Image Super-Resolution with SwinIR Trained on LSDIR: Baseline Results, Ablations, and Comparison to RCAN

Arda Öztürk (22103616), Umut Başar Demir (22102376), Süleyman Talha Belen (22202791)
CS 484 – Introduction to Computer Vision

***Abstract*—Single-Image Super-Resolution (SISR) reconstructs a high-resolution (HR) image from a degraded low-resolution (LR) input. We implement and train SwinIR, a Transformer-based restoration model that uses shifted-window self-attention. Training is performed from scratch on a restricted subset of LSDIR and evaluated on Manga109. Our $\times 2$ baseline achieves 36.18 dB PSNR and 0.9641 SSIM on Manga109, and performance decreases for $\times 3$ and $\times 4$ as expected. We also present ablation studies on $\times 2$ by varying channel width, number of RSTBs, layers per RSTB, and patch size. Finally, we compare SwinIR against RCAN trained under the same data regime and discuss trade-offs between capacity and data/compute budget.**

***Index Terms*—single-image super-resolution, SwinIR, Swin Transformer, LSDIR, Manga109, ablation study**

I. Introduction

Single-Image Super-Resolution (SISR) aims to recover an HR image from a single LR observation. As the scaling factor increases, more high-frequency information is missing, making the problem harder. CNN-based approaches are strong in local modeling, while Transformer-based approaches can capture longer-range dependencies. SwinIR combines shifted-window self-attention with residual Transformer blocks to achieve efficient and high-quality restoration.

II. Method: SwinIR

SwinIR consists of: (i) shallow feature extraction using a convolutional embedding, (ii) deep feature extraction via Residual Swin Transformer Blocks (RSTBs), and (iii) reconstruction using feature fusion and upsampling (pixel shuffle). Attention is computed within non-overlapping windows; shifting the window partition between layers enables cross-window information flow with manageable computation.

III. Datasets and Preprocessing

We train using LSDIR. For evaluation, we use Manga109 and generate LR inputs via bicubic downsampling from HR ground truth. During training, aligned LR–HR patches are extracted and simple augmentations (e.g., flips/rotations) are applied.

A. Why LSDIR is a Strong Training Dataset

LSDIR is a strong choice for training super-resolution models because it provides **large-scale, diverse, and high-quality** images that help the network learn generalizable priors. Compared to small benchmark datasets, large collections reduce overfitting to a narrow content distribution and expose the model to **many different edge types, textures, and natural statistics** (e.g., repeated patterns, smooth regions, high-frequency details). This is especially useful for Transformer-based models like SwinIR, which can benefit from seeing **wide visual diversity** in order to learn robust long-range dependencies.

Another advantage is that LSDIR is designed for **image restoration settings**: it supports training pipelines where we synthesize LR inputs (e.g., bicubic downsampling) from high-quality HR sources and train with aligned LR–HR pairs. Overall, LSDIR provides a realistic and diverse training distribution, which improves the chance that the learned SR mapping transfers well to a different evaluation domain such as Manga109. In our project, we use a designated subset of LSDIR consisting of 1000 images, which may reduce the achievable PSNR compared to paper's SwinIR which has around 2000 images.

IV. Why We Compare with RCAN

We chose **RCAN** as our main comparison model because it is a very strong and well-known CNN baseline for super-resolution. RCAN uses a **residual-in-residual** structure to train a very deep network, and it uses **channel attention** to give more focus to useful feature channels [1]. Because of this design, RCAN became one of the standard reference methods in super-resolution research. For that reason we trained our own RCAN on the same LSDIR subset.

We also chose RCAN because the **SwinIR paper compares against RCAN** and shows that SwinIR can achieve **better PSNR with fewer parameters** than strong CNN-based methods [2]. This makes RCAN a fair and meaningful baseline to compare with.

Fig. 1: Example images from the LSDIR dataset, illustrating its large visual diversity (scenes, objects, textures) which helps training restoration models to generalize [3].

V. Experimental Setup

A. Baseline Configuration

Our baseline uses: channels=180, #RSTBs=6, layers/RSTB=6, patch size 64×64. We report $\times 2$, $\times 3$, $\times 4$ results, and run ablations on $\times 2$.

B. Training Details

We train with Adam, L1 loss, initial learning rate 2×10^{-4}, batch size 8, and a MultiStepLR schedule. Training is performed in Google Colab on an NVIDIA A100 GPU and checkpoints are saved for convergence monitoring.

VI. Results

A. Quantitative Evaluation

Table I reports baseline PSNR/SSIM on Manga109. Performance decreases as the scale increases due to higher reconstruction ambiguity.

TABLE I: Baseline PSNR/SSIM on Manga109 for Three Upscaling Factors

Scale	PSNR (dB)		SSIM	
	Paper	Ours	Paper	Ours
$\times 2$	39.60	36.18	0.9792	0.9641
$\times 3$	34.74	31.54	0.9518	0.9468
$\times 4$	31.67	28.86	0.9226	0.9293

a) Discussion of Table I (Baselines).: Table I shows the expected degradation as the upscaling factor increases: $\times 2$ is relatively constrained (less missing information), while $\times 3$ and $\times 4$ require hallucinating more high-frequency details. Therefore, PSNR drops from 36.18 dB ($\times 2$) to 31.54 dB ($\times 3$) and 28.86 dB ($\times 4$), which matches the general SISR difficulty trend.

The gap to the paper numbers is mainly explained by **dataset size (1000 vs 2000 image)**: we trained from scratch on a **restricted LSDIR subset**. Our baseline SwinIR and paper's SwinIR are both trained on 500k iterations. In addition, even small evaluation details (e.g., border cropping, preprocessing) can shift PSNR/SSIM. For this reason, the most reliable takeaway here is the **relative trend across scales** rather than absolute parity with the paper. Also, human eye cannot notice the difference after 30dB PSNR.

b) Note on SSIM vs. PSNR.: SSIM and PSNR do not always move perfectly together: PSNR is pixel-wise error sensitive, while SSIM emphasizes structural similarity. Therefore, it is possible to observe slightly higher SSIM even when PSNR is lower, especially if the model produces smoother but structurally consistent outputs.

B. Qualitative Evaluation

SwinIR improves edge sharpness and recovers finer textures compared to bicubic interpolation. Error maps typically show larger errors around sharp boundaries, thin lines, and text-like structures.

a) Qualitative behavior and difficulties.: Visually, SwinIR tends to recover sharper edges and cleaner textures compared to bicubic interpolation, which is consistent with its ability to use contextual cues through shifted-window attention (see Fig. 3). However, the remaining artifacts are not random: the largest errors often concentrate on (i) **thin high-contrast lines**, (ii) **text-like strokes**, and (iii) **fine repeated patterns**. These regions are difficult because a small spatial shift in the reconstruction can cause large pixel-wise error, and because multiple HR solutions can be plausible given the same LR input.

In the error visualizations, high-error masks usually highlight boundaries and line art, indicating that the model is

learning a strong global prior but still struggles with **precise geometric alignment** at high frequencies. This is also why $\times 3$ and $\times 4$ look perceptually reasonable but score lower in PSNR: small misalignments are heavily penalized.

VII. Ablation Study on $\times 2$ (Manga109)

We vary one hyperparameter at a time and evaluate on Manga109 at $\times 2$. Tables II–V show the effects of channel width, #RSTBs, layers per RSTB, and patch size. Among these, increasing patch size provides the strongest PSNR gain in our runs.

TABLE II: Ablation on $\times 2$ (Manga109): Effect of Channel Width

Channels	PSNR (dB)
150	35.08
180 (baseline)	**35.16**
210	35.57

TABLE III: Ablation on $\times 2$ (Manga109): Effect of #RSTBs

#RSTBs	PSNR (dB)
4	34.85
6 (baseline)	**35.16**
8	35.46

TABLE IV: Ablation on $\times 2$ (Manga109): Effect of Layers per RSTB

Layers/RSTB	PSNR (dB)
4	34.70
6 (baseline)	**35.16**
8	35.54

TABLE V: Ablation on $\times 2$ (Manga109): Effect of Patch Size

Patch size	PSNR (dB)
56	34.99
64 (baseline)	**35.16**
72	36.08

a) Interpreting the ablations.: Overall, the ablations show a clear **capacity vs. budget trade-off**. Increasing channel width, the number of RSTBs, or the number of layers per RSTB consistently improves PSNR, which suggests that our baseline is not yet at a saturation point under the current data regime. At the same time, the gains are relatively moderate, indicating **diminishing returns**: as the model becomes larger, extra parameters help but each step adds less improvement per additional compute.

The patch size ablation produces the strongest gain in our runs. A larger patch gives the model **more spatial context per update**, which can be especially important for SwinIR because attention-based features benefit from seeing coherent structures (edges, repeated textures) within a single training crop. In other words, bigger patches make it easier for the model to learn **longer-range consistency**, reducing local ambiguity and improving reconstruction quality.

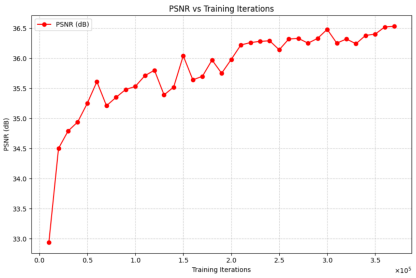
Practically, larger patches also change the effective training signal: they contain more diverse pixels/structures per iteration, which can stabilize optimization and improve generalization. The downside is higher memory/compute cost, so patch size becomes one of the most impactful knobs when training resources are limited.

b) Note on absolute numbers in ablations.: The ablation runs were performed with a shorter (50k iterations for each parameter analysis) training schedule compared to the final baseline checkpoint, so the absolute PSNR values may differ; the main purpose is to analyze **relative sensitivity** to each hyperparameter.

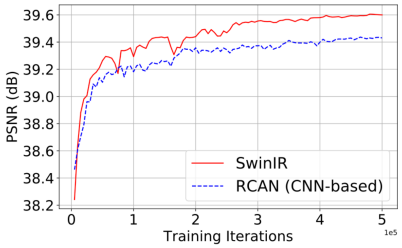
VIII. PSNR vs Training Iterations Comparison

In this section, we compare PSNR curves across training iterations. We include three plots: (1) our SwinIR PSNR curve, (2) SwinIR paper curve, and (3) our RCAN PSNR/SSIM curve. The goal is to see the general training behavior (fast improvement early, then slower improvement later).

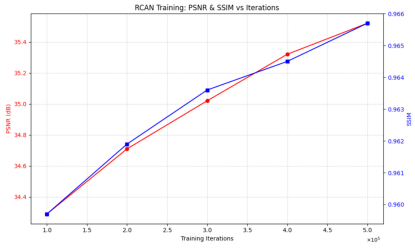
A. Comparison Plots (3 images)



(a) Our SwinIR: PSNR vs iterations.



(b) SwinIR paper: SwinIR vs RCAN.



(c) Our RCAN: PSNR & SSIM vs iterations.

Fig. 5: PSNR vs training iterations comparison (our results vs paper).

(a) Generated LR (×2) (b) HR Ground Truth (c) SwinIR Output

Fig. 2: Visual comparison of super-resolution results for baseline ×2 (Left: generated LR image, Middle: HR ground truth, Right: SwinIR output).

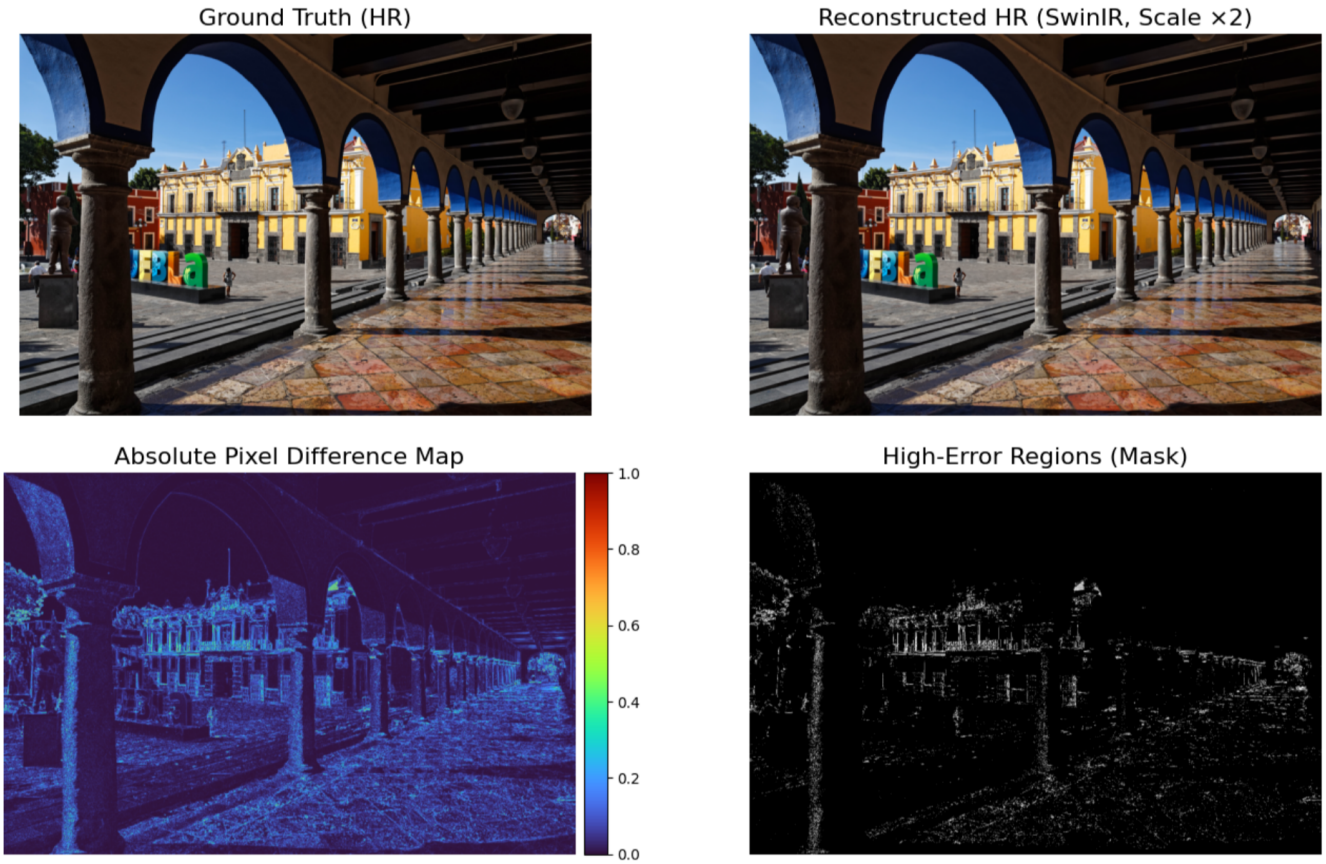


Fig. 3: Error analysis for baseline ×2: (top-left) HR ground truth, (top-right) SwinIR reconstruction, (bottom-left) absolute pixel difference map, (bottom-right) high-error regions mask.

(a) HR (GT) (b) LR (×2) (c) LR (×3) (d) LR (×4)

(e) SwinIR (×2) (f) SwinIR (×3) (g) SwinIR (×4)

Fig. 4: Qualitative comparison on one sample: HR ground truth, LR inputs, and SwinIR reconstructions for ×2, ×3, and ×4. All images are displayed at the same size for fair visual comparison.

B. Simple observations

- **Our SwinIR:** PSNR increases quickly early, then increases slowly and becomes almost stable later. Our SwinIR model beats RCAN (36.5 dB versus 35.6 dB).
- **Paper plot:** SwinIR stays slightly higher than RCAN near the end, and both curves saturate over time.
- **Our RCAN:** PSNR and SSIM increase together, but the improvement becomes smaller as iterations grow.

a) Learning dynamics from PSNR curves.: The curves show a typical restoration training pattern: a rapid PSNR increase at early iterations (the model quickly learns coarse restoration and basic edge recovery), followed by a slower improvement phase where the network refines high-frequency details. The later saturation suggests that additional gains likely require either (i) more data diversity, (ii) longer training with a lower learning rate, or (iii) stronger supervision signals depending on the target metric.

Comparing SwinIR and RCAN trends, SwinIR reaching a higher plateau is consistent with its stronger global modeling: attention helps to enforce longer-range consistency, which can be important in structured images (e.g., line art). However, the gap size is sensitive to training budget; under limited data/iterations, both models may saturate earlier than reported in large-scale paper settings.

Note: We arranged fewer checkpoints for RCAN to see the final result which we compare immediately. The paper values and our values may not be directly comparable because dataset, scale, and training settings can be different. Here, the main goal is to compare the trend of the curves.

IX. Comparison with RCAN

We compare SwinIR with RCAN trained under the same restricted LSDIR subset. SwinIR is more parameter-efficient in our setting and achieves higher PSNR on Manga109 for ×2.

TABLE VI: SwinIR vs. RCAN on Manga109 (×2)

Model	Params (M)	PSNR (dB)	SSIM
RCAN (original)	∼16	39.44	0.9786
RCAN (ours, LSDIR subset)	∼15.6	35.46	0.9660
SwinIR (original)	∼12	39.92	0.9797
SwinIR (ours, LSDIR subset)	∼11.8	36.18	0.9641

In our experiments, we also observe a similar trend: **our SwinIR model reaches higher PSNR than our RCAN model**, and it does this with a more parameter-efficient design. The SwinIR paper also reports that SwinIR can outperform

state-of-the-art methods while reducing the number of parameters [2]. So overall, RCAN is a great model, but SwinIR usually beats it in both **accuracy** (PSNR) and **efficiency**.

a) Discussion: accuracy vs. parameter efficiency.: Under the same restricted LSDIR training regime, SwinIR achieves higher PSNR than RCAN while using fewer parameters, which highlights a practical advantage: **better quality per parameter**. This suggests that SwinIR's shifted-window attention can utilize limited training data effectively by leveraging broader context, while RCAN's CNN features remain strong but more locally biased.

At the same time, the original paper numbers for both models are significantly higher than our re-trained versions. This reinforces the key experimental lesson: in super-resolution, **training regime matters as much as architecture**. With larger-scale training (more images/iterations), both RCAN and SwinIR improve substantially, and the paper-level gap becomes more reproducible.

X. Conclusion

We implemented and trained SwinIR for SISR on an LSDIR subset and evaluated on Manga109. Baselines show expected degradation as the upscaling factor increases. Ablations indicate that increasing capacity helps, while patch size provides the strongest improvement under our budget. Compared to RCAN trained with the same data regime, SwinIR yields better PSNR with fewer parameters in our runs. Future work includes scaling training data/iterations and exploring improved losses for perceptual quality.

A key takeaway is that most of the remaining gap (which is impossible to detect by human eyes) to paper-level results is explained by **data scale**, not a fundamental limitation of the implementation. Among the tested knobs, **patch size** gave the best improvement under limited compute, indicating that context per iteration is critical for SwinIR. Future improvements can prioritize on a larger LSDIR portion. However, with the half size of dataset, results show perfect reconstruction.

References

[1] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image Super-Resolution Using Very Deep Residual Channel Attention Networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[2] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and Y. Timofte, "SwinIR: Image Restoration Using Swin Transformer," in *ICCV Workshops (AIM Workshop)*, 2021.

[3] Y. Li *et al.*, "LSDIR: A Large-Scale Dataset for Image Restoration," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 1775–1787.

[4] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. ICCV*, 2021.

[5] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. ICPR*, 2010.