Efficient Diffusion Models for Image Super-Resolution

Leonardo Mariut - 1986191
Mohamed Zakaria Benjelloun Tuimy - 2190452
Neural Networks
Sapienza University of Rome

Table of contents

Motivation and Aim

Why Super-Resolution Matters

- SR is widely used in old photo restoration, medical imaging, and satellite image enhancement.
- These tasks need high detail recovery while keeping realistic structure and textures.

Problem with Diffusion-Based SR

- Standard diffusion models are high quality but very slow, requiring hundreds of sampling steps.
- This makes them impractical for real-world or time-sensitive applications.

Our Aim:

- Implement and extend ResShift, a diffusion model that works on the residual between LR and HR.
- Achieve 4x SR in ~15 steps, balancing speed and fidelity.
- Improve ResShift quality while maintaining similar inference time and performance.

Method Overview

ResShift (Baseline Idea)

- Builds a new diffusion Markov chain that bridges LR → HR through the residual.
- Starts from an upsampled LR image (not random noise).
- A custom noise schedule gradually shifts the residual to the target HR.

Why It's Efficient

- Since the LR image already contains most of the structure, the model only needs to add missing details (edges and textures).
- This creates a shorter path to the solution → fewer steps, faster inference, no major quality loss.

Architecture

- Dual-domain UNet, base channels 64, total params ≈1.39M.
- Spatial branch, DCT branch, wavelet branch → fused features to emphasize high-frequency detail.
- Skip connections across the encoder/decoder stages to preserve low-frequency structure and enable sharper reconstruction.
- L1 for fidelity and VGG perceptual loss for textures.

Implementation and Experiments

Dataset Setup

- DIV2K: 800 train / 200 val HR images.
- Training uses random 256x256 HR crops → LR by x4 bicubic downsample, then upsample to 256x256 for y_0 .
- Normalization [0,1] → [-1,1].

Evaluation protocol

- Metrics: PSNR, SSIM on DIV2K validation
- Baseline: bicubic upsampling.

Training Setup

- Predict residual with LR conditioning; η(t) schedule tuned for ~20 steps.
- Example hyperparams: NUM_EPOCHS = 380, Batch = 8, LR = 1e-4, freq-loss weight ≈ 2.0.
- Checkpoints saved every 10 epochs
- Evaluation performed every 10 epochs.
- Hardware: Nvidia GeForce RTX 5060 Ti 16Gb

Results

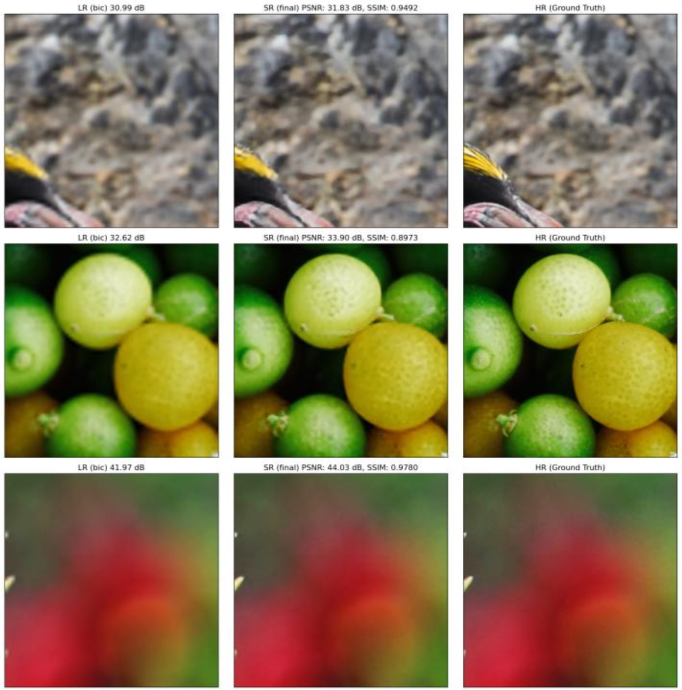
Quantitative Results

- ResShift (15 steps): ~29.5 dB PSNR, 0.85 SSIM.
- Bicubic Baseline: ~28.14PSNR, 0.75 SSIM.
- Confirmed clear fidelity gain from residual diffusion.

Inference Speed

- 15 steps gave a decent quality. Much faster than standard diffusion (100+ steps).
- Quality plateaus around 15–20 steps → extra steps add minimal improvement.

Results-2: Epoch 380 eval && SR image



Conclusions

Goal Achieved && Key Takeaways:

- Implemented ResShift and reproduced main claim: high-quality SR in ~15steps.
- Efficiency: Operating on residual = fewer steps, faster inference.
- Quality: Better textures and sharper details vs bicubic.
- Extensions: Frequency and wavelet domain improved fine detail; continuous scaling adds flexibility.

Bicubic:
PSNR = 25.26 dB
SSIM = 0.7043

SR:
PSNR = 25.04 dB
SSIM = 0.7208

Limitations & Future Work

Model Limitations

- Still slower than one-shot SR methods (CNNs/Transformers) since diffusion always requires multiple steps.
- Training requires careful noise schedule tuning, which makes it complex to reproduce.

Hardware Limitations

- Training and inference were run on limited GPU resources.
- Slower training times and smaller batch sizes restricted experimentation.
- Could not explore larger models or extensive hyperparameter tuning due to compute limits.

Dataset Limitations

- DIV2K dataset is relatively small (800 training images).
- Limited diversity → the model may not generalize perfectly to all real-world scenarios (e.g., medical or satellite images).
- More diverse and larger datasets would likely improve robustness and detail preservation.

Future Directions:

- Implement a training schedule that starts with only L1 for the first N epochs, then gradually adds perceptual and frequency losses.
- Test reducing frequency loss to 0.0 at early steps and linearly increasing it.

References

- **Papers:**

- Liu et al. (2024) – Arbitrary-Steps Image Super-Resolution with Time-Step Schedule & 2D-LUT. arXiv:2412.09013v1.
- Liang et al. (2023) – Implicit Diffusion Models for Continuous Super-Resolution. arXiv:2303.16491v2.
- Zhou et al. (2025) – DMNet/DDMN: Dual-domain Modulation Network for Lightweight Super-Resolution. arXiv:2503.10047v2.
- Johnson et al. (2016) – Perceptual Losses for Real-Time Style Transfer and Super-Resolution. arXiv:1603.08155.
- Mildenhall et al. (2020) – NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. arXiv:2003.08934.
- Ronneberger et al. (2015) – U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597.

- **Slides:**
- https://github.com/pietro-nardelli/sapienza-ppt-template

Thank you for the attention!