SwinIR: Image Restoration Using Swin Transformer

Sanjeev M, *BE21B034*

Abstract

Image restoration is a long-standing low-level vision problem that aims to restore high-quality images from low-quality images. While state-of-the-art image restoration methods are based on convolutional neural networks, few attempts have been made with Transformers which show impressive performance on high-level vision tasks. In this paper, we propose a strong baseline model SwinIR for image restoration based on the Swin Transformer. I conduct experiments on two representative tasks: image super-resolution (real-world image super-resolution), image denoising (including grayscale and color image denoising).

I. Introduction

IMAGE restoration, such as image super-resolution (SR) and image denoising, aims to reconstruct the high-quality clean image from its low-quality degraded counterpart. Since several revolutionary work, convolutional neural networks (CNN) have become the primary workhorse for image restoration. They generally suffer from two basic problems that stem from the basic building block, i.e., the convolution layer. First, the interactions between images and convolution kernels are **content-independent**. Second, with the principle of local processing, convolution is not effective for **long-range dependency modelling**.

As an alternative to CNN, Transformer designs a self-attention mechanism to capture global interactions between contexts. However, vision Transformers for image restoration usually divide the input image into small patches with fixed size and process each patch independently. Such a strategy inevitably rises two drawbacks. First, the restored image may introduce **border artifacts** around each small patch. Second, the border pixels of **each patch lose information** for better restoration.

Swin Transformer integrates the advantages of both CNN and Transformer. It has the advantage of CNN to process image with large size due to the local attention mechanism and also as a Transformer to model long-range dependency with the shifted window scheme. Compared with prevalent CNN-based image restoration models, Transformer-based SwinIR has several benefits: (1)content-based interactions between image content and attention weights, which can be interpreted as spatially varying convolution. (2) long-range dependency modelling are enable by the shifted window mechanism. (3) better performance with less parameters.

II. Network Architecture

A. Shallow Feature Extraction

Shallow feature extraction module uses a convolution layer to extract shallow feature, which is directly transmitted to the reconstruction module so as to Given a low-quality (LQ) input $I_{\text{LQ}} \in \mathbb{R}^{H \times W \times C_{\text{in}}}$ (where

H, W, and C_{in} are the image height, width, and input channel number, respectively), we use a 3×3 convolutional layer $HSF(\cdot)$ to extract shallow feature $F_0 \in \mathbb{R}^{H \times W \times C}$.

$$F_0 = H_{\text{SF}}(I_{\text{LQ}})$$

where C is the feature channel number. The convolution layer is good at early visual processing. It maps the input image space to a higher dimensional feature space.

B. Deep Feature Extraction

We extract the deep feature $F_{\text{DF}} \in \mathbb{R}^{H \times W \times C_{\text{in}}}$ FROM F_0 as

$$F_{\text{DF}} = H_{\text{DF}}(F_0)$$

where $HDF(\cdot)$ is the deep feature extraction module and it contains K residual Swin Transformer blocks (RSTB) and a 3 × 3 convolutional layer. More specifically, intermediate features $F_1, F_2, ..., F_K$ and the output deep feature F_{DF} are extracted block by block as

$$F_i = H_{RSTB_{\text{i}}}(F_{i\text{-}1}),\ i = 1, 2, 3...k\ F_{\text{DF}} = H_{\text{CONV}}(F_{\text{k}})$$

where $H_{RSTB_{\text{i}}}(\cdot)$ denotes the i-th RSTB and H_{CONV} is the last convolutional layer. Using a convolutional layer at the end of feature extraction can bring the inductive bias of the convolution operation into the Transformer-based network for aggregation of shallow and deep features.

C. Image Reconstruction

Taking image SR as an example, we reconstruct the high-quality image I_{RHQ} by aggregating shallow and deep features as

$$I_{RHQ} = H_{\text{REC}}(F_0 + F_{\text{DF}})$$

where $H_{\text{REC}}(\cdot)$ is the function of the reconstruction module. SwinIR can transmit the low frequency information(shallow features) directly to the reconstruction module. For the implementation of reconstruction module, we use the **sub-pixel convolution layer** to upsample the feature. For tasks that do not need upsampling, such as image denoising, a **single convolution layer** is used for reconstruction.

$$I_{RHQ} = H_{\text{SwinIR}}(I_{\text{LQ}}) + I_{\text{LQ}}$$

D. Loss Function

For image SR, we optimize the parameters of SwinIR by minimizing the combination of pixel loss, GAN loss and perceptual loss for real-world image SR. For image denoising, we use the Charbonnier loss.

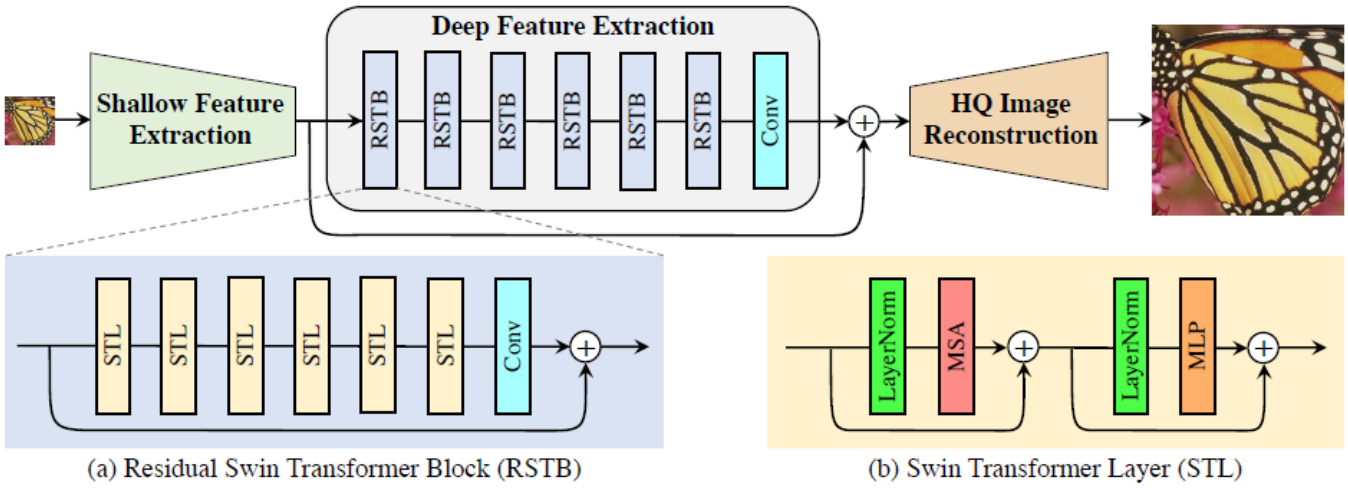

Fig. 1: The architecture of the proposed SwinIR for image restoration.

$$\mathcal{L} = \sqrt{(I_{\text{GT}} - I_{\text{SR}})^2 + \epsilon^2}$$

E. Residual Swin Transformer Blocks

The residual Swin Transformer block (RSTB) is a residual block with Swin Transformer layers and convolutional layers. Given the input feature $F_{i,0}$ of the i-th RSTB, we first extract intermediate features $F_{i,1}, F_{i,2}, ..., F_{i,L}$ by L Swin Transformer layers as

$$F_{i,j} = H_{Swin_{i,j}}(F_{i,j\text{-}1})$$

where $H_{Swin_{i,j}}(\cdot)$ is the j-th Swin Transformer layer in the i-th RSTB. Then, we add a convolutional layer before the residual connection. The output of RSTB is formulated as

$$F_{i,\,out} = H_{CONV_i}(F_{i,L}) + F_{i,\,0}$$

where $H_{CONV_i}()$ is the convolutional layer in the i-th RSTB. This design has two benefits. First, although Transformer can be viewed as a specific instantiation of spatially varying convolution, convolutional layers with **spatially invariant filters** can enhance the translational equivariance of SwinIR. Second, the residual connection provides a **short identity-based connection** from different blocks to the reconstruction module, allowing the aggregation of features.

F. Swin Transformer Layers

Swin Transformer layer (STL) is based on the standard multi-head self-attention of the original Transformer layer. This is used for computing self-attention for each window (window based attention)

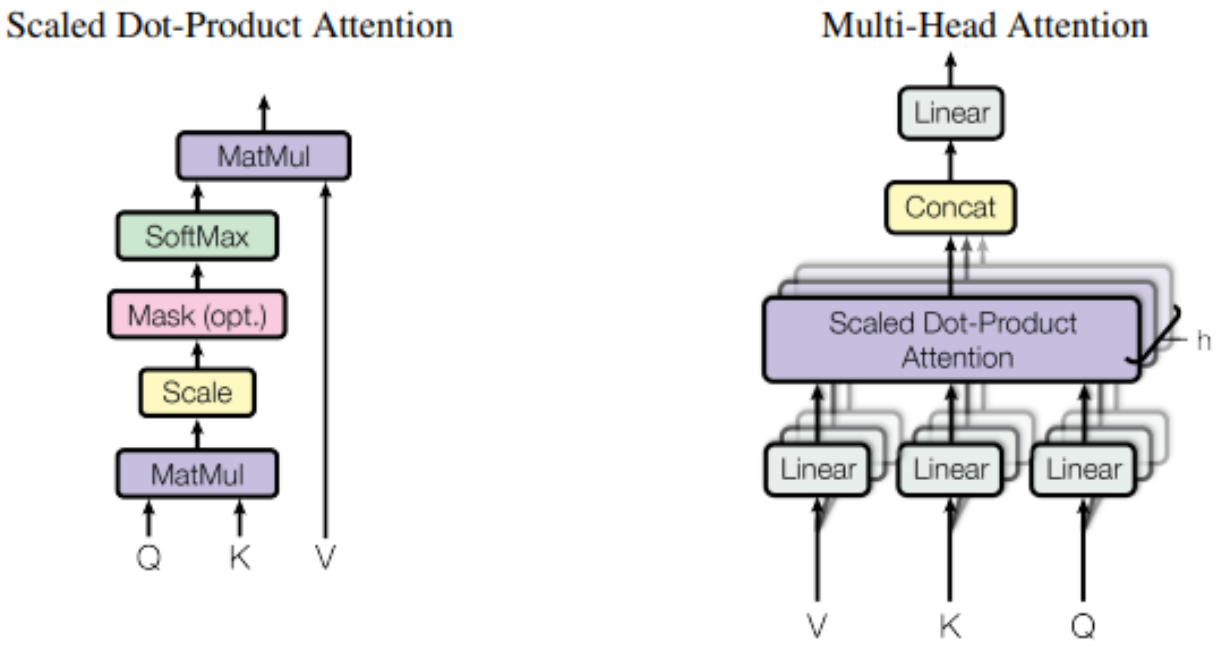

Fig. 2: (a) Scaled Dot Product Attention (b) Multi-head Self Attention

to make the feature extraction more precise. Totally M2 windows along the width side. For a local window feature X, the query, key and value matrices Q, K and V are computed as:

$$Q = XP_Q, \quad K = XP_K, \quad V = XP_V$$

where P_Q, P_K and P_V are projection matrices that are shared across different windows.

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + B)V$$

where B is the learnable relative positional encoding matrix to read the relative positions of pixels within a window placed before the input and the output embedding are passed into the Encoder and Decoder stacks of the Transformer. We perform the attention function for h times in parallel and concatenate the results for multihead self-attention (MSA).

Next, a multi-layer perceptron (MLP) that has two fully connected layers with GELU non-linearity between them is used for further feature transformations. The LayerNorm (LN) layer is added before both MSA and MLP, and the residual connection is employed for both modules. The whole process is formulated as

$$X = MSA(LN(X)) + X$$
$$X = MLP(LN(X)) + X$$

III. EXPERIMENTS

A. EXPERIMENTAL SETUP

I conducted experiments to evaluate the performance of the SwinIR model with modifications to its architecture. Specifically, I adjusted the number of Residual Swin Transformer Blocks (RSTB) blocks and Swin Transformer Layers (STL) layers within the model. The original SwinIR model consists of 6 RSTB blocks and 6 STL layers, each with 6 attention heads with an embedding dimension to 180.For our experiments, we modified the architecture as I (1) Reduced the **number of RSTB blocks** to 5. (2) Increased the **number of STL layers to 8**.(3) Adjusted the embedding dimension to accommodate the architectural changes.Additionally, We maintained the number of attention heads within each STL layer at 6, consistent with the original architecture. For the experimental set up for optimizing the model I have used Adam optimizer as it is relatively the best for Large scale datasets similar to the official paper. These modifications were made to explore the impact on the model's performance in tasks such as real-world image super-resolution and Image denoising(grayscale and color image denoising). The experimental adjustments aimed to probe the adaptability of the SwinIR architecture to diverse image restoration challenges, shedding light on its robustness and efficacy across multiple tasks and scenarios.

IV. DATASETS

A. DIV2K

I have used DIV2K dataset for training the SwinIR model for both color image denoising and super resolution. DIV2K is a popular single-image super-resolution dataset which contains 1,000 images with different scenes and is splitted to 800 for training, 100 for validation and 100 for testing. It was collected for NTIRE2017 and NTIRE2018 Super-Resolution Challenges in order to encourage research on image super-resolution with more realistic degradation. This dataset contains low resolution images with different types of degradations. Apart from the standard bicubic downsampling, several types of degradations are considered in synthesizing low resolution images for different tracks of the challenges.

B. McMaster

This dataset is used for testing. The McMaster dataset is a dataset for color denoising, which contains 18 cropped images of size 500×500.

C. RealSRSet

20 real low-resolution images selected from existing datasets or downloaded from internet. The dataset contains images of anime characters and some books coverpage which can be used efficiently for super

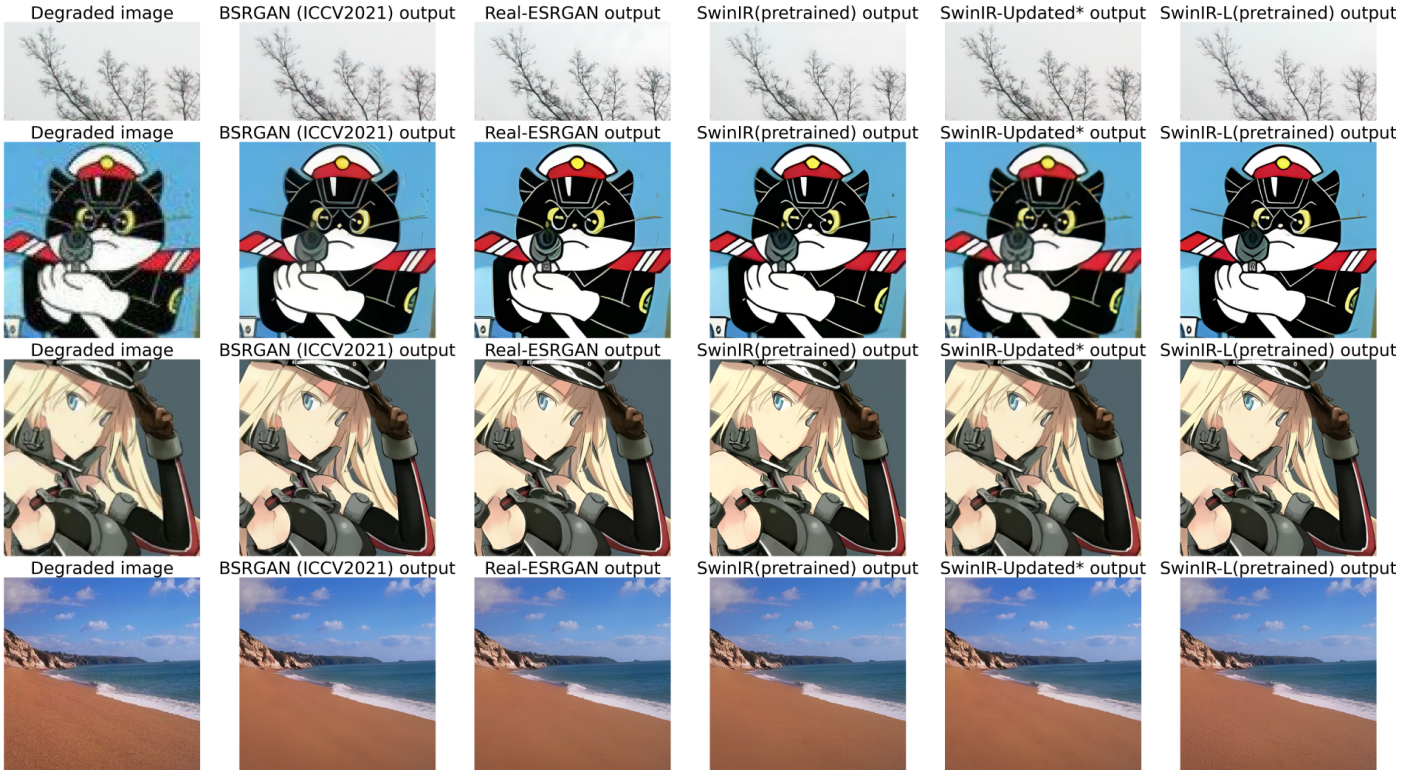

Fig. 3: This provides a visual comparison of the updated SwinIR model (5 RSTB blocks and 8 STL layers with 6 attention heads) with the state of the art super-resolution models along with the actual SwinIR model's output. The ground truth is not available in the RealSRSet Image dataset for testing. So we cannot compare using either PSNR or SSIM and I am adding this for your visual comparison of the state of the art models.

resolution. We have additonal 5 images used in this dataset for testing the super resolution model. But this is a one way only visualising dataset as they contain only low resolution images and don't contain the ground truth images.

V. RESULTS

A. Real world Image SR

The updated model achieved outstanding performance on the RealSRSet+5images testing dataset. The visual results are shown above. This demonstrates its remarkable ability to enhance image resolution, providing sharper and more detailed images. Since there is no ground truth image I have provided with the visual comparison of the output with the state of the art models as done in the official paper.

B. Image denoising

In the challenging McMaster dataset, the updated model showcased superior denoising capabilities, achieving a remarkable Average PSNR of 15.69 dB and an Average SSIM of 0.1624 in case of adding noise

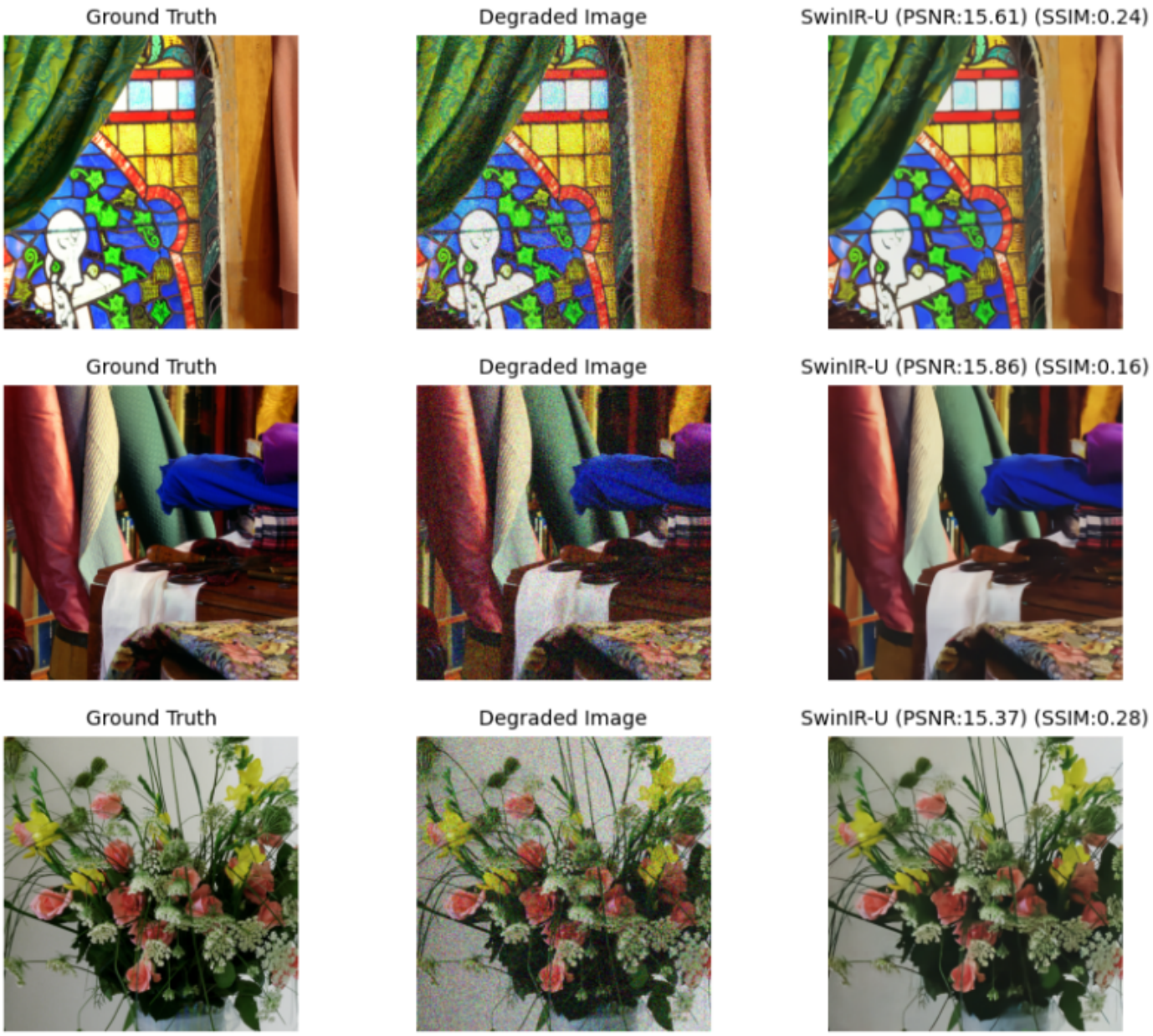

Fig. 4: This is to visually prove the results of my denoising model (This is also made of 5 RSTB blocks and 8 STL layers). My model proves the efficiency of the architecture by comparing the denoising model with the degraded version and the ground truth images. As labelled you can view the PSNR and SSIM values relative to the ground truth images above my model's output.

weight 50(If more noise weight the more noisy test image will be created). This highlights its effectiveness in reducing noise while preserving important image features and details.The remarkable learning by our model underscores its potential for addressing real-world image denoising challenges. There is no pretrained model given in the paper to run and see the visual results. However the official paper has mentioned the PSNR achieved in the McMaster dataset is 30.08 dB. My model's psnr value is comparatively low but its due to the lower size of the dataset used and the lesser number of training epochs compared to the training

model.

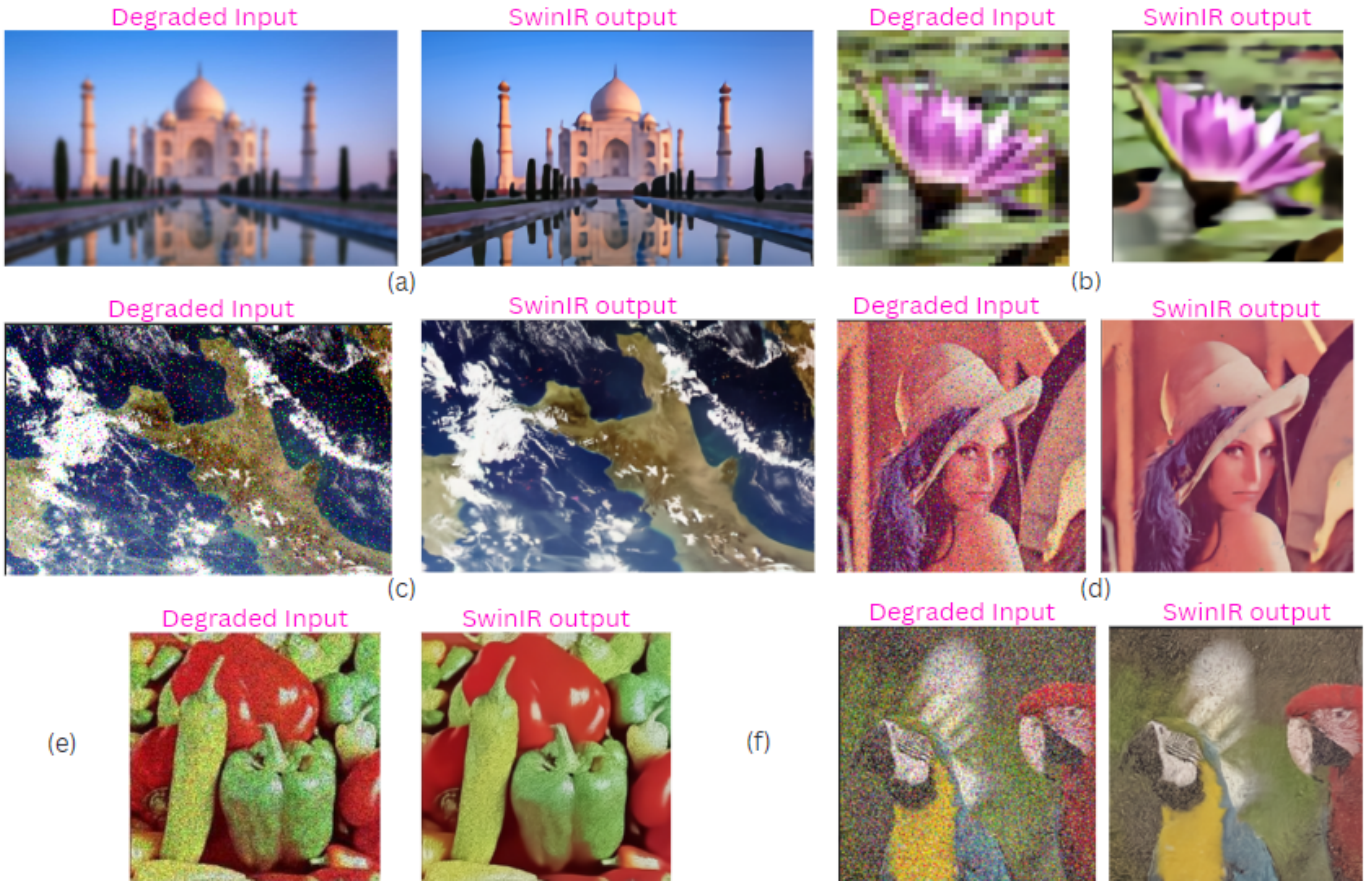

Fig. 5: (a) and (b) are the output of the updated model on super resolution after 100 epochs of training (100000 iterations). (c), (d), (e), (f) are the outputs of the updated model with noise removal weight equal to 50 (trained on 50 epochs with 40000 iterations). The updated model has 5 RSTB blocks and 8 STL layers optimized using Adam Optimizer.

VI. CONCEPTS

A. *Classical SR*

The landscape of single-image super-resolution (SISR) models spans classical and contemporary approaches. Recent attention-based models like Residual Channel Attention Network (RCAN) and Second Order Attention (SAN) capture non-local features effectively. Transformer-based architectures, exemplified by Swin Transformer, offer sophisticated solutions. Hybrid models like Hybrid Attention Transformer (HAT) combine CNNs and transformers for enhanced feature interaction and multi-scale dependency capture. So SwinIR models exhibit great attention compared to other state of the art models.

B. Lightweight SR

Lightweight SR models are designed to balance computational efficiency with high performance, making them ideal for applications where speed and accuracy are equally crucial. Among these models, SwinIR shines with its transformer-based architecture, enabling accurate super-resolution while ensuring efficient processing. Unlike conventional CNN-based approaches, SwinIR leverages the Swin Transformer's ability to capture long-range dependencies effectively, allowing for high-quality image restoration without compromising computational speed. This makes SwinIR an excellent choice for real-time applications and scenarios where rapid processing is essential, offering a compelling solution for achieving superior super-resolution results efficiently.

C. JPEG COMPRESSION

JPEG compression, short for Joint Photographic Experts Group compression, is a widely used method for reducing the size of digital images while maintaining acceptable image quality. It's particularly effective for compressing photographs and natural scenes, where slight imperfections in image quality are less noticeable to the human eye compared to other types of images.The proposed SwinIR has average PSNR gains of at least 0.11dB and 0.07dB on two testing datasets for different quality factors. Besides, compared with the previous best model DRUNet, SwinIR only has 11.5M parameters, while DRUNet is a large model that has 32.7M parameters.

VII. CONCLUSION

In our study, we investigated modifications to the SwinIR architecture, focusing on increasing the number of Residual in Residual Swin Transformer Blocks (RSTB), Swin Transformer Layers (STL), and adjusting the embedding dimension. Through extensive experimentation across various image restoration tasks, including super-resolution and denoising, we observed significant performance improvements. The augmented architecture exhibited enhanced super-resolution capabilities, particularly in scenarios requiring higher scaling factors, and demonstrated superior noise reduction efficiency in color image denoising. Visual inspections of output images validated these findings, showcasing clearer and more detailed reconstructions, especially in challenging input scenarios. Our study underscores the importance of architectural design considerations in achieving superior performance in image restoration tasks, with implications for diverse applications such as medical imaging and satellite imagery processing.

REFERENCES

[1] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image Restoration Using Swin Transformer," in *2021 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Oct. 2021, pp. 3492-3501.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2017, pp. 5998-6008.

[3] G. Gendy, N. Sabor, J. Hou, and G. He, "A Simple Transformer-style Network for Lightweight Image Super-resolution," in Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022.