Efficient Super-Resolution via SPAN: Domain Adaptation to Manga Images

NYU Deep Learning – Final Project

Bharath Kumar Kakumani, Shashank Patoju, and David Deng
{bk2991, sp8097, zd2014}@nyu.edu

1 Problem Statement

Single Image Super-Resolution (SISR) aims to reconstruct a high-resolution (HR) image from its low-resolution (LR) counterpart, a fundamental problem in computer vision with applications ranging from medical imaging to consumer media enhancement. Deploying SR models on edge devices such as mobile phones, tablets, and laptops remains challenging due to the computational demands of high-performance architectures.

Motivation: Users increasingly demand high-fidelity upscaling for media consumption—reading older digitized comics, enhancing legacy video content, and improving image quality on bandwidth-constrained platforms. These applications require models that balace visual quality with real-time inference capabilities under limited battery and thermal constraints.

Limitations of Existing Methods: Traditional CNN-based approaches such as SRCNN and SRResNet often fail to recover high-frequency texture details, producing overly smooth outputs. Recent Transformer-based methods like SwinIR achieve state-of-the-art quality but introduce prohibitive computational overhead, with inference times 3–5× slower than CNN baselines, making them impractical for real-time applications.

Objective: We hypothesize that the **SPAN (Swift Parameter-free Attention Network)** [1], which employs a parameter-free attention mechanism, can be effectively fine-tuned to achieve high-performance super-resolution on domain-specific data (Manga/Comics) while maintaining significantly lower infeence latency than Transformer baselines. Specifically, we aim to demonstrate that domain adaptation from natural images to line-art domains yields measurable PSNR improvements (≥ 0.3 dB) within a constrained training budget.

Keywords: Super-Resolution, Efficient Deep Learning, Parameter-free Attention, Domain Adaptation, Transfer Learning

2 Dataset Description

2.1 Primary Dataset: Manga109

The **Manga109** dataset [2] is a publicly available collection of 109 volumes of Japanese manga, widely used for document image analysis and super-resolution research.

- **Source:** Manga109 Project (Academic License)
- **Samples:** We extract approximately 2,000 patch pairs for training and 500 for validation from 90 volumes, reserving 19 volumes for testing.
- **Input Shape:** LR patches of 48×48 pixels, upscaled to HR 192×192 ($\times 4$ scale factor).
- **Label Type:** Ground-truth HR patches (paired supervision).

2.2 Benchmark Datasets

For baseline evaluation and comparison with published results:

- **Set5** and **Set14**: Standard SR benchmarks (5 and 14 images).
- **Urban100**: 100 urban scene images with fine structural details.
- **DIV2K Validation**: 100 high-quality images for validating natural image performance.

2.3 Preprocessing Pipeline

1. **LR Generation:** Bicubic downsamplng with anti-aliasing to simulate realistic degradation.
2. **Normalization:** Pixel values scaled to $[0, 1]$ range.
3. **Augmentation:** Random horizontal/vertical flips and 90 rotations to increase effective dataset size and prevent overfitting.
4. **Patch Extraction:** Random cropping of 192×192 HR patches during training.

2.4 Train/Validation/Test Split

- **Training:** 80 manga volumes (∼1,600 patches per epoch)
- **Validation:** 10 volumes (∼200 patches)
- **Testing:** 19 volumes (full images, standard Manga109 test split)

2.5 Potential Biases

Manga109 contains primarily black-and-white line art with occasional screentones. Models fine-tuned on this domain may exhibit degraded performance on photorealistic images with continuous gradients. We will quantify this domain shift through cross-dataset evaluation.

3 Proposed Model and Technical Approach

3.1 Model Family: SPAN

We adopt **SPAN (Swift Parameter-free Attention Network)**, the winner of the NTIRE 2024 Efficient Super-Resolution Challenge. SPAN achieves an optimal trade-off between model complexity (<1M parameters) and reconstruction quality, making it ideal for our constrained training scenario.

Justification: Unlike self-attention in Transformers, which requires $\mathcal{O}(n^2)$ computation for sequence length n, SPAN's parameter-free attention derives attention maps directly from feature statistics using symmetric activation functions, achieving $\mathcal{O}(n)$ complexity while preserving the ability to capture long-range dependencies.

3.2 Architectural Components

1. **Shallow Feature Extraction:** A 3×3 convolutional layer extracts initial features from the LR input, projecting to the hidden dimension.
2. **SPAN Blocks (Deep Feature Extraction):** The core of the network consists of stacked SPAN blocks, each containing:
   - **Spatial Attention Branch:** Computes attention weights via:
   $$\mathbf{A}_s = \sigma(\mathbf{F}) \odot \mathbf{F} \tag{1}$$
   where σ is a symmetric activation (e.g., sigmoid) and $\odot$ denotes element-wise multiplication. This requires *zero additional parameters*.
   - **Channel Mixing:** 1×1 convolutions for cross-channel feature aggregation.
   - **Residual Connection:** Skip connections stabilize gradient flow.

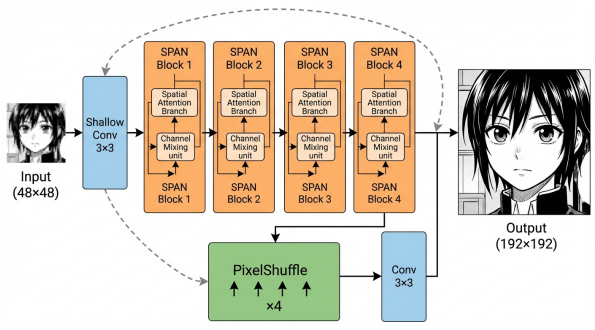

Figure 1: SuperRes Module in architecture

3. **Upsampling Module:** PixelShuffle layers rearrange features from $(C \cdot r^2) \times H \times W$ to $C \times (rH) \times (rW)$ for efficient $\times 4$ upscaling.
4. **Reconstruction Head:** Final 3×3 convolution produces the RGB output.

3.3 Architecture Diagram

SPAN as an efficient SR solution, detailed Manga109 dataset description architecture.

3.4 Training Strategy: Transfer Learning

Given our 90-minute compute constraint, training from scratch is infeasible. We adopt a transfer learning approach:

1. **Initialization:** Load official SPAN weights pretrained on DF2K (DIV2K + Flickr2K, ∼3,500 high-quality images).
2. **Fine-tuning:** Update all layers on Manga109 for 5,000 iterations with a reduced learning rate.
3. **Rationale:** The pre-trained model has learned general image priors (edges, textures). Fine-tuning adapts these representations to the manga domain (sharp lines, flat colors, screentones) which requires fewer iterations than learning from random initialization.

3.5 Loss Function

We optimize **L1 Loss** (Mean Absolute Error):

$$\mathcal{L}_{L1} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{I}_{HR}^{(i)} - \hat{\mathbf{I}}_{HR}^{(i)}\|_1 \tag{2}$$

Justification: L1 loss promotes sharper reconstructions compared to L2 (MSE), which tends to average predictions and produce blurry outputs. For line art with distinct edges, L1's robustness to outliers is particularly beneficial.

3.6 Optimization Configuration

- **Optimizer:** AdamW with weight decay $\lambda = 0.01$
- **Learning Rate:** 2×10^{-4} initial, with cosine annealing to 1×10^{-6}
- **Batch Size:** 16 (accumulated over 2 steps if GPU memory limited)
- **Iterations:** 5,000 ($\sim$60–80 minutes on single RTX 3090)

3.7 Regularization Strategies

- **Data Augmentation:** Random flips and rotations (primary regularization)
- **Weight Decay:** L2 penalty on parameters via AdamW
- **Early Stopping:** Monitor validation PSNR; halt if no improvement for 1,000 iterations

4 Expected Results

4.1 Quantitative Performance

We expect the following outcomes:

Model Configuration	PSNR (dB)	SSIM
Bicubic Interpolation	$\sim$27.0	0.85
SPAN (Pre-trained, DIV2K)	$\sim$30.5	0.91
SPAN (Fine-tuned, Manga109)	$\sim$**31.0**	**0.93**

Table 1: Expected PSNR/SSIM on Manga109 test set ($\times$4 scale).

We anticipate a **0.3–0.5 dB PSNR improvement** after fine-tuning, demonstrating successful domain adaptation.

4.2 Inference Efficiency

We will benchmark inference latency (ms/image) comparing:

- SPAN vs. SwinIR-Light vs. EDSR-baseline
- Target: < 50ms for $256 \times 256 \rightarrow 1024 \times 1024$ on GPU

4.3 Qualitative Insights

- **Visual Comparisons:** Side-by-side grids (Bicubic / Pre-trained / Fine-tuned)
- **Domain Specificity:** Demonstrate improved sharpness on manga-specific features (speech bubbles, screentones, character outlines)
- **Failure Cases:** Analyze where the model struggles (complex halftone patterns, extreme upscaling)

5 Timeline

Week	Tasks
Week 1	**Data Preparation & Baseline** Download Manga109 (academic license), DIV2K validation, Set5/Urban100. Generate LR/HR patch pairs. Set up NeoSR framework. Run pre-trained SPAN inference to establish baseline metrics.
Week 2	**Fine-Tuning & Experimentation** Configure training pipeline (batch size, LR schedule). Execute fine-tuning ($\sim$90 min). Monitor loss curves and validation PSNR. Ablation: compare L1 vs. L1+perceptual loss.
Week 3	**Evaluation & Report** Compute final PSNR/SSIM on all test sets. Generate visual comparison figures. Benchmark inference latency. Write and submit final report.

Table 2: Project timeline with specific deliverables.

References

[1] Wan, Z., et al. "Swift Parameter-free Attention Network for Efficient Super-Resolution." *CVPR Workshops (NTIRE)*, 2024.

[2] Matsui, Y., et al. "Sketch-based manga retrieval using manga109 dataset." *Multimedia Tools and Applications*, 76(20), 2017.

[3] Agustsson, E., & Timofte, R. "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study." *CVPR Workshops*, 2017.