Super Resolution on JPEG Compressed Image: Overcoming CNN Limitations via Swin Transformer Architecture

Sangwook Lee Cham Lee Muyeong Jeong
Department of Computer Science
Korea University
{vecum0814, dlcka1111, jmy3033}@korea.ac.kr

Abstract

This report presents a comparative study on super-resolution (SR) for JPEG-compressed images, focusing on overcoming the limitations of conventional CNN-based models by integrating the Swin Transformer architecture. We explore multiple hybrid structures of CNN and Transformer blocks to efficiently capture both local and global contexts. Experimental results demonstrate that Transformer-augmented models consistently outperform the CNN-only baseline in terms of PSNR, while maintaining moderate model complexity. Among the various designs, a parallel architecture of CNN and Swin Transformer yields the highest performance. Although super-resolution intuitively contributes to improved OCR accuracy by reducing blocking artifacts, edge distortions, and ringing effects, further experiments suggest that such improvements do not always translate into meaningful gains in OCR accuracy.

1. Introduction

Sharing travel experiences through images is a joyful and meaningful activity, especially in today's globally connected world. However, when using a smart phone in many regions, limited internet bandwidth makes it difficult to transmit high-quality photos in a timely manner. As a result, users often resort to compressing images using lossy formats such as JPEG, which introduces noticeable artifacts like blocking and ringing. These artifacts degrade the visual quality and may lead to dissatisfaction for both the sender and recipients. Despite this inconvenience, users are typically left with few practical alternatives when fast or stable mobile network access is unavailable. Therefore, this study focuses on improving the performance of super-resolution (SR) techniques for JPEG-compressed images.

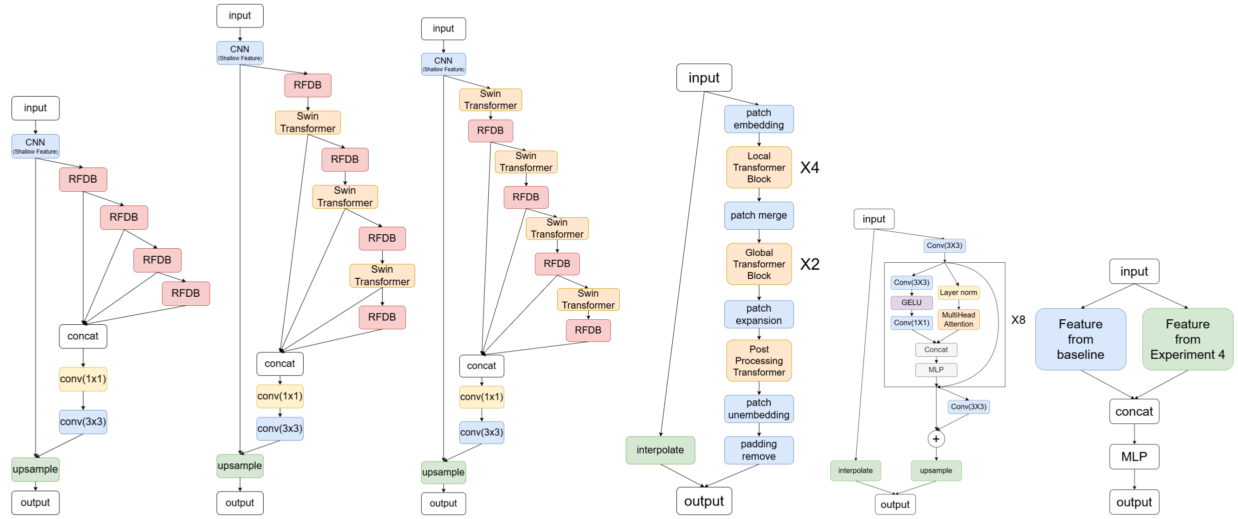
In this study, adopted Peak Signal-to-Noise Ratio (PSNR) as an evaluation metric for super-resolution performance. PSNR is a widely used metric that quantifies pixel-level differences between reconstructed and ground-truth images, making it more aligned with machine-level fidelity than human perception. We speculated that this limitation may cause a mismatch between high PSNR scores and actual visual or functional quality, particularly in tasks where human-level understanding matters. To bridge this gap, we considered Optical Character Recognition (OCR) accuracy as a complementary evaluation method. Our reasoning was that if certain characters are difficult for humans to recognize due to compression artifacts, OCR systems may also fail to read them correctly.

However, based on the feedback from our professor, we learned that PSNR is already a sufficiently robust indicator in this context, and the difference in OCR performance between low-quality (LQ) and high-quality (HQ) images might not be as significant as expected. To verify this, we conducted additional experiments comparing OCR results on both LQ and HQ images. This allowed us to more thoroughly assess whether super-resolution meaningfully contributes to machine-level understanding tasks such as OCR.

2. Related work

For our baseline model, we selected the Residual Feature Distillation Network (RFDN)[1]. Given that the motivation involved smartphone usage, we considered lightweight super-resolution models to be more appropriate for this study. RFDN, being a compact and efficient CNN-based architecture, aligns well with this requirement. Furthermore, since RFDN is purely convolutional, we hypothesized that its performance could be further improved by integrating Transformer modules, which are known to capture long-range dependencies more effectively.

There are various Transformer variants designed for vision tasks. In this study, we employed architectures inspired by the Vision Transformer (ViT)[2] as well as the Swin Transformer[3]. The Swin Transformer divides images into non-overlapping patches and applies self-attention within local windows, effectively capturing local features, sup-



(a) Baseline (b) RFDB first (c) Transformer first (d) Transformer only (e) Integrated parallel (f) Separate parallel

Figure 1. The structure of the model used in the super resolution experiment

plementing a limitation of ViT. By shifting these windows across layers, it incrementally incorporates global context as well. Additionally, Swin Transformer is more computationally efficient than the original (Vanilla) Transformer, making it well-suited for high-resolution image processing.

3. Experiments

The super-resolution experiment requires both high-quality (HQ) and low-quality (LQ) images. As the source of HQ images, we utilized the DIV2K dataset[5], which contains 800 training images and 100 testing images at 2K resolution. To simulate realistic image degradation scenarios—such as those found in messaging platforms like KakaoTalk—we generated two LQ versions of each training image by applying JPEG compression with quality levels of 95 and 90. This resulted in a total of 1,600 LQ training samples. In addition, the input resolution was reduced to one-fourth, and a ×4 super-resolution reconstruction was applied. The model was trained on this augmented dataset, and PSNR was evaluated on the reconstructed output.

3.1. Experiment 1

In Experiment 1, we evaluated the performance of the baseline model, RFDN. Following the original architecture of RFDN, as shown in Figure 1a, the model extracts and aggregates features using only convolutional layers. The PSNR score initially measured at 25.592, and increased to 26.082 after 20 training epochs.

3.2. Experiment 2

In Experiment 2, we extended the baseline model by incorporating a Swin Transformer. To sequentially process both local and global contexts, we alternated the application of RFDB and the Swin Transformer in a repeated manner as shown in Figure 1b. The PSNR score initially measured at 26.073, and increased to 26.250 after 20 training epochs.

3.3. Experiment 3

In Experiment 3, the Transformer module was applied prior to the RFDB module. We hypothesized that applying the Transformer after the RFDB might reduce overall effectiveness, as local context would be processed before global context. To test this hypothesis, we applied the Transformer before the RFDB modules in Experiment 3, as shown in Figure 1c. The PSNR score initially measured at 26.078, and increased to 26.283 after 20 training epochs.

3.4. Experiment 4

Inspired by the novel paper ”Attention is All You Need” [4], we removed all CNN components and employed only Transformer blocks in this experiment, as illustrated in Figure 1d. Additionally, We utilized ViT rather than Swin Transformer to specifically investigate the impact of fully leveraging global self-attention in the super-resolution task. The objective was to capture both local and global context using solely self-attention mechanisms. To achieve this, we divided each input into 4x4 patches for the Local Transformer and subsequently merged them into 8x8 patches for the Global Transformer, allowing the model to capture features over a significantly larger area. The PSNR score initially measured at 26.208, and increased to 26.225 after 20 training epochs.

Experiment	PSNR	20 Epoch PSNR	Gain (vs. Exp1)	Params	Param Increase (vs. Exp1)
Baseline	25.592	26.082	-	433,448	-
RFDB first	26.073	26.250	+0.168	1,124,913	x2.6
Transformer first	26.078	26.283	+0.201	1,161,435	x2.68
Transformer Only	26.208	26.225	+0.143	765,440	x1.76
Integrated parallel	26.491	26.716	+0.634	1,004,035	x2.31
Separate parallel	19.885	25.589	-0.473	1,204,638	x2.77

Table 1. PSNR and parameter comparison across different SR model variants.

3.5. Experiment 5

In Experiment 5, we aimed to leverage the advantages of both CNN and Transformer simultaneously. In Experiments 2 and 3, local and global contexts were processed sequentially within a single pathway. Inspired by this, in Experiment 5, as shown in Figure 1e, we designed the model to process local and global contexts in parallel pathways, allowing them to be learned simultaneously. The features extracted from each branch are then fused to produce a single output. The PSNR score initially measured at 26.491, and increased to 26.716 after 20 training epochs.

3.6. Experiment 6

Motivated by the feedback from our professor, Experiment 6 aimed to examine the potential synergy of integrating separately optimized models for local and global context processing, considering that their parameter counts were also comparable. In Experiment 5, the two branches were designed to be trained simultaneously. However, in Experiment 6, as shown in Figure 1f, we used the separately trained baseline model and the model from Experiment 4. From the baseline model, features were extracted before the upsampling step, while from the model in Experiment 4, features were extracted before the unembedding step. These two sets of features were then fused to generate the final output. The PSNR score initially measured at 19.885, and increased to 25.589 after 20 training epochs.

3.7. Experiment 7

In Experiment 7, we aimed to verify whether there is a performance difference in OCR between low-quality (LQ) and high-quality (HQ) images, based on the professor's feedback. We used the "AI Hub Korean OCR Dataset"[6] for the experiments, and employed the ko-trocr model[8], an adaptation of TrOCR[7] for the Korean language. To create LQ images, we applied JPEG compression to HQ images and cropped them to match the reduced resolution. OCR was then performed separately on both HQ and LQ sets.

4. Results

The results of Experiments 1 through 6 concerning SR are summarized in Table 1. The PSNR score of the baseline model was lower than that of the models incorporating Transformer blocks. This result suggests that architectures leveraging Transformers to capture global context outperform those relying solely on CNNs. However, it is also possible that the baseline model's lower performance is attributable to its significantly smaller number of parameters.

A comparison between Experiments 2 and 3 shows that Experiment 3 achieved a higher PSNR score. This observation implies that processing global context prior to local context may lead to better overall performance. It suggests that capturing the global structure early, followed by local detail refinement, could be a more effective strategy.

In Experiment 4, the initial PSNR was higher than in Experiments 2 and 3, likely due to the Transformer's strong ability to rapidly capture global dependencies. However, the subsequent performance improvement during training was limited. This limitation primarily stems from the Transformer's lack of an explicit spatial inductive bias, in contrast to CNNs. As a result, Transformers tend to struggle with efficiently modeling local context and fine-grained details. Furthermore, due to their high model capacity and sensitivity to dataset size, Transformers are prone to overfitting on small datasets, leading to unstable training dynamics and restricted restoration quality in prolonged training scenarios.

Although Experiment 6 initially underperformed compared to the baseline model, this outcome suggests that additional training is required to effectively fuse independently learned features. Notably, Experiment 6 exhibited the largest PSNR improvement among all experiments. This finding indicates that the initial performance degradation resulted from suboptimal feature fusion, but with continued training, the fusion became more effective, yielding significant performance gains.

The best-performing model was the integrated parallel architecture from Experiment 5, which outperformed Experiments 2, 3, and 6 despite having fewer parameters. In this design, CNNs are responsible for restoring local details, while the Transformer captures global patterns. Sequentially processing both feature types in a single branch can lead to role ambiguity and unstable learning. In contrast, the parallel configuration minimizes such interference and facilitates more stable and effective learning, resulting in superior performance.

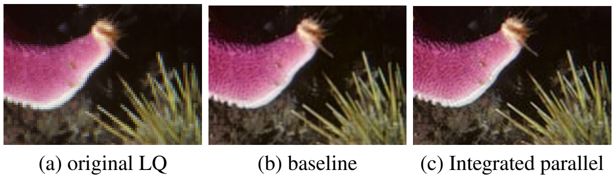

(a) original LQ (b) baseline (c) Integrated parallel

Figure 2. SR results of the baseline and Experiment 5

We illustrate the results of the SR task using images. Figure 2a shows an crop of the original low-quality (LQ) image used in the SR task. Figure 2b and Figure 2c present the reconstructed results obtained from the baseline model and the best-performing Integrated parallel model, respectively. As shown in Figure 2b, faint blocking artifacts remain visible, whereas in Figure 2c, such artifacts are barely distinguishable to the naked eye. This indicates that the application of the transformer not only improves the PSNR score but also enhances visual quality.

Metric	**HQ**	**LQ**
WER (mean)	0.3813	0.3889
CER (mean)	0.1439	0.1448
Top-1 Accuracy	0.6187	0.6111

Table 2. Comparison of OCR performance metrics between high-quality (HQ) and low-quality (LQ) images.

The results of Experiment 7 are summarized in Table 2. Although HQ images exhibited slightly better performance in terms of WER, CER, and Top-1 Accuracy, the differences were minimal. This indicates that the OCR model maintains nearly consistent performance regardless of compression level. Consequently, our initial assumption that 'OCR performance would be better on HQ images than on LQ images' was found to be incorrect. Given the negligible performance difference, we conclude that OCR metrics are not meaningful indicators for evaluating the effectiveness of super-resolution tasks that aim to improve PSNR.

5. Conclusion

We aimed to improve super-resolution (SR) performance on JPEG-compressed images by integrating Transformer[3] modules into conventional CNN-based architectures. Through experiments with various hybrid and sequential designs, we explored how global and local contexts can be effectively captured and fused.

Notably, processing global context prior to local refinement, as in Experiment 3, proved more effective than the reverse order. Experiment 6 demonstrated the highest PSNR improvement despite its initially low performance, suggesting that effective fusion of independently learned features requires sufficient training.

Among the SR models, a integrated parallel architecture in Experiment 5 that processes local and global features separately and then merges them showed the most promising results.

Additionally, we hypothesized that improved SR quality might enhance OCR accuracy by reducing compression artifacts such as blocking and ringing. However, through Experiment 7, we observed minimal differences in OCR performance between high- and low-quality images. This suggests that modern OCR models, especially those based on Transformer encoders, are already robust against moderate degradation due to compression.

References

[1] Jie Liu., Jie Tang., Gangshan Wu. (2020). *Residual Feature Distillation Network for Lightweight Image Super-Resolution*. arXiv:2009.11551. https://arxiv.org/abs/2009.11551

[2] Alexey Dosovitskiy., Lucas Beyer., Alexander Kolesnikov., Dirk Weissenborn., Xiaohua Zhai., Thomas Unterthiner., Mostafa Dehghani., Matthias Mindere.r, Georg Heigold., Sylvain Gelly., Jakob Uszkoreit., Neil Houlsby. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv:2010.11929. https://arxiv.org/abs/2010.11929

[3] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. In ICCV. https://arxiv.org/abs/2103.14030

[4] Ashish Vaswani., Noam Shazeer., Niki Parmar., Jakob Uszkoreit., Llion Jones., Aidan N. Gomez., Lukasz Kaiser., Illia Polosukhin. (2017). *Attention Is All You Need*. arXiv:1706.03762. https://arxiv.org/abs/1706.03762

[5] DIV2K High Resolution Images: https://gts.ai/dataset-download/div2k-high-resolution-images/

[6] AI Hub Korean OCR Dataset: https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=105

[7] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D. A. F., Zhang, C., Li, Z., & Wei, F. (2021). *TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models*. arXiv:2109.10282. https://arxiv.org/abs/2109.10282

[8] Huggingface - TrOCR: https://huggingface.co/ddobokki/ko-trocr