SRGAN Project

Authors: Abdallah Jamal Al-Harrem, Hossam Shehadeh

Executive Summary

This report presents a comprehensive analysis of a dual-implementation SRGAN (Super-Resolution Generative Adversarial Network) system designed for 4× single image super-resolution. Two distinct models were developed and evaluated: one specialized for facial imagery using the CelebA dataset, and another for general natural images using the DIV2K dataset. The implementation employs a sophisticated two-stage training methodology, combining pixel-wise pre-training with adversarial fine-tuning to achieve superior perceptual quality while maintaining computational efficiency.

Key Achievements:

- CelebA Model: Achieved 42.38 dB PSNR and 0.0368 LPIPS on facial imagery
- DIV2K Model: Achieved 25.15 dB PSNR and 0.1577 LPIPS on natural images
- Robust training framework with comprehensive evaluation metrics
- Domain-specific optimizations for facial super-resolution

1. Introduction and Problem Statement

Single Image Super-Resolution (SISR) represents a fundamental challenge in computer vision, aiming to reconstruct high-resolution images from their low-resolution counterparts. Traditional interpolation methods often produce blurry results lacking high-frequency details. This project addresses the need for perceptually realistic super-resolution that preserves fine details and textures while maintaining computational efficiency.

The primary objective is to develop a GAN-based super-resolution system capable of producing visually compelling 4× upscaled images across different domains. The emphasis is on perceptual quality rather than pure pixel-wise reconstruction accuracy, acknowledging that human visual perception often differs from mathematical similarity metrics.

2. Literature Review and Theoretical Foundation

The SRGAN architecture builds upon several key innovations in deep learning:

Generative Adversarial Networks (GANs): The adversarial training paradigm enables the generation of realistic textures and fine details that traditional MSE-based approaches cannot capture.

Perceptual Loss Functions: The integration of VGG-based content loss provides feature-level supervision, encouraging the generator to produce images that are perceptually similar to ground truth rather than pixel-identical.

Residual Learning: The generator employs deep residual connections to facilitate gradient flow and enable training of very deep networks without degradation.

Sub-pixel Convolution: Efficient upsampling through pixel shuffle operations reduces computational overhead while maintaining reconstruction quality.

3. Methodology and Architecture

3.1 Two-Stage Training Strategy

The implementation employs a sophisticated two-stage training methodology designed to ensure stability and optimal performance:

Stage 1 - SRResNet Pre-training:

- Generator trained independently using L1 pixel loss

- Establishes a strong baseline with high PSNR performance
- Provides stable initialization for subsequent adversarial training
- Duration: Variable epochs until convergence

Stage 2 - Adversarial Fine-tuning:

- Joint training of generator and discriminator
- Multi-component loss function balances pixel accuracy with perceptual realism
- Discriminator provides adversarial feedback to enhance texture generation

3.2 Generator Architecture (SRResNet)

The generator follows a deep residual architecture optimized for super-resolution:

LAYER STAGE	COMPONENTS
INPUT LAYER	Conv(9×9, 64) + PReLU
RESIDUAL BODY	16 × ResidualBlock: Each Block = [Conv(3×3, 64) + BN + PReLU] × 2 + Skip Connection
POST-RESIDUAL	Conv(3×3, 64) + BN + Global Skip Connection
UPSAMPLING	2 × SubPixelConv (upscale=2): Each Block = Conv(3×3, 256) + PixelShuffle(2) + PReLU
OUTPUT LAYER	Conv(9×9, 3) + Tanh

Key Design Choices:

- **16 Residual Blocks:** Balances model capacity with computational efficiency
- **PReLU Activation:** Addresses vanishing gradient problem in deep networks
- **Batch Normalization:** Stabilizes training and accelerates convergence
- **Sub-pixel Convolution:** Efficient 4× upsampling through learned interpolation
- **Global Skip Connection:** Preserves low-frequency information throughout the network

3.3 Discriminator Architecture

The discriminator employs a VGG-inspired architecture for binary classification:

Layer Stage	Components
Input	96×96×3 HR/SR Image
Convolutional Blocks	8 × ConvolutionalBlock: • Pattern: [Conv(3×3) + BN + LeakyReLU(0.2)] • Channels: 64 → 128 → 128 → 256 → 256 → 512 → 512 → 512 • Stride alternates between 1 and 2 for downsampling
Global Average Pooling	AdaptiveAvgPool2d(6×6)
Classifier	Linear(18432 → 1024) + LeakyReLU→ Linear(1024 → 1)

Architecture Rationale:

- **Progressive Channel Growth:** Captures hierarchical features from edges to textures
- **Strided Convolutions:** Efficient spatial downsampling without max pooling
- **LeakyReLU:** Prevents gradient vanishing in discriminator
- **Adaptive Pooling:** Ensures consistent feature map size regardless of input variations

3.4 Perceptual Loss Network (TruncatedVGG19)

A pre-trained VGG19 network truncated at conv5_4 layer serves as a fixed feature extractor:

- **Frozen Parameters:** Prevents feature drift during training
- **ImageNet Pre-training:** Leverages learned representations for natural images
- **Layer Selection:** conv5_4 balances semantic understanding with spatial resolution
- **Normalization:** Inputs normalized to VGG's expected range

4. Loss Function Design

4.1 Pre-training Loss

$$L_{pretrain} = L1(I_{SR}, I_{HR})$$

Simple L1 pixel loss encourages accurate reconstruction while being less prone to blurring than L2 loss.

4.2 Adversarial Training Loss

Generator Loss (Multi-component):

$$L_G = \alpha \cdot L_{pixel} + \beta \cdot L_{content} + \gamma \cdot L_{adversarial}$$

Where:

$L_{pixel} = L1(I_{SR}, I_{HR})$ # Pixel fidelity

$L_{content} = MSE\big(\varphi(I_{SR}), \varphi(I_{HR})\big)$ # Perceptual similarity

$L_{adversarial} = -\log\Big(D\big(G(I_{LR})\big)\Big)$ # Adversarial realism

Discriminator Loss:

$$L_D = -\log\big(D(I_{HR})\big) - \log\Big(1 - D\big(G(I_{LR})\big)\Big)$$

Loss Weight Configuration:

- Content Loss Weight (α): 0.006
- Adversarial Loss Weight (γ): 0.001
- Pixel Loss Weight: 1.0 (implicit)

These weights were carefully tuned to balance pixel accuracy with perceptual quality.

5. Dataset Analysis and Preprocessing

5.1 CelebA Configuration

- **Total Images:** 30,000 high-quality facial images
    - **Training Split:** 24,000 images (80%)
    - **Validation Split:** 4,500 images (15%)
    - **Test Split:** 1,500 images (5%)

Domain-Specific Optimizations:

- **Dynamic Crop Margin:** Intelligent cropping ensures facial features remain centered
- **Rotation Disabled:** Preserves facial alignment for optimal results
- **Crop Size:** 96×96 HR patches optimized for facial detail preservation

5.2 DIV2K Configuration

- **Training Images:** 800 high-resolution natural images
- **Validation Images:** 100 images
- **Coverage:** Diverse natural scenes including landscapes, objects, and textures

General Image Optimizations:

- **Random Rotations:** 0°, 90°, 180°, 270° augmentation for robustness
- **Horizontal Flipping:** 50% probability for data diversity
- **Bicubic Downsampling:** High-quality LR generation for training

5.3 Preprocessing Pipeline

Spatial Processing:

1. **Random/Center Cropping:** 96×96 patches from HR images
2. **Bicubic Downsampling:** 4× reduction to create 24×24 LR inputs
3. **Augmentation:** Domain-appropriate transformations

Intensity Processing:

1. **Tensor Conversion:** PIL → PyTorch tensors
2. **Normalization:** [0,255] → [-1,1] range for stable training
3. **Batch Formation:** Efficient DataLoader with configurable batch size

6. Training Configuration and Implementation

6.1 Optimization Strategy

Optimizers: Adam with β_1=0.9, β_2=0.999

- Pre-training Learning Rate: 1e-4
- GAN Generator Learning Rate: 1e-4
- GAN Discriminator Learning Rate: 1e-4

Training Parameters:

- Batch Size: 16 (optimized for memory efficiency)

- HR Crop Size: 96×96 pixels
- LR Input Size: 24×24 pixels
- Scaling Factor: 4×

6.2 Advanced Training Features

Comprehensive Checkpointing:

- `checkpoint_latest.pth`: Resumable training state
- `checkpoint_best_psnr.pth`: Best PSNR performance
- `checkpoint_best_ssim.pth`: Best structural similarity
- `checkpoint_best_lpips.pth`: Best perceptual quality
- Milestone checkpoints at regular intervals

Multi-Platform Logging:

- Console output with progress tracking
- CSV metrics history for analysis
- JSON training summaries
- Weights & Biases (W&B) integration for real-time monitoring
- Matplotlib visualization generation

Robust Training Infrastructure:

- Automatic GPU detection and utilization
- Memory usage monitoring and reporting
- Error handling and recovery mechanisms
- Configurable validation frequency

7. Evaluation Metrics and Methodology

7.1 Quantitative Metrics

Peak Signal-to-Noise Ratio (PSNR):

- Measures pixel-wise reconstruction accuracy
- Higher values indicate better pixel fidelity
- Computed in dB scale for standard reporting

Structural Similarity Index (SSIM):

- Evaluates structural preservation and human perception alignment
- Range [0,1] with 1 indicating perfect similarity
- Considers luminance, contrast, and structure

Learned Perceptual Image Patch Similarity (LPIPS):

- Deep learning-based perceptual distance metric
- Lower values indicate better perceptual similarity

- Correlates strongly with human perception studies

7.2 Validation Strategy

Comprehensive Evaluation Pipeline:

- Post-epoch validation on entire validation set
- Per-image metric calculation with statistical analysis
- Mean and standard deviation reporting for robustness assessment
- Multi-metric optimization with separate best model tracking

Test Set Evaluation:

- CelebA: 1,500 held-out facial images
- Set14: Standard benchmark for natural image super-resolution
- Cross-domain evaluation for generalization assessment

8. Experimental Results and Analysis

8.1 CelebA Model Performance

8.1.1 CelebA Test Set Results

Model	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
SRResNet (PSNR-Opt)	**42.70**	**0.9669**	0.0833
SRGAN (Best PSNR)	42.38	0.9642	0.0566
SRGAN (Best SSIM)	42.39	**0.9661**	0.0633
SRGAN (Best LPIPS)	41.43	0.9572	**0.0368**

Key Observations:

- SRResNet achieves highest pixel-wise accuracy (42.70 dB PSNR)
- SRGAN variants trade pixel accuracy for perceptual quality
- Best LPIPS model shows 56% improvement in perceptual similarity
- Minimal SSIM degradation demonstrates preserved structural integrity

8.1.2 Set14 Cross-Domain Evaluation

Model	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
SRResNet (PSNR-Opt)	23.01	0.6916	0.3847
SRGAN (Best PSNR)	**24.03**	0.6718	0.2927
SRGAN (Best SSIM)	**24.37**	**0.6902**	0.3225
SRGAN (Best LPIPS)	24.29	0.6681	**0.2426**

Cross-Domain Analysis:

- Significant PSNR improvement: +1.36 dB with SRGAN (Best SSIM)
- 37% perceptual quality improvement with Best LPIPS model
- Demonstrates effective generalization beyond facial imagery

- SSIM preservation indicates robust structural understanding

8.2 DIV2K Model Performance

8.2.1 Set14 Natural Image Results

Model	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
SRResNet (PSNR-Opt)	23.25	0.6576	0.3616
SRGAN (Best PSNR)	**25.15**	**0.7191**	0.1901
SRGAN (Best SSIM)	**25.15**	**0.7191**	0.1901
SRGAN (Best LPIPS)	24.80	0.7038	**0.1577**

Performance Highlights:

- Exceptional PSNR improvement: +1.90 dB over SRResNet baseline
- Substantial SSIM enhancement: +0.0615 improvement
- Outstanding perceptual quality: 56% LPIPS improvement
- Best PSNR and SSIM models achieved identical performance

8.3 Comparative Model Analysis

8.3.1 Domain Specialization Effects

CelebA vs DIV2K on Set14:

- DIV2K model shows superior generalization (+1.12 dB PSNR improvement)
- CelebA model demonstrates facial domain specialization benefits
- Both models significantly outperform baseline SRResNet

Training Data Efficiency:

- CelebA: 24,000 training images achieved strong facial performance
- DIV2K: 800 diverse images enabled better generalization
- Quality over quantity principle validated for diverse natural images

8.3.2 Metric Trade-off Analysis

PSNR vs Perceptual Quality:

- Clear trade-off between pixel accuracy and perceptual realism
- LPIPS-optimized models sacrifice ~1-2 dB PSNR for 30-50% perceptual improvement
- SSIM remains relatively stable across optimization targets

Optimization Target Impact:

- Multi-metric optimization yields distinct model characteristics
- Best LPIPS models consistently achieve superior perceptual quality
- Best PSNR models maintain competitive performance across all metrics

9. Technical Implementation Achievements

9.1 Software Engineering Excellence

Robust Training Framework:

- Production-ready codebase with comprehensive error handling
- Automatic checkpoint recovery and training resumption
- Multi-platform logging and monitoring integration
- Memory-efficient implementation with GPU optimization

Advanced Features:

- Dynamic learning rate scheduling capabilities
- Comprehensive metric tracking and visualization
- Automated model selection based on multiple criteria
- Extensible architecture for future enhancements

9.2 Novel Technical Contributions

Domain-Adaptive Data Processing:

- CelebA-specific dynamic crop margin optimization
- Intelligent augmentation strategy selection
- Flexible dataset configuration system

Enhanced Evaluation Framework:

- Statistical robustness with standard deviation reporting
- Multi-metric optimization with separate model tracking
- Cross-domain evaluation capabilities
- Comprehensive performance analysis tools

10. Discussion and Insights

10.1 Architectural Design Validation

The two-stage training approach proved highly effective, with pre-training providing stable initialization for adversarial fine-tuning. The 16-block residual architecture strikes an optimal balance between model capacity and computational efficiency.

10.2 Loss Function Effectiveness

The multi-component loss function successfully balances pixel accuracy with perceptual quality. The carefully tuned weights (0.006 for content, 0.001 for adversarial) prevent adversarial training instability while enabling texture generation.

10.3 Dataset-Specific Insights

CelebA Specialization: Domain-specific optimizations significantly improve facial super-resolution quality, though at the cost of general applicability.

DIV2K Generalization: Despite limited training data (800 images), the diverse DIV2K dataset enables superior generalization to natural images.

10.4 Evaluation Metric Implications

The results demonstrate the importance of multi-metric evaluation. PSNR optimization produces numerically superior but perceptually inferior results, while LPIPS optimization yields visually compelling outputs with acceptable pixel accuracy trade-offs.

11. Limitations and Future Directions

11.1 Current Limitations

Resolution and GPU Constraints: Training was limited to 96×96 crops due to available GPU memory. This prevented testing the model's full potential on higher resolutions like **128×128 or 256×256**, which would require more powerful hardware and significantly longer training.

Poor Domain Transfer: The CelebA model is highly specialized for faces and performs poorly on general images. The DIV2K model is a much better generalist, confirming a clear trade-off between specialization and flexibility.

GAN Training Sensitivity: The adversarial training process is difficult to stabilize and requires careful hyperparameter tuning to avoid visual artifacts and mode collapse.

11.2 Future Directions

Scale to Higher Resolutions: Implement **Progressive Training** to efficiently train the model on **128×128 or 256×256** images, overcoming current hardware limits to generate more detailed outputs.

Upgrade Model Architecture: Replace the current generator blocks with more powerful **Residual-in-Residual Dense Blocks (RRDB)** from ESRGAN and integrate **Self-Attention Mechanisms** to improve texture quality.

Automate Smart Cropping: Develop a **saliency-based cropping** system for general datasets. This would automatically find and train on important objects, applying the successful logic from our CelebA pipeline to any image.

12. Conclusion

This comprehensive SRGAN implementation demonstrates the effectiveness of adversarial training for perceptually realistic super-resolution. The dual-model approach successfully addresses both domain-specific optimization (facial imagery) and general applicability (natural images). Key achievements include:

1. **Superior Perceptual Quality:** Up to 56% improvement in LPIPS scores over baseline methods
2. **Robust Architecture:** 16-block residual generator with efficient sub-pixel upsampling
3. **Advanced Training Framework:** Production-ready implementation with comprehensive monitoring
4. **Multi-Metric Optimization:** Balanced approach considering pixel accuracy and perceptual quality
5. **Cross-Domain Validation:** Demonstrated generalization capabilities across different image domains

The results validate the importance of perceptual loss functions in super-resolution and highlight the trade-offs between pixel accuracy and visual quality. The implementation provides a solid foundation for future research and practical applications in single image super-resolution.

13. References and Acknowledgments

This implementation builds upon the foundational work of several key papers in deep learning and computer vision. We primarily extend the concepts introduced by **Ledig et al. in "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network" (2017)**, which established the SRGAN framework. Our perceptual evaluation is heavily informed by the LPIPS metric proposed by **Zhang et al. in "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" (2018)**. The development of our robust training pipeline and comprehensive evaluation represents our primary engineering contribution to this established field.

Dataset Acknowledgments:

- CelebA: Large-scale CelebFaces Attributes Dataset

- DIV2K: DIVerse 2K resolution high quality images
- Set14: Standard benchmark dataset for single image super-resolution evaluation

Furthermore, this project was made possible by a suite of powerful open-source tools. We extend our sincere thanks to the developers of PyTorch for providing an excellent deep learning framework. The experiment tracking and visualization provided by the Weights & Biases platform were instrumental in monitoring, analyzing, and ensuring the reproducibility of our results.