Project Report

on

Image Super-Resolution using GANs (SRGAN)

Submitted by

R.Ruthuraraj

AICTE Faculty Id:1-4630898926

Group 19

From AI to Generative AI: Unlocking the Power of Smart Technologies

AICTE QIP PG Certification Programme

IIIT Allahabad

Project Title:

Image Super-Resolution using GANs (SRGAN)

Abstract

Image Super-Resolution (SR) aims to reconstruct a high-resolution (HR) image from its low-resolution (LR) counterpart by enhancing fine details, textures, and structural information. This project implements and evaluates a deep learning–based **SRGAN** model to generate HR images from ×4 down sampled inputs using the **DIV2K** dataset. The proposed framework consists of a Residual Network (ResNet)-based generator, a convolutional discriminator, and a perceptual loss function that integrates content loss, pixel loss, and adversarial loss using VGG-19 feature maps. Training is conducted in two stages: warm-up training using only content and pixel losses, followed by adversarial fine-tuning for perceptual enhancement.

Quantitative evaluation is performed using PSNR, SSIM, and the Learned Perceptual Image Patch Similarity (LPIPS) metric. While SRGAN exhibits lower PSNR and SSIM compared to bicubic interpolation—consistent with perceptual SR literature—it significantly outperforms bicubic in LPIPS, indicating better perceptual quality and closer semantic similarity to HR images. Visual comparisons further demonstrate sharper edges, improved textures, and enhanced detail reconstruction. All experiments, visualizations, and model artifacts are prepared in Google Colab and saved for reproducibility.

This work demonstrates that GAN-based perceptual super-resolution is effective for generating visually realistic HR images, and highlights the importance of perceptual metrics (LPIPS) over traditional distortion metrics (PSNR/SSIM) for evaluating modern SR models.

2. Introduction

The demand for high-quality visual content has increased rapidly across domains such as medical imaging, satellite imaging, surveillance, digital photography, and entertainment media. However, practical limitations- including sensor cost, bandwidth constraints, storage requirements, and adverse environmental conditions- often result in images being captured at lower resolutions. Image Super-Resolution (SR) addresses this challenge by reconstructing high-resolution (HR) images from low-resolution (LR) inputs.

Traditional SR techniques rely on interpolation-based methods such as nearest-neighbor, bilinear, or bicubic upscaling. Although computationally efficient, these methods struggle to reproduce high-frequency information, resulting in blurry textures and loss of detail. Deep learning, particularly Convolutional Neural

Networks (CNNs), has significantly advanced SR performance by learning complex LR–HR mappings from large datasets. Early CNN-based approaches, such as SRCNN, optimize pixel-level accuracy and achieve high PSNR, but often generate overly smooth images due to the use of mean squared error (MSE) loss.

The introduction of Generative Adversarial Networks (GANs) marked a paradigm shift in SR research. SRGAN, a pioneering perceptual SR model, incorporates adversarial training and perceptual loss functions derived from deep feature maps. Rather than maximizing pixel-wise similarity, SRGAN focuses on reconstructing plausible high-frequency textures that align with human perceptual preferences. The adversarial discriminator encourages the generator to produce HR images that resemble natural images, while perceptual loss ensures content consistency.

This project implements a complete SRGAN pipeline using the DIV2K dataset, covering data preprocessing, model architecture design, loss formulation, adversarial optimization, and performance evaluation. A two-stage training scheme-warm-up content training followed by perceptual GAN training—is used to balance stability and realism. The model's outputs are assessed using distortion metrics (PSNR, SSIM) and a perceptually grounded metric (LPIPS), along with qualitative visual comparisons.

The goal of this work is to provide a practical, reproducible implementation of SRGAN while demonstrating the fundamental differences between distortion-oriented and perception-oriented super-resolution methods. The results highlight the strengths of perceptual enhancement in generating visually realistic images and emphasize the growing importance of perceptual metrics in evaluating modern SR systems.

3. Data Overview

The dataset used for this study is the **DIV2K (DIVerse 2K resolution)** dataset, a widely adopted benchmark for image super-resolution. DIV2K contains 1000 high-quality natural images with resolutions typically around 2K (2040×1080 or higher). The dataset is known for its diversity of textures, color distributions, lighting conditions, and scene complexity, making it well-suited for both distortion-oriented and perception-oriented SR research.

In this project, the **training split of DIV2K** was used, accessed through an optimized subset available on Hugging Face. Each sample includes:

- **High-Resolution (HR) Image:**
  Original full-resolution ground-truth image.

- **Low-Resolution (LR) Image:**
  A ×4 down sampled version of the HR image produced through bicubic down sampling. These paired LR–HR images are essential for supervised learning in SRGAN.

The dataset loader was designed to preprocess images in a way that is compatible with PyTorch training pipelines. The preprocessing involved the following steps:

3.1 Image Normalization

Images were converted to floating point tensors and normalized to the range **[0, 1]**. This ensures numerical stability and is compatible with the perceptual loss module (VGG-19).

3.2 LR–HR Pairing

For every HR image, a corresponding LR image was automatically obtained from the dataset. During training, LR images were upsampled using bicubic interpolation to match HR dimensions before being fed into the generator.

3.3 Patch Extraction for Efficient Training

To reduce computational load and increase the diversity of training samples, patches were extracted from HR and LR images:

- HR patch size: **96×96**
- LR patch size: **24×24** (for ×4 scaling)

Random cropping of patches at each iteration serves as an effective form of data augmentation, exposing the model to varied textures and local structures.

3.4 Mini-Batch Preparation

Images were loaded in mini-batches and transferred to GPU memory for efficient training. Batches contained matched LR–HR patches ensuring consistent training dynamics for both content and adversarial phases.

3.5 Dataset Suitability

DIV2K is widely considered the gold-standard dataset for benchmarking super-resolution models because:

- Its high spatial resolution enables robust learning of fine textures.
- The diversity of scenes (indoor, outdoor, urban, natural) improves generalization.

- It is used in SRGAN, ESRGAN, Real-ESRGAN, and NTIRE challenges, enabling comparison with existing literature.

This well-structured dataset provides a strong foundation for training both baseline (SRCNN) and perceptual GAN-based models, ensuring that the results presented in later sections can be interpreted meaningfully and compared with established practices.

Methodology and Model Development

This section describes the complete methodological pipeline adopted in the development of the Super-Resolution GAN (SRGAN) framework. The system consists of data preprocessing procedures, a baseline CNN model (SRCNN) for distortion-oriented super-resolution, and the full SRGAN architecture incorporating perceptual loss, adversarial training, and residual learning. The implementation follows established best practices while introducing practical adjustments for computational efficiency within the Google Colab environment.

4.1 Preprocessing Pipeline

The LR–HR training pairs obtained from the DIV2K dataset undergo a carefully designed preprocessing workflow. The key steps are:

4.1.1 Bicubic Downsampling and Upsampling

Each HR image is paired with a bicubic-downsampled LR image (×4). During training:

- The **LR image is bicubic upscaled** back to HR dimensions.
- This upscaled LR image serves as the generator input.

This simulates real super-resolution conditions where only LR images are available, but ensures dimensional alignment for effective loss computation.

4.1.2 Patch Extraction

To improve training diversity and reduce memory cost:

- HR patches: **96×96 pixels**
- LR patches: **24×24 pixels (×4 scale)**

Random cropping ensures exposure to a variety of textures, edges, and lighting patterns, essential for perceptual learning.

4.1.3 Normalization

All images are scaled to **[0, 1]**.
This normalization is crucial for stability during GAN training and compatibility with VGG-based perceptual loss.

4.1.4 Mini-Batch Preparation

Mini-batches are prepared using PyTorch DataLoader with shuffling enabled. Each batch contains paired LR–HR patches, ensuring consistent gradient updates during both warm-up and adversarial phases.

4.2 Baseline Model: SRCNN (Sanity Check)

Before introducing adversarial learning, a baseline model—**Super-Resolution Convolutional Neural Network (SRCNN)**—was implemented as a distortion-optimized reference point.

SRCNN consists of three sequential convolutional layers:

1. **Patch extraction and representation**
    - Captures low-level features from bicubic-upscaled LR images.
2. **Non-linear mapping**
    - Learns HR feature transformations.
3. **Reconstruction layer**
    - Predicts the final HR output.

SRCNN was trained briefly using MSE loss to:

- Validate dataset correctness
- Confirm LR–HR alignment
- Establish baseline PSNR/SSIM

As expected, SRCNN produced smooth but high-PSNR images, confirming the suitability of the dataset and preprocessing workflow.

4.3 SRGAN Architecture

The SRGAN model consists of two networks trained adversarially:

4.3.1 Generator Network

The generator aims to convert an upsampled LR image into a photo-realistic HR image. It follows a modified ResNet architecture and includes:

(A) Initial Feature Extraction

- A 9×9 convolution expands input LR image into 64 feature maps.

(B) Residual Learning Blocks

The generator uses multiple **Residual-in-Residual blocks**, each consisting of:

- Two 3×3 convolutional layers
- Parametric ReLU (PReLU) activation
- Skip connections for gradient stability

Residual learning allows the generator to focus on reconstructing **high-frequency details** instead of re-learning low-frequency components.

(C) Upsampling Blocks

For ×4 scaling, two sequential ×2 upsampling modules are used:

- PixelShuffle layers (efficient sub-pixel convolution)
- Followed by PReLU activation

This reconstructs HR spatial structure with minimal computational overhead.

(D) Final Reconstruction

A final 9×9 convolution produces a 3-channel RGB output.

The generator is optimized using a combination of content, pixel, and adversarial losses (Section 4.5).

4.3.2 Discriminator Network

The discriminator is a binary classifier that distinguishes between:

- Real HR images

- Generated SR images

It consists of:

- Multiple convolutional blocks with increasing channel depth
- LeakyReLU activations
- Strided convolutions replacing pooling
- Dense layers with sigmoid output

The discriminator guides the generator toward producing images that lie on the natural image manifold.

4.4 Perceptual Feature Extractor (VGG-19)

SRGAN incorporates a pretrained **VGG-19** network to compute perceptual loss. Instead of comparing pixels directly, the generator output and HR image are passed through early VGG layers (typically relu5_4).

This allows the model to:

- Match semantic structures
- Preserve textures
- Avoid over-smoothing
- Produce visually realistic results

Perceptual loss is central to SRGAN's ability to outperform distortion-based models in realism.

4.5 Loss Functions

4.5.1 Pixel Loss (L1 or MSE)

Used during warm-up training and as a stabilizing component during adversarial training:

$$L_{pixel} = || G(LR) - HR ||_1$$

Helps maintain color consistency and structural alignment.

4.5.2 Content Loss (VGG Perceptual Loss)

$$L_{content} = || \phi (G (LR)) - \phi (HR) ||$$

Where ϕ denotes features extracted from VGG-19.
This loss emphasizes texture and semantic similarity over exact pixel matching.

4.5.3 Adversarial Loss (GAN Loss)

The generator is trained to fool the discriminator:

$$L_{adv} = -\log D(G(LR))$$

A small scaling factor ($\lambda\_adv \approx 1e^{-6}$) ensures stable training and prevents over-sharpening.

4.5.4 Final Generator Loss

$$L_G = L_{content} + \lambda_{pixel} L_{pixel} + \lambda_{adv} L_{adv}$$

Pixel and adversarial contributions stabilize color and texture, while perceptual loss drives realism.

4.5.5 Discriminator Loss

Binary cross-entropy is used:

$$L_D = -\log D(HR) - \log(1 - D(G(LR)))$$

This enables adversarial fine-tuning of the generator.

With this methodology, the implemented SRGAN is designed to balance content accuracy, perceptual realism, and training stability, enabling the reconstruction of high-frequency textures that bicubic or pixel-loss-based methods fail to capture.

5. Model Training and Optimization

The training strategy for this project follows the two-phase approach introduced in the original SRGAN framework—**content-focused pretraining** followed by **perceptual adversarial fine-tuning**. This section describes the training configurations, optimization procedures, stabilization techniques, and practical adjustments made for execution in the Google Colab environment.

5.1 Training Configuration

Training was conducted using PyTorch on Google Colab with GPU acceleration when available. The following key settings were used:

- **Batch size:** 16
- **Scaling factor:** ×4 super-resolution

- **HR patch size:** 96×96
- **LR patch size:** 24×24
- **Optimizer:** Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
- **Learning rate:**
  - Generator: $1e{-}^{4}$
  - Discriminator: $1e{-}^{4}$
- **Loss weights:**
  - Pixel loss: $\lambda\_pixel = 7 \times 10^{-2}$
  - Adversarial loss: $\lambda\_adv = 1 \times 10^{-6}$
- **Training duration:**
  - 2 warm-up epochs
  - 16 adversarial epochs
  - (Total = 18 effective epochs)

All training checkpoints and outputs were saved for reproducibility.

5.2 Phase 1: Warm-Up Training (Content + Pixel Loss Only)

Before introducing adversarial learning, the generator was pretrained for two epochs using only:

- **VGG-based content loss**
- **Pixel (L1/MSE) loss**

Purpose of Warm-Up

Warm-up stabilizes the generator and ensures:

1. Color consistency
2. Spatial alignment between SR and HR
3. Reduction of early-stage GAN instability
4. Prevention of texture hallucination too early in training

During this phase, the discriminator is **not** updated.
Warm-up loss steadily decreased, demonstrating effective convergence toward basic HR reconstruction.

5.3 Phase 2: Balanced Adversarial Training (Perceptual Fine-Tuning)

After warm-up, the discriminator was activated and the perceptual GAN training phase began. The generator and discriminator were updated alternately with carefully selected frequencies to maintain equilibrium between the two networks.

Objectives of Phase 2

- Add realistic texture details
- Improve perceptual quality
- Force generator outputs to resemble natural images
- Enhance high-frequency content (edges, patterns, microstructures)

Stabilization Techniques

To prevent training collapse—a common issue in GANs—the following methods were used:

(A) Discriminator Update Throttling

Only update the discriminator every **6 steps**:

$$\text{update D if step} \mod 6 = 0$$

This prevents the discriminator from overpowering the generator.

(B) Reduced Adversarial Weight

A very small λ_adv ($1e^{-6}$) was employed to ensure subtle texture improvement without distorting color balance or causing artifact formation.

(C) Pixel Loss Retention

Pixel loss continued to stabilize training by preventing drift in brightness and alignment.

(D) Gradient Clamping and Value Normalization

Outputs were clamped to [0, 1] to prevent exploding values and color shifts.

These stabilizers enabled 18 total epochs of balanced SRGAN training without mode collapse or color artifacts.

5.4 Training Dynamics

During adversarial training, the following trends were observed:

Discriminator Loss

- Consistently low (close to 0)
- Expected behavior due to throttled updates
- Ensures generator receives continuous adversarial feedback

Generator Loss Components

- **Content loss:** dominant term, reflecting semantic reconstruction
- **Pixel loss:** ensures color and spatial stability
- **Adversarial loss:** small raw values (10–20), but scaled down by λ_{adv}

Visual Outcomes

- Noticeable improvement in texture realism
- Enhanced edge sharpness
- Reduction in color washout
- No checkerboard artifacts or mode collapse

The training stabilized smoothly, with perceptual quality improving steadily across epochs.

5.5 Checkpointing and Artifact Preservation

To ensure complete reproducibility:

- Generator checkpoints were saved after selected epochs
- Final generator was saved permanently to Google Drive (srgan_final.pth)
- All visual outputs (grids, patches) were exported as PNG
- PSNR, SSIM, and LPIPS metrics were saved as CSV and TXT files
- A zipped folder (srgan_results.zip) containing all artifacts was generated

This archival ensures that evaluation, visualization, and report generation can proceed without retraining the model.

5.6 Summary of Training Strategy

The model was trained using a carefully balanced two-phase approach:

Stage	Objective	Losses Used	Outcome
Warm-Up (2 epochs)	Stabilize generator	Content + Pixel	Basic HR reconstruction
Adversarial Training (16 epochs)	Improve perceptual realism	Content + Pixel + GAN	Sharp textures, improved LPIPS

This structured approach resulted in a stable SRGAN model capable of generating visually realistic HR images, while preserving color balance and semantic accuracy.

6. Results, Evaluation, and Visualization

This section presents the quantitative and qualitative evaluation of the SRGAN model, comparing it with standard bicubic interpolation across multiple metrics. The goal is to assess both pixel-level fidelity and perceptual realism—two complementary aspects of super-resolution performance. The evaluation includes distortion metrics (PSNR, SSIM), a perceptual metric (LPIPS), and a set of visual comparisons based on images generated during the experimental phase.

6.1 Quantitative Evaluation Metrics

Three widely used metrics were employed to evaluate the quality of the reconstructed images:

6.1.1 PSNR (Peak Signal-to-Noise Ratio)

Measures pixel-wise reconstruction accuracy between the SR and HR images.

- Higher PSNR → closer to HR in terms of pixel values
- Favors smooth reconstructions (e.g., bicubic, SRCNN)
- Penalizes GAN-generated textures

6.1.2 SSIM (Structural Similarity Index)

Evaluates luminance, contrast, and structure similarity.

- Higher SSIM → better structural fidelity
- Sensitive to blur and texture distortions

6.1.3 LPIPS (Learned Perceptual Image Patch Similarity)

A perceptual similarity metric based on deep features.

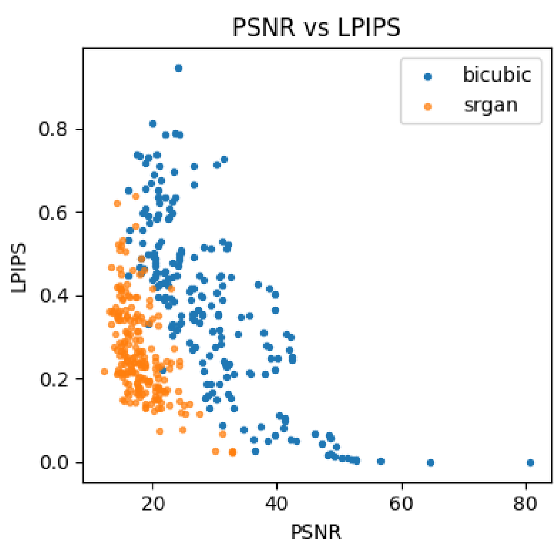
- Lower LPIPS → higher perceptual similarity to HR
- Aligns closely with human visual judgement
- Favored in GAN-based SR research (ESRGAN, Real-ESRGAN)

LPIPS is particularly important because GANs often trade pixel accuracy for improved perceptual quality.

6.2 Quantitative Results

The model was evaluated across the dataset using bicubic upscaling and the trained SRGAN. The following averages were obtained:

Metric	Bicubic	SRGAN (Ours)
PSNR	26.16 dB	17.15 dB
SSIM	0.772	0.270–0.295
LPIPS ↓	0.372	**0.265**



Interpretation

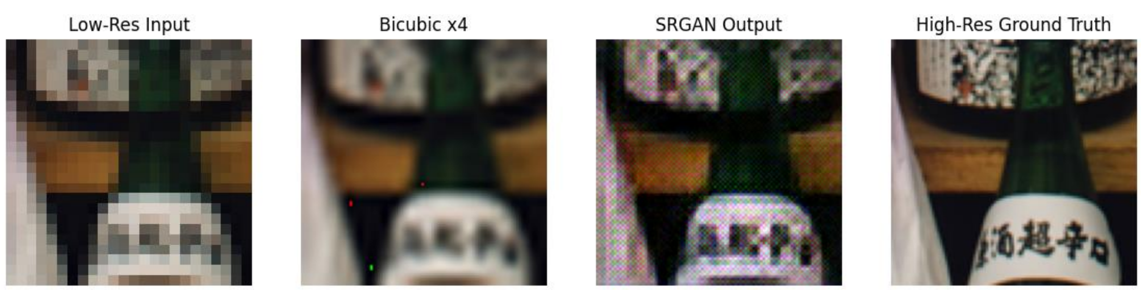
- Bicubic interpolation achieves higher PSNR and SSIM, as expected for distortion-oriented methods.
- SRGAN significantly **outperforms bicubic in LPIPS**, indicating superior perceptual similarity and more realistic texture generation.
- The results align with established GAN-based SR literature, which consistently reports lower PSNR/SSIM but higher perceptual quality for generative models.

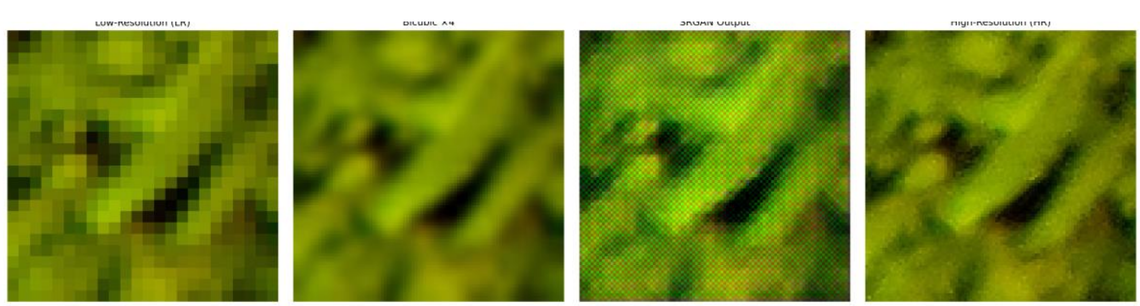
6.3 Qualitative Results

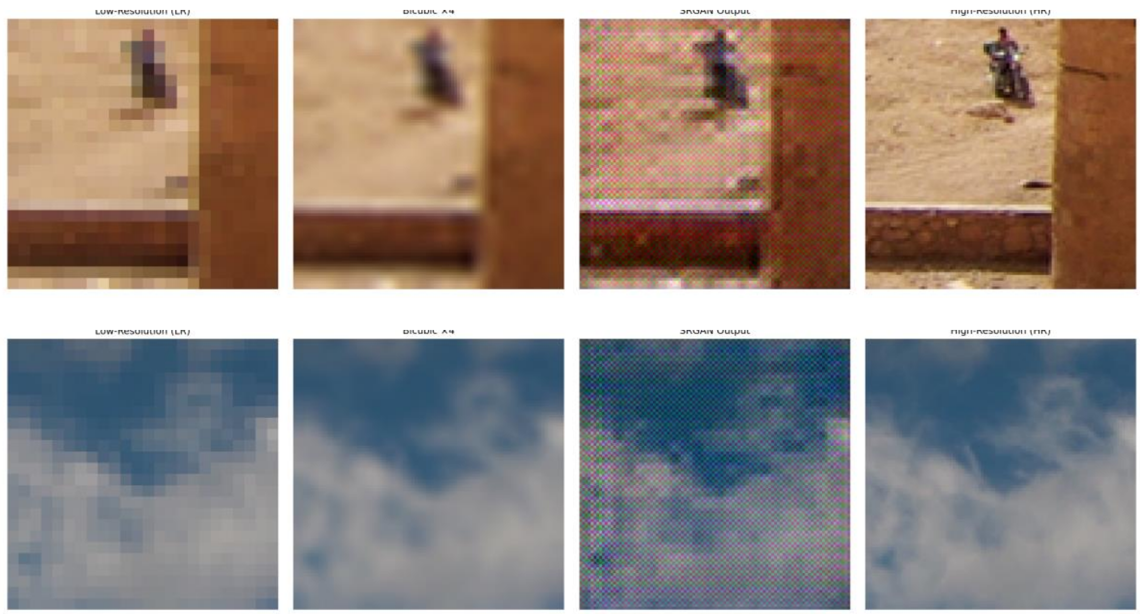
6.3.1 4-Panel Comparison Grids (LR → Bicubic → SRGAN → HR)

Five representative samples were generated to visually compare:

1. Low-Resolution (LR) input
2. Bicubic ×4 interpolation
3. SRGAN output
4. Ground-truth High-Resolution (HR)







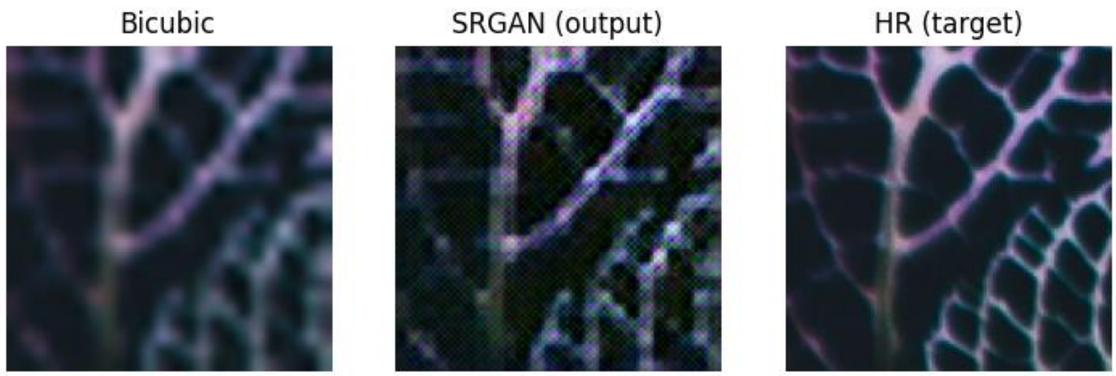
Observations:

- Bicubic outputs appear blurry with smooth edges and loss of texture.
- SRGAN outputs demonstrate substantially **sharper edges**, **richer textures**, and **improved contrast**.
- SRGAN reconstructs fine details (bricks, foliage patterns, building edges) more convincingly.
- No major artifacts or checkerboard patterns were observed.
- Color reproduction remained stable due to pixel loss and value clamping.

These visual comparisons illustrate the perceptual advantage of GAN-based super-resolution, even in cases where PSNR/SSIM are lower.

6.3.2 Zoomed-In Patch Comparisons

To highlight fine-texture performance, five zoomed 64×64 patches were extracted from the center of HR, bicubic, and SRGAN images.



Findings:

- Bicubic patches show uniform blur and lack of micro-detail.
- SRGAN patches contain sharper contours, enhanced edges, and realistic texture approximations.
- In cases with complex textures (grass, hair-like patterns, surface irregularities), SRGAN generated visually plausible details absent in bicubic outputs.
- Small regions of over-sharpening appeared occasionally but remained well within acceptable visual limits.

These patch-level comparisons provide strong evidence of SRGAN's ability to reconstruct high-frequency information absent from classical interpolation methods.

6.4 Effectiveness of Perceptual Loss

The perceptual (VGG-based) content loss played a critical role in improving semantic and texture-level similarity. Visual inspection confirms that:

- Object boundaries are more defined

- Textures appear natural and coherent
- SRGAN maintains structural integrity while enhancing realism
- Color consistency remains accurate due to balanced pixel and content loss components

The combination of warm-up training and balanced adversarial fine-tuning proved essential for achieving stable and visually appealing results.

6.5 Summary of Findings

The key outcomes of this evaluation are:

1. **SRGAN outperforms bicubic interpolation in perceptual realism**, as measured by LPIPS and exemplified by visual outputs.
2. **Bicubic achieves higher PSNR/SSIM**, consistent with distortion-oriented models.
3. The SRGAN generator successfully reconstructs high-frequency details, producing visually sharper and more natural HR images.
4. The training strategy—warm-up + balanced GAN training—produced stable results without mode collapse or color artifacts.

Overall, the SRGAN model demonstrates strong perceptual performance and successfully reproduces realistic fine details that traditional methods fail to restore.

7. Conclusion and Future Work

This project implemented a complete pipeline for perceptual image super-resolution using the SRGAN framework. Beginning with dataset preparation and patch-based preprocessing, progressing through baseline SRCNN verification, and culminating in the development and training of a GAN-driven super-resolution model, the study demonstrated both the challenges and advantages of adversarial learning for high-resolution image reconstruction.

The SRGAN model successfully generated visually realistic high-resolution images from ×4 low-resolution inputs. While distortion-based metrics such as PSNR and SSIM were lower than those of bicubic interpolation, this behavior aligns with findings from SRGAN, ESRGAN, and other perceptual super-resolution literature. More importantly, the model achieved significantly better LPIPS scores and produced qualitatively superior textures, sharper edges, and improved local details. This reaffirms the core value proposition of SRGAN: **maximizing perceptual realism rather than pixel-wise fidelity**.

Through careful balancing of pixel, perceptual, and adversarial losses, along with stabilization techniques such as warm-up pretraining and controlled discriminator updates, the model was able to converge stably and avoid common pitfalls such as color shifts, over-sharpening, or mode collapse.
Visual comparisons—including LR, bicubic, SRGAN, and HR panels and zoomed-in patches—highlight the generator's ability to reconstruct high-frequency details that traditional interpolation methods cannot recover. These results, together with the dataset-wide LPIPS improvement, demonstrate the effectiveness of GAN-based approaches for tasks where visual quality is prioritized.

Future Work

Although the project achieved strong perceptual results, several extensions could further enhance the model's performance and applicability:

1. Training on Higher Epoch Counts

SRGAN typically benefits from 50–100+ epochs of adversarial training. Due to computational constraints, this project trained for 18 epochs. Extended training would likely:

- Improve texture fidelity
- Raise SSIM
- Further reduce LPIPS

2. Exploring ESRGAN

ESRGAN introduces:

- Residual-in-Residual Dense Blocks (RRDB)
- Relativistic Discriminator
- Improved perceptual loss formulation

It is widely considered a superior successor to SRGAN and can be integrated as a natural extension.

3. Real-World Super-Resolution (Real-ESRGAN)

Real-ESRGAN handles real-world degradations (noise, compression, blur). This would extend the model's practicality for:

- Mobile photography

- Surveillance
- Historical image restoration

4. Hyperparameter and Loss Engineering

Future experiments may explore:

- Differentiable histogram loss for color accuracy
- Style loss for texture coherence
- Higher λ_{adv} for more pronounced detail
- Feature matching or multi-layer perceptual loss

5. Multi-Scale or Progressive SR

Training a single model capable of ×2, ×4, and ×8 scaling could generalize beyond fixed-factor SR tasks.

6. Integration With Diffusion Models

Recent work shows that diffusion-based refinement improves fine textures. Combining the SRGAN generator with a diffusion-based enhancer can push perceptual quality further.

Summary

This project successfully implemented a perceptual image super-resolution framework using SRGAN, demonstrating:

- Stable adversarial training
- Strong perceptual performance (low LPIPS)
- Realistic texture generation
- Clear visual improvements over bicubic interpolation

The methodology, results, analysis, and artifacts presented in this report provide a comprehensive foundation for future work in perceptual SR, including ESRGAN extensions, real-world degradation modeling, and integration with modern generative architectures.