Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/154/2025.TJ23130

SwinUnetSR: a Transformer-based Encoder-Decoder Structure with a Lightweight Upsampler for Face Super Resolution

Jiahao Song

Department of Computer and Science, Boston University, Boston, USA
jhsong97@outlook.com

Abstract: Face Super-Resolution (FSR) is a critical task aimed at enhancing low-resolution (LR) face images into high-resolution (HR) ones while preserving essential facial features. This task has broad applications, including identity recognition in surveillance systems and facial detail restoration for biometric authentication. In this paper, a novel hybrid architecture, SwinUnetSR, is proposed for FSR tasks. Built on the SwinV2-B transformer, the model integrates it within a Convolutional Networks for Biomedical Image Segmentation (U-Net) framework, followed by a lightweight upsampler for HR image reconstruction. The encoder, based on SwinV2-B, leverages hierarchical attention mechanisms to efficiently process global contextual information and down-sample facial features. U-Net serves as the decoder, where skip connections fuse the compressed features with SwinV2-B outputs. A lightweight upsampler then upscales the feature maps into HR images. Experimental results demonstrate that SwinUnetSR achieves high Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) scores, indicating its effectiveness. While computational resource limitations and data scarcity prevent it from outperforming state-of-the-art models, the high evaluation scores confirm the feasibility of SwinUnetSR for FSR. Furthermore, the model holds promise for future expansion to more complex scenarios. Code is available at: https://github.com/KbKuuhaku/swin-unet-sr.

Keywords: Face Super-Resolution, U-Net, Peak Signal-to-Noise Ratio, Structural Similarity.

1. Introduction

Tracing back to the beginning of the 21st century, Baker first proposed a traditional Face Super-Resolution (FSR) method, face hallucinating, to enhance the quality of low-resolution (LR) images of faces captured by surveillance cameras [1]. Since then, other researchers have improved upon Baker's method with traditional machine learning, such as Principal Component Analysis (PCA), Eigen transformation, and some vanilla neural networks learning local feature representations of faces [2-5]. However, due to the limitations of traditional machine learning, these methods could only reach excellent performance in the experimental phase but are unable to satisfy the high-accuracy demands of the industry.

Around 2012, with the rapid growth of deep learning techniques, researchers developed various deep learning models and conducted experiments on SR to make improvements. In the domain of FSR, Zhou et.al presented a Bi-channel Convolutional Neural Network (Bi-channel CNN) and greatly

DOI: 10.54254/2755-2721/154/2025.TJ23130

surpassed results from previous works [6]. Unfortunately, Bi-channel CNN was still unable to capture the high-frequency details of facial representations. Building upon the vanilla CNN structure, Goodfellow et al. proposed Generative Adversarial Networks (GANs) for better image generation [7]. GAN has also been applied to SR. Super-Resolution GAN (SRGAN) along with the perceptual loss could recover LR images into photo-realistic ones, which contain more natural textures and fewer artifacts than the ones generated from vanilla CNN-based models [8].

Nevertheless, CNN-based models, including GAN, all suffer from capturing global dependencies among features due to the core mechanism of the sliding kernel. Transformer, which is prevalent in the field of Natural Language Processing (NLP), is able to solve this problem. Transformer utilizes self-attention modules to capture the contextual relations of the whole sequence. Meanwhile, some researchers attempted to apply the Transformer framework to computer vision. The challenge is that images cannot be directly learned by the Transformer and need to be converted into vector form. To tackle this challenge, Dosovitskiy et al. introduced the Vision Transformer (ViT) [9]. In ViT, an image is split into multiple patches (similar to tokens in NLP), then gets mapped into a vector representation and thus could be learned by the Transformer. ViT performs well on image classification and other simple computer vision tasks but struggles to deal with complex tasks such as object detection, semantic segmentation, etc. Furthermore, the quadratic time complexity of attention modules makes the training process much slower than the one of the CNN models. Spotting the weakness of ViT, Ze et al. presented the Swin Transformer, which speeds up the computation of self-attention by leveraging shifted windows [10]. By computing self-attention inside a local window and alternating two different window partitioning configurations between consecutive blocks, Swin Transformer decreases the latency and outperforms ViT and its variants on recognition tasks [11, 12].

In this paper, the primary goal is to evaluate the feasibility of using the Swin Transformer for FSR. A novel encoder-decoder architecture, named SwinUnetSR, is proposed, incorporating a lightweight upsampler. The encoder, based on the SwinV2-B backbone, performs image down-sampling while extracting high-frequency features in a global context. The decoder, inspired by the U-Net structure, is tasked with up-sampling the compressed features from the encoder and facilitating a connection between the encoder and decoder through skip connections. Finally, a lightweight upsampler reconstructs high-resolution (HR) images from the features recovered by the Convolutional Networks for Biomedical Image Segmentation (U-Net) decoder. Compared to other well-established models, such as Image Restoration Using Swin Transformer (SwinIR), SwinUnetSR demonstrates superior computational efficiency, requiring fewer resources while maintaining competitive performance. The overall pipeline of SwinUnetSR is illustrated in Figure 1.

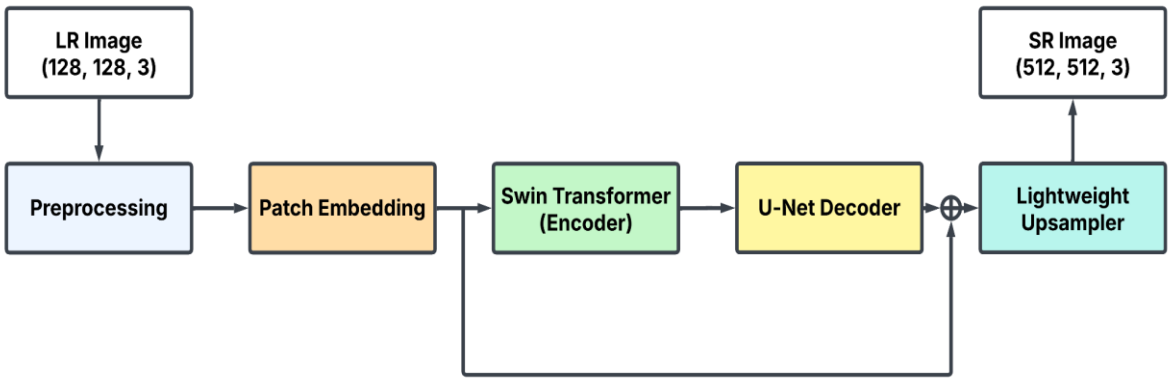

Figure 1: SwinUnetSR pipeline (picture credit: original)

DOI: 10.54254/2755-2721/154/2025.TJ23130

2. Methodology

2.1. Dataset description and preprocessing

The dataset is called Flickr-Faces-High-Quality (FFHQ) [13]. FFHQ was originally crawled from Flickr, where nearly 10 billion photos are hosted. The crawled images were first pruned by several filters, then aligned and cropped into RGB images of 1024x1024 resolution. In this study, a subset of the FFHQ dataset is used for training and testing. This subset can be found on Kaggle and consists of 52,000 out of 70,000 images. Images are down-sampled from 1024x1024 to 512x512 and thus it requires less computational power and time on training. Although it is a subset of FFHQ, it still covers a wide range of ages, ethnicities, and image backgrounds (Figure 2).

Figure 2: Samples of FFHQ (3x10)

There are mainly two parts in this dataset. The first part is for training and evaluation, which has 30,000 images in total. Specifically, the training, validation, and testing sets are further separated in proportion to 80%, 10%, and 10% respectively. This ensures the model does not overfit on the training set. The second part includes 22,000 images, which are used to test the model on unseen images. This will further prove that the model does not overfit on the validation set.

Training SwinUnetSR requires both LR and HR images. FFHQ provides HR images but LR ones are required to be constructed with traditional algorithms. Specifically, these LR images (128x128) are down-sampled from HR images (512x512) using bicubic interpolation. Then, before sending the data to the model, LR and HR images are rescaled to the range of 0 to 1. Finally, since SwinV2-B is pre-trained on the ImageNet dataset, it is better to normalize LR images with the mean and standard deviation from ImageNet [14].

2.2. Proposed approach

This paper introduces SwinUnetSR to test the feasibility of integrating Swin Transformer into the U-Net framework. SwinUnetSR mainly consists of three parts: Swin Transformer V2 (encoder), U-Net (decoder), and a lightweight upsampler. Swin Transformer V2 is a transformer-based model pre-trained on ImageNet and performs well on general image tasks. The downsampling module at the end of each stage of the Swin Transformer fits well on the U-Net structure. By replacing Residual Network (ResNet) blocks with the Swin Transformer, the U-Net structure could learn globally contextual representations of images during encoding. However, the challenge of applying the Swin Transformer directly into the U-Net is that there is a patch embedding at the start of the Swin Transformer, which reduces the image size. Therefore, features extracted from the U-Net decoder still need an upsampler to restore back to the original size of the image. In addition, since this is an image SR task, the upsampler will also include the x4 upscaling to get the SR image. The detailed architecture of SwinUnetSR is shown below (Figure 3).

DOI: 10.54254/2755-2721/154/2025.TJ23130

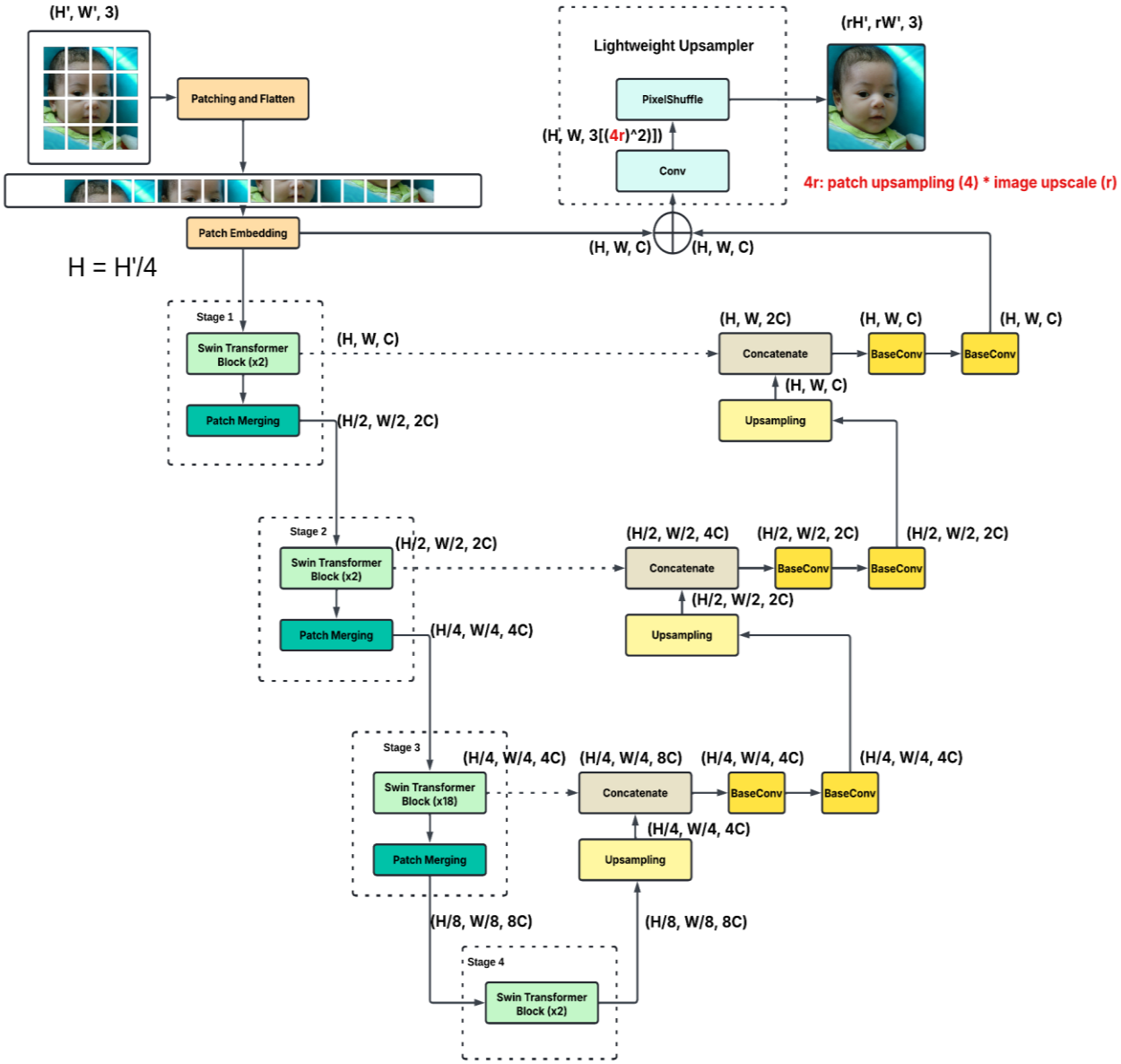

Figure 3: SwinUnetSR (picture credit: original)

2.2.1. Patch embedding

A naive way of passing images into the Transformer is to construct pixel embeddings. However, if pixels are treated the same as language tokens in NLP, an LR image, 128x128 for example, will have 16,384 tokens. On top of that, self-attention, the core module of the Transformer, has a time complexity of $O(T^2)$ for each image, where T is the total number of tokens, making it nearly impossible to train the Transformer on raw pixels. In order to reduce the time complexity, ViT splits the image into patches (Figure 4). Flattened patches are then linearly projected as patch embeddings. Thus, the overall time complexity reduces from $O(T^2)$ to $O[(\frac{T}{P^2})^2]$, where P is the patch size.

DOI: 10.54254/2755-2721/154/2025.TJ23130

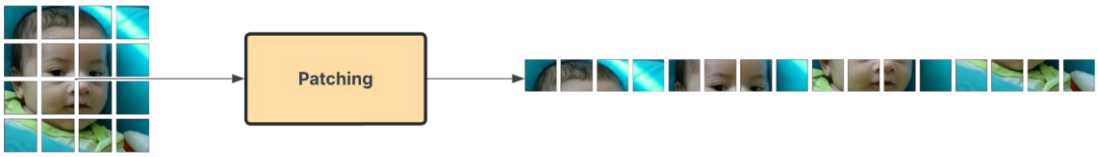

Figure 4: Image patching (picture credit: original)

2.2.2. Swin transformer

Swin Transformer uses a shift-window mechanism in self-attention and reduces the time complexity of self-attention into linear. Compared with ViT, it is more computationally efficient to use the Swin Transformer as an encoder to down-sample images and extract features. Moreover, the author of the Swin Transformer has released the pre-trained checkpoints for both V1 and V2. Fine-tuning a pre-trained Swin Transformer to adapt to the FSR task requires less data than training from scratch.

In SwinUnetSR, the architecture of the Swin Transformer can be divided into two parts. The first part is patch embedding. Images are split into 4x4 patches, which are then linearly projected as patch embedding. The other part is about the staging (Figure 5). At the start of each stage, there are multiple Swin Transformer blocks. At each block (shown in the left part of Figure 5), features are first normalized using Layer Normalization. Then, Window-multi-head self attention (W-MSA) or Shifted Window-MSA (SW-MSA), which alternates between consecutive blocks, gets executed to efficiently extract globally contextual features. The output features from MSA are then combined with the input of the Swin Transformer block using residual connections. Finally, after going through multiple blocks, there is a patch merging layer, which concatenates neighbor patches and down-samples the concatenated result.

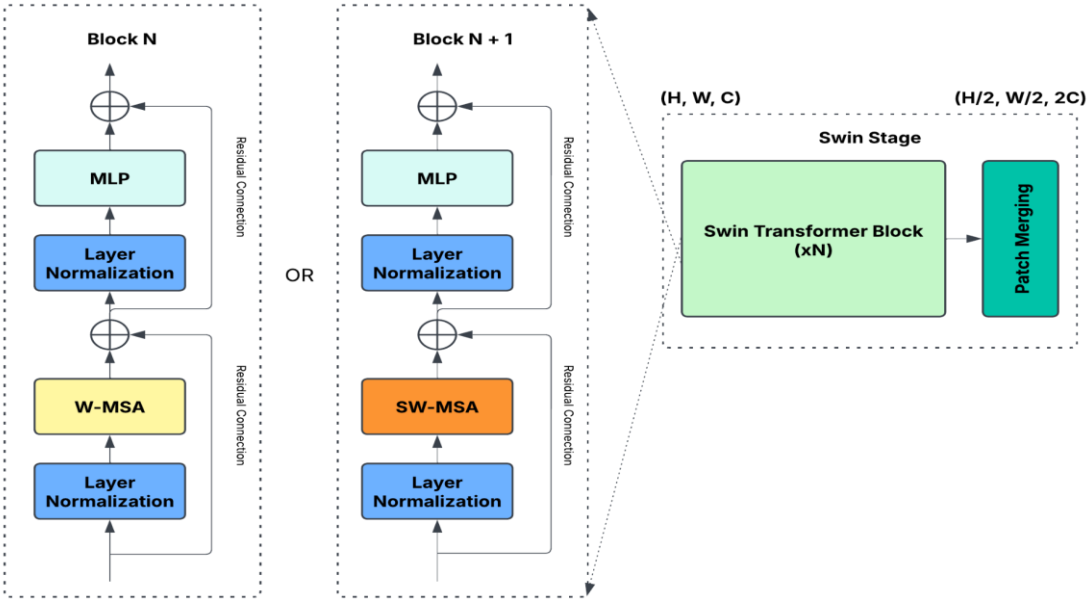

Figure 5: Swin stage (picture credit: original)

DOI: 10.54254/2755-2721/154/2025.TJ23130

2.2.3. U-Net and lightweight upsampler

U-Net is a U-shape encoder-decoder model. The decoder up-samples features and utilizes skip connections to associate output features with the ones from each layer of the encoder. As demonstrated in the right part of Figure 4, at each layer of the decoder, the output from the previous layer will first get up-sampled by the deconvolution. Then, the deconvolved results concatenate with features from the same level of encoder via skip connections. Finally, the output goes through two consecutive convolutional blocks (BaseConv), each with Rectified Linear Unit (ReLU) activation.

To upscale decoder outputs to the size of HR images, a lightweight upsampler is needed. The lightweight upsampler will first increase the number of channels from C to $(4r)^2C$, where r is the upscale factor. Then, the pixel shuffle is applied to rearrange elements such that the input gets upscaled by 4r, where 4r considers both the patch upsampling (4) and the image upscaling (r). Finally, at the end of the model, the output is denormalized to the range of 0 to 1 to align with the ground-truth images.

2.2.4. Loss function

L1 loss is one way of optimizing the FSR model. The pixel-wise L1 loss between LR images and HR images is as follows:

$$L = \|f_{SR}(I_{LR}) - I_{HR}\|_1 \tag{1}$$

where I_{LR} and I_{HR} are the scaled images of LR (input of SwinUnetSR) and HR (ground truth), and f_{SR} stands for the forward propagation of SwinUnetSR transforming LR images to SR images. L1 pixel loss computes the absolute difference of each pixel between SR and HR images. As a result, after minimizing the L1 loss, SR images will be close to HR images in terms of pixel value.

2.2.5. PSNR and SSIM

Peak Signal-to-Noise Ratio (PSNR) stands for Peak-Signal-to-Noise Ratio. It measures the fidelity of SR images compared with the original HR images in terms of noises. PSNR is in a logarithmic space so it is more sensitive to detailed differences between two images. PSNR (in dB) is computed with the mean of $PSNR_i$:

$$PSNR_i = 10\,log_{10}(\frac{MAX^2}{MSE_i}) \tag{2}$$

where MAX means the maximum range of pixel values (255) in an image, and MSE_i is the mean squared error between two images.

Structural Similarity (SSIM) measures the similarity between any two signals and is often computed locally when referring to the image similarity [15]. To compute local SSIMs, it starts with convolving images with a Gaussian Kernel whose size is 11 and whose standard deviation is 1.5. Then for each local SSIM, it can be expressed as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{3}$$

where x is the convolved result from image$x$ (same for$y$), $\mu_x$is the mean of image$x$,$\sigma_x$is the standard deviation of image$x$, and$\sigma_{xy}$is the correlation coefficient of image pair x and y. C_1andC_2 are constants. Finally, take the average of all $SSIM_j$ and can get the Mean of SSIM (MSSIM).

DOI: 10.54254/2755-2721/154/2025.TJ23130

3. Results and discussion

3.1. Results analysis

Experiments are conducted with SwinUnetSR under the setting shown in figures. During the training, L1 loss / PSNR / SSIM curves are recorded. To begin with, average L1 pixel loss is recorded with respect to iterations. As illustrated in Figure 6, L1 loss starts around 0.2 and drops down significantly during the first 2,000 iterations. The reason L1 loss starts with a small number is that SwinUnetSR is fine-tuned based on SwinV2-B, which has already learned a good representation of images in general. After the first 2,000 iterations, L1 loss decreases smoothly and approaches slowly to 0.02. Afterwards, there are some small improvements in L1 Loss but it stays above 0.02.

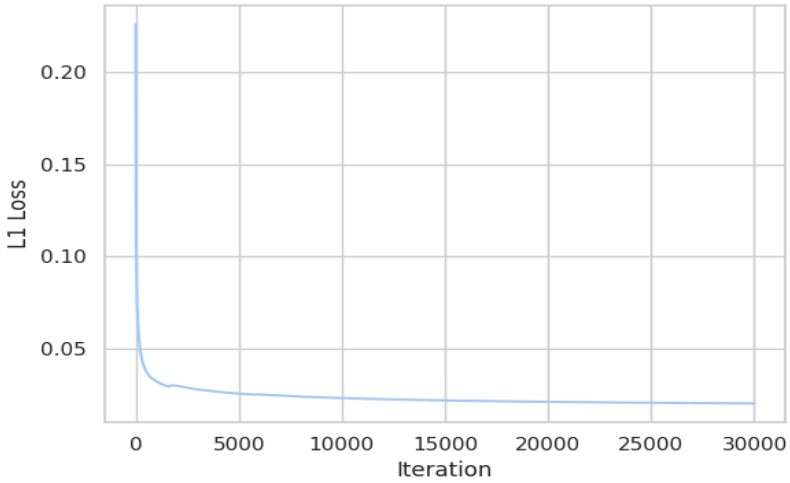

Figure 6: Average L1 loss over iterations (picture credit: original)

PSNR and SSIM (Figure 7) are evaluation metrics measuring the overall quality of SR images. Both of them are measured on the training set and the validation set. For PSNR, it fluctuates between 27 and 29 dB in the first 5 epochs. Then, the slope of the curve gradually becomes flat and reaches below 30 dB. SSIM shares the same trend with PSNR and stays under 0.84.

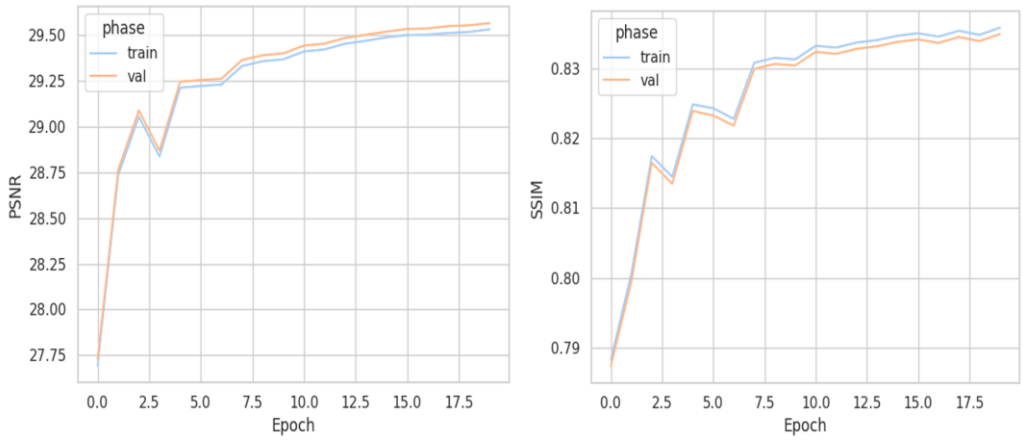

Figure 7: PSNR and SSIM over epochs (picture credit: original)

Additionally, there is no obvious gap between the curves of training and validation, which means the model does not overfit on the training set. Moreover, the model is tested on 22,000 unseen FFHQ

DOI: 10.54254/2755-2721/154/2025.TJ23130

images (Table 1), where PSNR reaches 29.51 and SSIM is over 0.84, Therefore, the model also does not overfit on the validation set.

Table 1: Evaluation results on unseen 22,000 FFHQ images

Number of images	PSNR	SSIM
22,000	29.51	0.84

3.2. Discussion

In summary, SwinUnetSR gets high PSNR and SSIM on FFHQ 512x512 (x4 scale). However, it does not surpass the boundary of 30 dB on PSNR. This might be caused by the downside of the dataset and model. First, due to the limitation of memory and Graphics Processing Unit (GPU) resources, only 30,000 images are used during the training. This might lead to incomplete learning and can be further improved by adding more data, such as CelebFaces Attributes High-Quality (CelebA-HQ). Second, the encoder of SwinUnetSR down-samples features after each stage. This reduces both the GPU computations and the memory burden caused by high-resolution features. At the same time, however, some high-frequency features might be lost during downsampling, leading to a limited supersampling in the end.

4. Conclusion

This paper introduces SwinUnetSR, a novel Swin-Transformer-based architecture designed for FSR tasks. The model integrates the strengths of SwinV2-B, U-Net, and a lightweight upsampler, maintaining an encoder-decoder structure while utilizing a separate upsampler for HR image reconstruction. The effectiveness of SwinUnetSR was demonstrated through fine-tuning SwinV2-B on a partial FFHQ dataset, achieving high PSNR and SSIM scores for super-resolved images. These results validate the feasibility of the proposed model in generating high-quality SR images. Importantly, SwinUnetSR is not limited to human faces; it shows promise for broader applications. Future work will expand the model's capabilities to more complex scenarios, including manga and game character images, further exploring its potential in diverse image synthesis tasks.

References

[1] *Baker, S., & Kanade, T. (2000). Hallucinating faces. In Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 580, 83-88.*

[2] *Ce, L., Heung, Y.S., & Zhang, C.S., (2001). A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, I-I.*

[3] *Lepcha, D. C., Goyal, B., Dogra, A. (2023). Image super-resolution: A comprehensive review, recent trends, challenges and applications. Information Fusion, 91, 230-260.*

[4] *Chang, H., Yeung, D.Y., & Xiong, Y.M., (2004). Super-resolution through neighbor embedding. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, I-I.*

[5] *Jiang, J., Hu, R., Wang, Z. (2014). Noise Robust Face Hallucination via Locality-Constrained Representation. IEEE Transactions on Multimedia, 16(5), 1268-1281.*

[6] *Zhou, E., Fan, H., Cao, Z., et al. (2015). Learning Face Hallucination in the Wild. Proceedings of the AAAI Conference on Artificial Intelligence, 29(1).*

[7] *Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2020). Generative adversarial networks. Commun. ACM, 63(11), 139–144.*

[8] *Zhou, C., Li, Q., Li, C., et al. (2024). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. International Journal of Machine Learning and Cybernetics, 1-65.*

[9] *Alexey, D., Lucas, B., Alexander, K., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In 9th International Conference on Learning Representations, 3-7.*

DOI: 10.54254/2755-2721/154/2025.TJ23130

[10] *Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In IEEE/CVF International Conference on Computer Vision, 9992-10002.*

[11] *Liu, Z., Hu, H., Lin, Y., et al. (2022). Swin Transformer V2: Scaling Up Capacity and Resolution. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11999-12009.*

[12] *Liang, J., Cao, J., Sun, G., et al. (2021). SwinIR: Image Restoration Using Swin Transformer. In IEEE/CVF International Conference on Computer Vision Workshops, 1833-1844.*

[13] *Karras, T., Laine, S., & Aila, T. (2021). A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(12), 4217-4228.*

[14] *Chang, Y., Wang, X., Wang, J., et al. (2024). A survey on evaluation of large language models. ACM transactions on intelligent systems and technology, 15(3), 1-45.*

[15] *Zuo, C., Qian, J., Feng, S., et al. (2022). Deep learning in optical metrology: a review. Light: Science & Applications, 11(1), 1-54.*