QwinSR: An All-MLP Shifted Window Model for Image Super Resolution

Mark Bauer
Grainger College of Engineering
University of Illinois Urbana-Champaign
markb5@illinois.edu

Quinn Ouyang*
College of Fine and Applied Arts
University of Illinois Urbana-Champaign
qouyang3@illinois.edu

Abstract

Famous for its shifted window representation, the Swin Transformer backbone architecture has consistently showcased state-of-the-art performance across a range of well-established computer vision benchmarks. To enhance computational efficiency, the Swin Mixer architecture adopts this structure but substitutes its attention layers with simpler multi-layer perceptrons (MLPs). In line with this design, we introduce QwinSR, an application of this all-MLP architecture tailored for single image super-resolution. This adaptation simplifies the Swin Transforer-based image restoration model, SwinIR. QwinSR leverages SMLs to extract essential features and subsequently aggregates them within a compact convolutional neural network, facilitating image reconstruction. We anticipate that our model will yield competitive accuracy-to-computation metrics, particularly when compared to SwinIR and other leading models in the field. Our (untested) code is available at https://github.com/quinnouyang/QwinSR.

1 Introduction

Despite advances in modern photography and image transmission technology, resolution loss is often an unavoidable or necessary compromise. This produces a need and desire for techniques to construct higher fidelity images from lower resolution sources, which we call super resolution (SR) imaging [3]. However, SR is a more niche category than other computer vision tasks (such as classification and semantic segmentation) [1], making this field ripe for new research. Given the simple objective and need for research on SR, we propose a simplified all-MLP model based on the shifted window design from the Swin Transformer by Liu et al. [2]: QwinIR.

1.1 Related Work

1.1.1 Single Image Super-Resolution

"Super-resolution" generally refers to the process of enhancing the visual detail and fidelity of an image by predicting pixels to increase its resolution [16]. In other literature, SR is roughly interchangeable with the more general terms "upsampling" and "reconstruction." Note that the latter implies that an exact / true higher resolution image exists for a lower resolution one, which typically comes from artificial downsampling (classic image SR) [2, 3]. Unlike the more common variant of this task, multiple image SR, "single image" specifically refers to a process that relies on only one source image rather than several. Effective multiple image SR techniques exploit the related images for more pixel information to construct from, which can apply in the real-world (e.g. a burst shot, video frames, etc.) [16, 17]. However, we focus on single image SR as it is a more fundamental and challenging task [17].

*"QwinSR" is a play on Swin using Quinn's name. This was Mark's idea.

Preprint. Under review.

Bicubic Interpolation SRResNet SRGAN Original

Figure 1: Visual comparisons of traditional and learning-based classic image SR approaches to a $\times 4$-upscaled original high resolution image.

Traditional algorithmic approaches for SR directly interpolate pixels from a lower resolution source, typically assuming a downsampling process to generalize the reconstruction. Popular basic algorithms include bicubic and nearest-neighbor interpolation which are fast and eschew the long training times associated with learning-based models, but the absence of trained priors obviously limits their ability to hallucinate new pixels [16, 17].

1.1.2 Learning-based Approaches

On the other hand, learning-based models spanning a variety of architectures have far surpassed the generalized interpolations that traditional approaches limit themselves to. Models based on convolutional neural networks (e.g. SRCNN, SRResNet), general adversarial networks (e.g. ESRGAN, SRGAN, Real-ESRGAN), and transformers (e.g. SwinIR) have all effectively competed for state-of-the-art performance [3].

For SR related tasks, CNNs are effective at global feature extraction but have lately been usurped by transformers, which tend to be more robust and lightweight. As a consequence, pure CNN models often require added complexity just to match the performance of transformer models [18]. It is quite effective to utilize hybrid CNN-transformer models, in which CNNs are used for global feature extraction while transformers focus on local feature extraction. This is demonstrated in the architecture of SwinIR [3].

GANs tend to be tedious to fine tune and also have many trainable parameters, making the training process very long [2]. Similarly, diffusion models like StableSR have gigantic datasets and are quite complex [19]. The focus of this research project is on simplicity, and as such we do not consider these models.

1.1.3 All-MLP

"All-MLP" describes a model as primarily relying on only multi-layer perceptrons (MLPs), often in contrast to neural networks or transformers. These models tend to have significantly fewer parameters and simpler architectures than other models, offering an ideal tradeoff between accuracy and training time / running time [4, 14]. Proposed by Tolstikhin et al., the MLP-Mixer is one of the first all-MLP demonstrations for computer vision. Despite its apparent computational limitations, MLP-Mixer has boasted results comparable to those from state-of-the-art implementations for image classification. Inspired by this work, we chose to follow the same all-MLP design philosophy and apply it on SR as a more high level benchmark.

1.1.4 Shifted Window

Swin Transformer: First popularized by Liu et al., the shifted window approach is the computational representation in a groundbreaking computer vision model: the Swin Transformer. Critically, it limits "self-attention computation to non-overlapping local windows while also allowing for cross-window connection," eliminating global self-attention and thus reducing from quadratic to linear computational complexity relative to image size [4]. Variants of the Swin Transformer have achieved

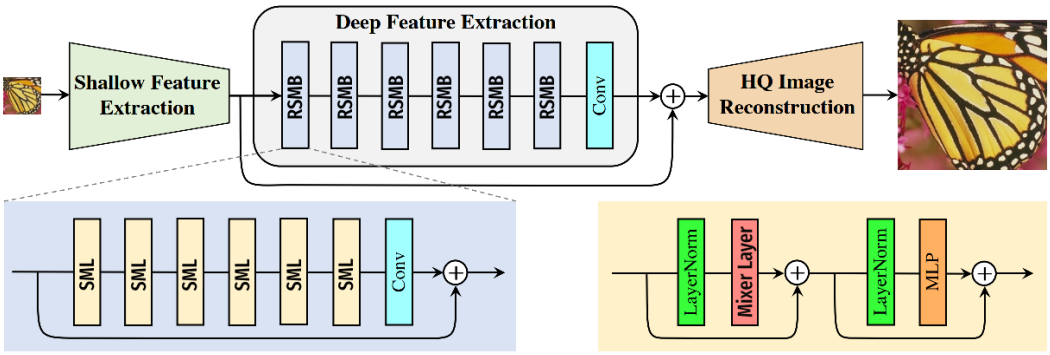

Figure 2: The architecture of QwinIR. Figure based off of that of SwinIR.

state-of-the-art results in a variety of common tasks, including image classification and semantic segmentation [3].

Swin-Mixer: The modularity of the shifted window technique in the Swin Transformer enables replacing its attention layers with MLPs, yielding the Swin-Mixer. Specifically, it replaces each of its multi-head attention layers with the all-MLP layers presented by Tolstikhin et al., converting the Swin Transformer Blocks into Swin Mixer Blocks. Swin-Mixer surpasses the already classification results of MLP-Mixer and other all-MLP models [4, 14], but little research exists on how it performs elsewhere, e.g. for SR.

SwinIR: A pioneering model for attention-based SR, SwinIR uses basic convolutional layers for shallow feature extraction and Swin Transformer blocks for deep feature extraction. These extracted features are then passed through a basic upsampling convolution that involves a convolutional layer and a pixel shuffle. SwinIR impressively surpasses other top-performing models based on CNNs, GANs, etc. in classical SR [3].

2 Model Architecture

The architecture of QwinIR is novel, in that it combines the model architectures of SwinIR and Swin-Mixer. Similar to SwinIR, our model involves a 3 step architecture: shallow feature extraction, deep feature extraction, and image reconstruction. We illustrate this in Figure 2.

2.1 Shallow Feature Extraction

QwinIR uses the exact same method for shallow feature extraction as SwinIR. That is, given a low quality input image, we use a 3×3 convolutional layer to perform shallow feature extraction. This step is taken because convolutional layers are quite good at local feature extraction, making them useful in the early stages of image processing. The authors of the SwinIR paper note that the use of this layer not only leads to more stable optimization and better results, but it also provides a simple way of mapping input images to higher dimensional feature spaces [**?**].

2.2 Deep Feature Extraction

Next, we perform deep feature extraction by passing the extracted shallow features through a series of residual Swin mixer blocks (RSMBs) and a 3×3 convolutional layer (these blocks are called "residual" because they each contain a skip connection, where the results of each block are aggregated with the input features passed to the block). Each RSMB contains a series of Swin mixer layers (SMLs) followed by a 3×3 convolutional layer, and each Swin mixer layer is simply a Swin Transformer layer that has had its multi stage attention head replaced with a mixer layer, as we depict in Figure 2.2 [4].

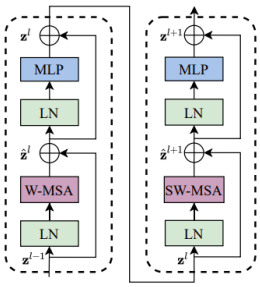

Figure 3: Two original successive Swin Transformer blocks with attention heads included (W-MSA and SW-MSA). Swin Mixer blocks replace each of these with a Mixer Layer.

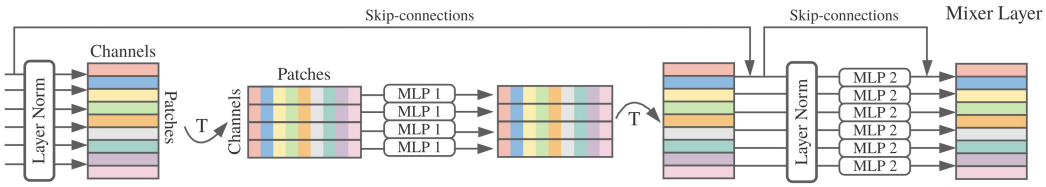

Figure 4: Mixer layer from MLP-Mixer, consisting of token-mixing MLP, channel-mixing MLP, skip-connections, dropout, and layer normalization

Mixer Layer: Each mixer layer is composed of 2 MLPs, with two fully connected layers and GELU nonlinearities. Mixer layers contain one channel-mixing MLP and one token-mixing MLP. The channel-mixing MLPs come first, which allow communication between different input channels and operate on each input token independently. The results of the channel-mixing MLPs are passed to the token-mixing MLPs, which allow communication between different tokens and operate on each channel independently [14]. By utilizing this mixing architecture, features are able to be extracted without the explicit need for attention. Hence, it makes logical sense for deep feature extraction to be accurate even when attention heads in the Swin Transformer blocks are replaced with mixer layers. However, due to the removal of attention heads, these blocks are technically no longer transformers, and hence we do not refer to them as such.

2.3 Reconstruction

Using a skip connection, the originally extracted shallow features and the newly extracted deep features are aggregated and passed into an image reconstruction module. By using a skip connection, we can ensure that high frequency features are not lost during the reconstruction process [**?**]. For image reconstruction, we implement a sub-pixel convolutional neural network (ESPCN), the same technique utilized by SwinIR [7]. The ESPCN consists of two convolution layers for feature extraction, and a sub-pixel convolution layer that aggregates the low resolution features. From here, the ESPCN then constructs the high resolution image in a single step.

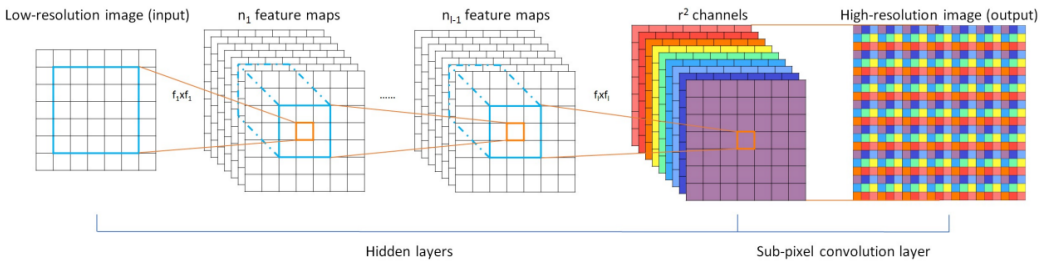

Figure 5: The network architecture for ESPCN

3 Experiments

Unfortunately, we were unable to conduct a formal experimental procedure due to resource and time constraints However, barring those factors, our experimental phase would proceed as follows: for our experimental procedure, we intend to test the ability of our model to perform image upscaling. Given that our model is derived from the SwinIR architecture, it is logical for our experiments to replicate those performed in the SwinIR paper. It should be noted that in the original SwinIR paper, the experimental procedure consists of image super resolution, image denoising, and JPEG compression artifact reduction. Since our architecture focuses only on super resolution, we ignore the other 2 experimental components.

Ablation Study: The first stage of our experimental procedure will be an ablation study. This study involves altering and removing certain components from our model to understand the contributions of those components to our results. In this stage, our model will be trained on the DIV2K [11] dataset and will be tested on the Manga109 [12] dataset. The first portion of this stage will observe the effects of changing the channel number, number of Swin-MLP blocks, and number of SMLs per block on model performance. To perform this step, we will test a variety of combinations for values of these hyperparameters and observe the relative correlations between the values of each hyperparameter and peak signal to noise ratio (PSNR) for the upscaled images. The second portion of this stage will observe the effects of changing patch size and the number of training images. To perform this step, we will test a variety of combinations for values of these hyperparameters and observe how each value affects PSNR and time for convergence. The third portion of this stage will observe the effects of the residual connection and convolutional layer in each SMB. Like in the SwinIR paper, we will test 4 residual connection variants in each block: no residual connection (i.e. no convolution operation), a 1×1 convolution layer, a 3×3 convolution layer, and three 3×3 convolution layers. For each of these hyperparameter values, the corresponding PSNRs will be assessed and compared.

Results In the second stage of our experimental procedure, we will select the optimal combination of hyperparameters as determined in the previous stage, and we will use them to test our model on the Set5 [8], Set14 [9], BSD100 [10], Urban100 [13], and Manga109 [12] datasets. The benchmarks we record in this stage will be PSNR (peak signal to noise ratio, lower indicates less grainy output images), SSIM (structural similarity index measure, which compares how similar an upscaled image is to the original high quality image a model seeks to reproduce), the number of model parameters, and the number of multiply-accumulate operations. These measurements will be compared to the benchmarks recorded in the SwinIR paper (this is convenient, because the SwinIR paper contains benchmarks for numerous SOTA architectures, not just for the SwinIR architecture). By performing this comparison, we will gain conclusive evidence as to how our model performs relative to state of the art models–in particular, if our model has a viable accuracy/cost ratio.

4 Conclusion

In this paper we propose QwinIR, an image restoration model that combines the architectures of SwinIR and Swin-Mixer. Like SwinIR, this model consists of three phases: shallow feature extraction, deep feature extraction, and image reconstruction. Unlike SwinIR, our model uses residual Swin-Mixer blocks for deep feature extraction, where each block consists of Swin-Mixer layers, a convolution layer, and a residual connection. While we were unable to conduct a formal experimental phase due to resource and time constraints, we believe that this model has potential to perform at a similar level to state of the art image upscaling models. For future work, we would like to conduct a formal testing phase on a more substantial cluster of GPUS, as GCP and our personal computers proved to be insufficient for training our model. Furthermore, we would like to experiment with the placement of skip connections within our network to potentially remove the need for an initial convolutional layer for shallow feature extraction, similar to the SwinSeg architecture described by Zhang et al. [6].

Acknowledgments and Disclosure of Funding

We would like to acknowledge Paris Smaragdis, Krish Subramani (please be gentle to us), Evan Matthews, Nicolas Prate, and Quinn's sister's cat, Sesame.

We have no funding.

References

[1] Papers with code: Computer vision. URL https://paperswithcode.com/area/computer-vision.

[2] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[3] S. C. Park, M. K. Park, and M. G. Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20(3):21–36, 2003. doi: 10.1109/MSP.2003.1203207.