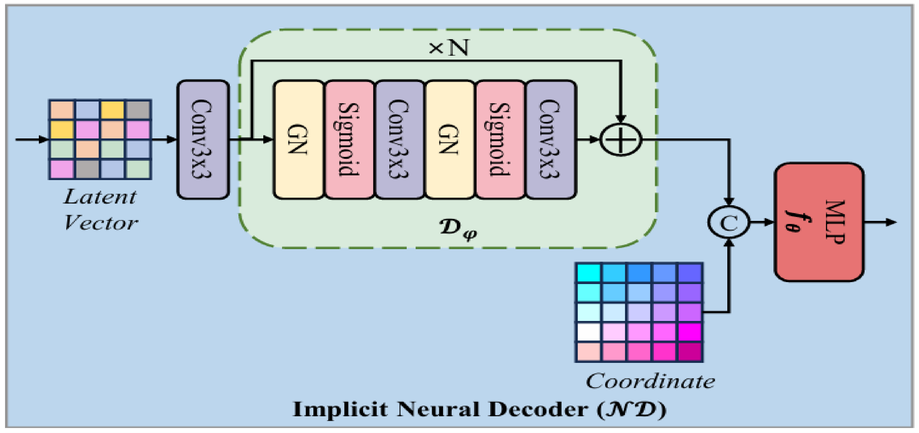
The Latent Diffusion Model (LDM) enhances learning and inference speed compared to pixel-space diffusion models. It relies on features from a pre-trained auto-encoder, which can introduce errors that affect the LDM and, ultimately, decoding effectiveness. To mitigate this, we propose a two-stage alignment process to reduce misalignment between the models, improving output quality. We aim to generate images as close to the ground truth as possible by incorporating loss functions weighted by time steps for better alignment. Additionally, we modified the objective function to combine denoising and reconstruction losses and used a reverse DDIM process for faster sampling.



Two-Stage Alignment Process

ENCODER-DECODER

Our model consists of an encoder, denoising diffusion, and a decoder. The encoder-decoder follows a basic auto-encoder structure using convolutional and transposed convolutional networks. The encoder extracts an image into a latent vector, while the decoder reconstructs it back to the image space. The decoder has a symmetric structure with Res-Blocks and removes the up-sampling layer. An MLP is added to generate arbitrary-scale images by mapping latent features to RGB values and coordinates. The latent vector is interpolated using Euclidean distance, with the MLP modelled using four layers of 256 units.



Implicit Neural Decoder ($\mathcal{ND}$)

IMPLEMENTATION AND DATASET DETAILS

Implementation: We trained the implicit neural decoder and diffusion model using the Adam optimizer with learning rates of 5e-5 and 1e-6, respectively, on a 24GB NVIDIA RTX 4090 GPU.

Evaluation: We tested on human face datasets (FFHQ, CelebA-HQ), general scenes (LSUN), and ultra-high-resolution datasets (DIV2K, Flickr2K). Evaluation metrics included FID, Precision, Recall, and Self-SSIM for image generation and PSNR and LPIPS for SR to assess image quality and perceptual consistency. FPS was used to measure inference speed.

CONCLUSION

We proposed an Implicit Neural Decoder with a latent diffusion model for efficient, arbitrary-scale image generation and up-sampling. Our method enables faster training and inference, improves image quality through a two-stage alignment, and achieves high-fidelity, diverse outputs.

REFERENCES AND CODES

[2403.10255] Arbitrary-Scale Image Generation and Upsampling using Latent Diffusion Model and Implicit Neural Decoder

Official Implementation

Implementation for the paper