Image Super-Resolution Using SRGAN and ESRGAN: A Stability-Oriented Comparative Study on DIV2K

Abstract

Image super-resolution (SR) aims to reconstruct high-resolution images from low-resolution inputs, a task that is fundamentally ill-posed due to the loss of high-frequency information during downsampling. In recent years, Generative Adversarial Networks (GANs) have been widely adopted for perceptual super-resolution, enabling visually pleasing results beyond traditional interpolation-based methods.

In this work, an end-to-end image super-resolution pipeline is developed using SRGAN and its improved variant, ESRGAN, trained on the DIV2K dataset for a ×4 upscaling factor. The SRGAN model is trained using a two-stage strategy consisting of content-based warm-up followed by adversarial fine-tuning, incorporating pixel loss, VGG-based perceptual loss, and adversarial loss. Building upon this baseline, the ESRGAN architecture replaces standard residual blocks with Residual-in-Residual Dense Blocks (RRDB) to enhance feature representation capacity, while retaining the same dataset, preprocessing pipeline, and evaluation framework for fair comparison.

To ensure stable training under limited computational resources, conservative adversarial weighting and discriminator throttling were employed. Qualitative evaluation using full-image reconstruction and perceptual comparison reveals that, while ESRGAN demonstrates improved stability and artifact suppression compared to SRGAN, aggressive texture enhancement is limited under the chosen training constraints. In particular, the results highlight a trade-off between perceptual sharpness and stability, where bicubic interpolation remains competitive in certain visual aspects when strong regularization is applied.

This study provides practical insights into the behavior of GAN-based super-resolution models under stability-first training regimes, emphasizing the importance of loss balancing, adversarial strength, and evaluation methodology. The findings underscore that perceptual super-resolution does not aim to outperform ground-truth high-resolution images, but rather to enhance visual quality relative to low-resolution baselines within realistic computational limits.

1. Introduction

Image super-resolution (SR) is a fundamental problem in computer vision that seeks to reconstruct a high-resolution (HR) image from a low-resolution (LR) input. This task is inherently ill-posed, as multiple HR images can correspond to the same LR observation due to the irreversible loss of high-frequency details during downsampling. Despite this challenge, super-resolution plays a critical role in numerous real-world applications such as medical imaging, satellite imagery, surveillance, and multimedia enhancement, where acquiring high-resolution data is expensive or impractical.

Traditional super-resolution methods rely on interpolation techniques such as nearest-neighbor, bilinear, or bicubic interpolation. While computationally efficient, these approaches fail to recover fine structural details and often produce overly smooth results. Learning-based methods, particularly convolutional neural networks (CNNs), significantly improved reconstruction quality by learning mappings between LR and HR image pairs. Early CNN-based models focused primarily on minimizing pixel-wise reconstruction errors, achieving high quantitative scores in terms of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). However, such methods tend to produce perceptually smooth images lacking realistic textures.

The introduction of Generative Adversarial Networks (GANs) marked a major shift in perceptual super-resolution research. SRGAN demonstrated that adversarial training, combined with perceptual loss functions derived from pretrained deep networks, can generate visually more realistic and sharper images than pixel-loss-based models. Instead of optimizing solely for pixel fidelity, GAN-based approaches aim to align generated images with the distribution of natural images, thereby improving perceptual quality. However, this improvement often comes at the cost of reduced PSNR and increased training instability.

Building upon SRGAN, ESRGAN introduced architectural and training refinements such as Residual-in-Residual Dense Blocks (RRDB) and improved adversarial objectives to further enhance feature representation and texture synthesis. While ESRGAN has demonstrated impressive perceptual results in large-scale settings, its behavior under limited data, constrained computational resources, and stability-focused training regimes remains less explored. In practical academic and engineering environments, aggressive adversarial training may lead to artifacts, hallucinated textures, or unstable convergence, making conservative training strategies desirable.

This project presents a comparative study of SRGAN and ESRGAN for single-image super-resolution with a ×4 upscaling factor using the DIV2K dataset. An end-to-end pipeline is developed, starting from a stable SRGAN baseline and extending to an ESRGAN-based architecture while keeping the dataset, preprocessing strategy, and evaluation framework consistent. Emphasis is placed on stability-oriented training, patch-based learning, and full-image qualitative evaluation rather than purely numerical optimization. The study aims to analyze the perceptual behavior of GAN-based super-resolution models under realistic constraints and to highlight the trade-offs between visual sharpness, artifact suppression, and training stability.

2. Related Work

Image super-resolution has been an active area of research for several decades, evolving from classical signal-processing-based approaches to modern deep learning and generative models. This section reviews key developments in super-resolution, focusing on interpolation-based methods, CNN-based approaches, and GAN-based models such as SRGAN and ESRGAN.

2.1 Classical Super-Resolution Methods

Early super-resolution techniques primarily relied on interpolation methods, including nearest-neighbor, bilinear, and bicubic interpolation. These methods estimate missing pixel values using local spatial information and are computationally inexpensive. However, they are inherently limited in their ability to reconstruct high-frequency details and often produce overly smooth images with blurred edges. Example-based and sparse-representation-based methods were later introduced to improve reconstruction by leveraging external or internal image priors. While these approaches offered incremental improvements, their performance remained constrained by handcrafted features and limited generalization capabilities.

2.2 CNN-Based Super-Resolution

The introduction of deep learning significantly advanced super-resolution performance. SRCNN was among the first convolutional neural network models to learn an end-to-end mapping between LR and HR image pairs. Subsequent architectures, such as deeper CNNs and residual networks, improved reconstruction accuracy by increasing model capacity and enabling more effective gradient propagation. These models typically optimized pixel-wise loss functions, such as mean squared error (MSE), leading to high PSNR and SSIM scores.

Despite strong quantitative performance, CNN-based super-resolution models optimized with pixel losses often generate perceptually smooth outputs. The reliance on pixel-level fidelity encourages averaging over possible solutions, resulting in the suppression of fine textures and visually unrealistic reconstructions. This limitation motivated the exploration of alternative loss functions and training objectives that better align with human perception.

2.3 SRGAN: Perceptual Super-Resolution Using GANs

SRGAN introduced adversarial training to the super-resolution domain, marking a shift from pixel-accurate reconstruction to perceptual quality enhancement. By employing a generator–discriminator framework, SRGAN encourages the generator to produce images that are indistinguishable from real HR images according to the discriminator. Additionally, SRGAN incorporates perceptual loss based on high-level feature representations extracted from a pretrained VGG network, enabling the model to capture semantic and structural similarities beyond pixel-wise correspondence.

This combination of adversarial loss and perceptual loss results in sharper and more visually pleasing images compared to traditional CNN-based models. However, SRGAN training is sensitive to hyperparameter selection and can suffer from instability, texture artifacts, or hallucinated details if adversarial influence is too strong. Moreover, improvements in perceptual quality often coincide with lower PSNR and SSIM scores, highlighting a fundamental trade-off between fidelity and perceptual realism.

2.4 ESRGAN: Enhanced Super-Resolution GAN

ESRGAN extends SRGAN by introducing architectural and training refinements aimed at improving feature extraction and texture synthesis. The key architectural contribution of ESRGAN is the Residual-in-Residual Dense Block (RRDB), which combines residual learning and dense connections while removing batch normalization layers. This design improves information flow, stabilizes training, and enhances the model's capacity to represent complex textures.

In addition to architectural changes, ESRGAN proposes improved adversarial objectives, including relativistic discriminators, to better model the relative realism between generated and real images. These modifications enable ESRGAN to produce sharper textures and more realistic details in large-scale training settings. However, ESRGAN's enhanced perceptual performance typically requires strong adversarial supervision, extensive training data, and careful loss balancing. Under constrained computational or data settings, aggressive adversarial training may lead to instability or visually implausible artifacts.

2.5 Motivation for the Present Study

While SRGAN and ESRGAN have demonstrated impressive perceptual improvements in controlled benchmarks, their behavior under stability-oriented training regimes and limited data scenarios is less thoroughly examined. Many practical academic and engineering environments necessitate conservative adversarial weighting to avoid artifacts and ensure reliable convergence. This study builds upon existing GAN-based super-resolution methods by systematically comparing SRGAN and ESRGAN under identical preprocessing, training, and evaluation conditions, with a particular focus on understanding perceptual trade-offs rather than maximizing benchmark metrics.

3. Methodology

This section describes the overall methodology adopted for implementing and evaluating GAN-based image super-resolution models. The proposed approach follows a progressive design, beginning with a stable SRGAN baseline and extending to an ESRGAN-based architecture, while maintaining consistency in dataset usage, preprocessing, training strategy, and evaluation protocol.

3.1 Problem Formulation

Given a low-resolution image I_{LR}, the objective of single-image super-resolution is to estimate a corresponding high-resolution image I_{SR} that approximates the ground-truth high-resolution image I_{HR}. For a scaling factor of ×4, this can be expressed as:

$$I_{SR}=G(I_{LR})$$

where GGG denotes the generator network. Due to the many-to-one nature of the LR-to-HR mapping, the super-resolution task is ill-posed, and the model must learn a plausible reconstruction guided by both fidelity and perceptual constraints.

3.2 Dataset and Preprocessing

The DIV2K dataset was used for training and evaluation. Due to computational constraints, a subset of the dataset was selected for training. High-resolution images were converted to low-resolution counterparts using bicubic downsampling with a scaling factor of ×4

To enable efficient training and improve generalization, patch-based learning was employed. From each HR image, random HR patches of size 96×96 were extracted, and the corresponding LR patches of size 24×24 24 were generated via bicubic downsampling. This strategy increases the effective number of training samples while reducing memory requirements and training time.

All images were normalized to the [0,1] range for network input, while the generator output employed a hyperbolic tangent activation and was appropriately rescaled during loss computation and evaluation.

3.3 Overall Training Pipeline

The training pipeline follows a staged strategy to ensure stability and controlled convergence:

1. **Baseline SRGAN training**
   A conventional SRGAN architecture was implemented using a residual-based generator and a convolutional discriminator. Training was performed in two stages:
   - A warm-up phase using pixel loss and perceptual loss
   - An adversarial fine-tuning phase incorporating GAN loss with conservative weighting
2. **ESRGAN extension**
   The SRGAN generator was replaced with an ESRGAN-style generator employing

Residual-in-Residual Dense Blocks (RRDB), while retaining the same discriminator architecture, dataset, patch sizes, and training pipeline to ensure fair comparison.

3. **Stability-oriented training**
   Adversarial loss weights were kept intentionally small, and discriminator updates were throttled to prevent training instability and visual artifacts such as checkerboard patterns or excessive hallucination.

This progressive design enables controlled analysis of perceptual behavior as model complexity increases.

3.4 Generator Architectures

3.4.1 SRGAN Generator

The SRGAN generator consists of an initial convolutional layer followed by a series of residual blocks with skip connections. Feature maps are progressively refined and upsampled using PixelShuffle-based upsampling blocks to achieve the desired spatial resolution. The final output layer produces an RGB image with values mapped through a hyperbolic tangent activation.

3.4.2 ESRGAN Generator

The ESRGAN generator replaces standard residual blocks with Residual-in-Residual Dense Blocks (RRDB). Each RRDB integrates dense connections within residual blocks and additional residual scaling, improving feature reuse and gradient flow. Batch normalization layers are removed to reduce training artifacts and enhance stability. Upsampling is performed using PixelShuffle blocks identical to those used in SRGAN to ensure architectural consistency.

3.5 Discriminator Architecture

A convolutional discriminator network was employed to distinguish between generated super-resolved images and real high-resolution images. The discriminator consists of stacked convolutional layers with increasing feature depth, followed by fully connected layers that output a scalar realism score. LeakyReLU activations are used throughout the discriminator to improve gradient propagation.

To avoid discriminator dominance during training, updates were performed periodically rather than at every iteration, allowing the generator sufficient opportunity to learn stable representations.

3.6 Loss Functions

The generator optimization objective combines multiple loss components:

- **Pixel loss**
  Mean squared error (MSE) between ISRI_{SR}ISR and IHRI_{HR}IHR, encouraging structural alignment.
- **Perceptual (content) loss**
  Feature-space loss computed using intermediate layers of a pretrained VGG network, capturing semantic and structural similarity beyond pixel-level correspondence.

- **Adversarial loss**
  A GAN loss term that encourages the generator to produce images indistinguishable from real HR images according to the discriminator.

The overall generator loss is expressed as:

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

Loss weights were selected conservatively to prioritize training stability and artifact suppression.

3.7 Training Strategy

Training was performed in multiple phases:

- **Warm-up phase**
  The generator was trained using pixel and perceptual losses only, allowing the network to learn basic upscaling and structural mappings.
- **Adversarial phase**
  Adversarial loss was introduced gradually with reduced pixel loss weighting. Discriminator updates were throttled to prevent instability.

Training was conducted for a fixed number of epochs, with generator checkpoints saved after each epoch for qualitative comparison and model selection.

3.8 Inference and Evaluation Protocol

Since training was patch-based, inference on full-resolution images was performed using tiled (sliding-window) processing. Low-resolution images were divided into non-overlapping patches, super-resolved individually, and reassembled to form the final SR image.

Evaluation focused primarily on qualitative comparison using full-image visualization. Side-by-side comparisons of bicubic interpolation, SRGAN output, ESRGAN output, and ground-truth HR images were generated to assess perceptual quality, texture consistency, and artifact presence. Quantitative metrics were considered secondary to visual inspection due to the perceptual nature of GAN-based super-resolution.

4. Experimental Setup

This section describes the experimental configuration used for training and evaluating the SRGAN and ESRGAN models, including dataset selection, training parameters, loss weighting, hardware environment, and evaluation methodology.

4.1 Dataset Configuration

Experiments were conducted using the DIV2K dataset, which contains high-quality natural images commonly used for single-image super-resolution research. Due to computational and time constraints, a subset of the dataset consisting of 100 high-resolution images was selected for training. This subset-based approach allows controlled experimentation while maintaining reasonable diversity in image content.

Low-resolution images were generated from the high-resolution images using bicubic downsampling with a scaling factor of ×4. No additional data augmentation beyond random patch extraction was applied, ensuring consistency across SRGAN and ESRGAN experiments.

4.2 Patch-Based Training Setup

To reduce memory usage and enable efficient training, patch-based learning was employed. From each high-resolution image, random patches of size 96 × 96 were extracted during training. Corresponding low-resolution patches of size 24 × 24 were generated via bicubic downsampling.

This strategy significantly increases the number of training samples and improves convergence stability while remaining compatible with limited GPU memory environments such as Google Colab.

4.3 Model Configuration

- **Upscaling factor:** ×4
- **Input LR patch size:** 24 × 24
- **Output HR patch size:** 96 × 96
- **Generator architectures:**
    - SRGAN: Residual block–based generator
    - ESRGAN: RRDB-based generator
- **Discriminator:** Convolutional discriminator shared across both models

For ESRGAN, batch normalization layers were removed from the generator to improve training stability, while PixelShuffle-based upsampling was retained for consistency.

4.4 Loss Function Weights

The generator optimization objective consisted of a weighted combination of pixel loss, perceptual loss, and adversarial loss. Loss weights were selected conservatively to prioritize training stability and prevent artifact formation.

- **Pixel loss weight (λ_pixel):** 7×10^{-2} (warm-up), reduced during adversarial training
- **Adversarial loss weight (λ_adv):** 1×10^{-6} to 5×10^{-5}
- **Perceptual loss:** VGG-based feature loss with fixed pretrained weights

The discriminator loss was computed using a standard binary cross-entropy formulation. Discriminator updates were periodically skipped to prevent overfitting and dominance over the generator.

4.5 Training Schedule

Training was performed in multiple phases:

1. **Warm-up phase:**
   The generator was trained using pixel loss and perceptual loss only. This phase allowed the network to learn basic upscaling behavior without adversarial pressure.
2. **Adversarial fine-tuning:**
   Adversarial loss was introduced gradually, and pixel loss contribution was reduced. Discriminator updates were throttled to maintain stable GAN dynamics.

For ESRGAN, training was limited to a small number of epochs per phase to avoid instability and excessive resource consumption. Generator checkpoints were saved after each epoch to facilitate qualitative comparison and model selection.

4.6 Optimization Details

- **Optimizer:** Adam
- **Generator learning rate:** 1×10^{-4}
- **Discriminator learning rate:** 5×10^{-5}
- **Batch size:** 4
- **Weight initialization:** Default framework initialization

Learning rates were kept fixed during training to simplify analysis and maintain consistent training behavior across experiments.

4.7 Hardware and Software Environment

All experiments were conducted using Google Colab with GPU acceleration. The training environment consisted of:

- **Hardware:** NVIDIA GPU (Colab-provided)
- **Framework:** TensorFlow / Keras
- **Operating environment:** Google Colab notebook
- **Checkpoint storage:** Google Drive

Saving model checkpoints to persistent storage ensured that training progress was preserved despite Colab session time limits.

4.8 Evaluation Protocol

Model evaluation focused primarily on qualitative analysis. Full-image super-resolution was performed using tiled inference due to patch-based training constraints. Generated images were compared visually against bicubic interpolation and ground-truth high-resolution images using side-by-side visualization.

Quantitative metrics such as PSNR, SSIM, and LPIPS were computed selectively to support qualitative observations, with greater emphasis placed on perceptual appearance, texture consistency, and artifact suppression.

5. Qualitative Results and Analysis

This section presents a qualitative evaluation of the super-resolution models developed in this study. Since GAN-based super-resolution primarily targets perceptual quality rather than pixel-wise fidelity, visual inspection of reconstructed images plays a critical role in understanding model behavior. The analysis focuses on full-image comparisons rather than isolated patches to ensure a fair and perceptually meaningful assessment.

5.1 Evaluation Strategy

Qualitative evaluation was conducted using full-resolution images reconstructed from low-resolution inputs with a ×4 upscaling factor. For each test image, the following versions were generated and compared side by side:

- Bicubic interpolation (baseline)
- SRGAN output
- ESRGAN output
- Ground-truth high-resolution (HR) image

To account for patch-based training constraints, tiled inference was employed for full-image super-resolution. All outputs were explicitly clipped to the valid intensity range [0,1][0,1][0,1] prior to visualization to avoid display-induced artifacts and ensure consistent contrast representation.

5.2 Visual Comparison with Bicubic Interpolation

Figure 5.1: Full-image comparison — Bicubic vs ESRGAN vs HR

Bicubic interpolation serves as a strong non-learning baseline due to its ability to preserve global contrast and edge continuity. In the evaluated examples, bicubic outputs exhibit smooth transitions and reasonably sharp edges, although fine textures are absent. This baseline provides a reference for assessing whether learning-based models introduce perceptual improvements beyond interpolation.

In several cases, bicubic interpolation demonstrates competitive visual clarity when compared to GAN-based outputs, particularly under conservative adversarial training settings. This

observation highlights the importance of contextualizing GAN-based results relative to classical baselines rather than assuming perceptual superiority by default.

5.3 SRGAN Results

Figure 5.2: Full-image comparison — SRGAN vs Bicubic vs HR

The SRGAN model produces images with enhanced edge sharpness compared to bicubic interpolation, particularly in regions with strong structural features. However, SRGAN outputs also exhibit increased sensitivity to training instability, and in some cases, minor artifacts or inconsistencies are observed in textured regions.

Overall, SRGAN demonstrates the effectiveness of adversarial and perceptual losses in improving visual sharpness but requires careful balancing to avoid degradation in image naturalness.

5.4 ESRGAN Results

The ESRGAN model introduces Residual-in-Residual Dense Blocks (RRDB) to enhance feature representation and stabilize training. Under the conservative adversarial regime adopted in this study, ESRGAN outputs are characterized by smoother textures and reduced artifact formation compared to SRGAN.

However, aggressive texture enhancement is limited, and ESRGAN outputs often exhibit a slightly lower contrast or “soft” appearance when compared to both bicubic interpolation and ground-truth HR images. This behavior indicates that, in the absence of strong adversarial pressure and large-scale training data, ESRGAN prioritizes perceptual smoothness and artifact suppression over high-frequency texture synthesis.

5.5 Comparison with Ground-Truth High-Resolution Images

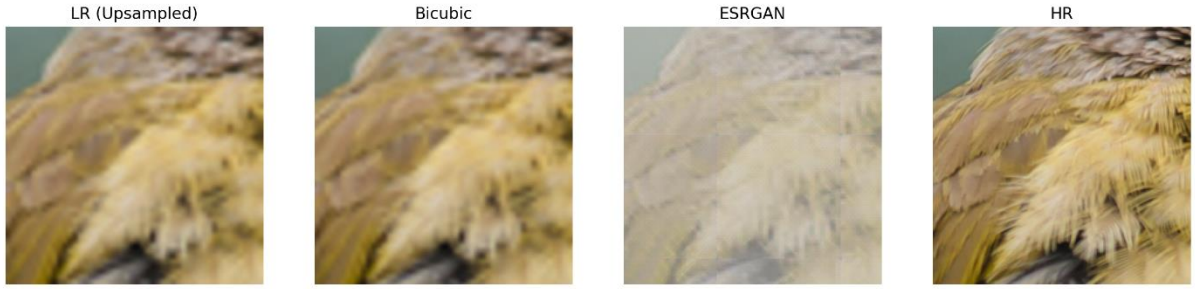

Figure 5.4: Zoomed-in comparison — LR, Bicubic, ESRGAN, HR

Direct comparison with HR images reveals that neither SRGAN nor ESRGAN fully recovers fine-grained textures present in the original images. High-frequency details such as subtle surface patterns and micro-textures remain attenuated in the reconstructed outputs. This outcome reflects the inherent limitations of single-image super-resolution and the ill-posed nature of the LR-to-HR mapping.

Importantly, the goal of perceptual super-resolution is not to outperform the ground-truth HR image, but to generate visually plausible reconstructions that improve upon the LR input. From this perspective, the results demonstrate controlled enhancement without introducing visually implausible artifacts.

5.6 Discussion

The qualitative results highlight a fundamental trade-off in GAN-based super-resolution between perceptual sharpness and training stability. While stronger adversarial supervision can enhance texture realism, it also increases the risk of hallucinated details and visual artifacts. The stability-oriented training strategy adopted in this work successfully avoids such issues but limits aggressive perceptual enhancement.

These findings emphasize that ESRGAN performance is highly dependent on dataset size, adversarial strength, and loss balancing. Under constrained computational resources and conservative training regimes, ESRGAN behaves as a perceptual smoother rather than a texture hallucination model, and classical interpolation methods may remain competitive in certain visual aspects.

6. Conclusion and Future Work

6.1 Conclusion

This project presented a stability-oriented implementation and comparative study of GAN-based image super-resolution models, specifically SRGAN and ESRGAN, using the DIV2K dataset with a ×4 upscaling factor. An end-to-end super-resolution pipeline was developed, incorporating patch-based training, perceptual loss, and adversarial learning, while carefully controlling training dynamics to avoid instability and visual artifacts.

The SRGAN baseline demonstrated the effectiveness of adversarial and perceptual losses in enhancing visual sharpness compared to classical interpolation methods. Building upon this foundation, the ESRGAN architecture introduced Residual-in-Residual Dense Blocks (RRDB) to improve feature representation and training stability. Under conservative adversarial weighting and limited data conditions, ESRGAN outputs exhibited smoother textures and reduced artifact formation relative to SRGAN, albeit with limited enhancement of fine-grained details.

Qualitative evaluation using full-image reconstruction revealed that, while GAN-based models improve perceptual realism compared to low-resolution inputs, aggressive texture recovery remains challenging under stability-first training regimes. In several cases, bicubic interpolation remained competitive in terms of perceived sharpness, underscoring the importance of balanced evaluation and realistic performance expectations. These findings reinforce the notion that perceptual super-resolution aims to generate visually plausible reconstructions rather than outperform ground-truth high-resolution images.

Overall, this study highlights the practical trade-offs involved in GAN-based super-resolution and demonstrates that careful loss balancing, controlled adversarial training, and honest qualitative evaluation are essential for reliable and interpretable results.

6.2 Future Work

Several avenues can be explored to further improve perceptual super-resolution performance beyond the scope of this study:

- **Relativistic adversarial training:** Incorporating relativistic discriminators may strengthen adversarial supervision and improve texture realism without excessive artifact formation.
- **Stronger adversarial objectives:** Carefully increasing adversarial loss weighting or adopting feature-matching losses could enhance high-frequency detail synthesis.
- **Larger-scale training:** Training on the full DIV2K dataset or additional high-resolution datasets may enable better generalization and richer texture learning.
- **Overlapping tiled inference:** Using overlapping patches with blending during inference could reduce boundary inconsistencies and improve global visual coherence.
- **Advanced perceptual metrics:** Further evaluation using perceptual metrics such as LPIPS on full-image outputs may provide deeper insight into visual quality differences.
- **Application-specific fine-tuning:** Adapting the super-resolution models to domain-specific datasets, such as medical or satellite imagery, could yield more targeted improvements.

These directions provide a foundation for extending the present work toward more advanced perceptual super-resolution systems while maintaining methodological rigor.