Low Rank Adaptation in Deep Networks for Image Super Resolution

Gugan Kathiresan
Khoury College of Computer Sciences
Northeastern University
Boston, USA
kathiresan.g@northeastern.edu

Anirudh Muthuswamy
Khoury College of Computer Sciences
Northeastern University
Boston, USA
muthuswamy.a@northeastern.edu

Aditya Varshney
Khoury College of Computer Sciences
Northeastern University
Boston, USA
varshney.ad@northeastern.edu

***Abstract*—The task of super resolution or Single Image Super Resolution (SSIR) has been a popular choice in computer vision research for a long time. As newer, complex and more efficient models are developed, the goal of achieving SSIR remains persistent. The recent class of generative and attention based models have proven to be effective at perform SSIR. However, models like transformers, Generative Adversarial Networks (GANs) and Diffusion networks are highly dependent on the distribution of the training data and often do not provide realistic generations in newer domains. Fine-tuning these models on the new data sets are computationally heavy tasks. Further, it still not guaranteed to yield satisfactory results if not trained for large periods of time on large data sets. The popular Low Rank Adaptation (LoRA) approach has picked up in recent works to address the computational costs of Large Language Models (LLMs). In this study, we apply LoRA to popular deep networks that aim to achieve single image super resolution when moving from different data set domains.**

***Index Terms*—Super Resolution, Computer Vision, Deep Learning, Diffusion, Transformers, GAN, Low Rank Adaptation**

I. Introduction

In 2015, Dong et al. [1] presented a pioneering study that captured the attention of the computer vision field. Their work centered on utilizing deep learning techniques to enhance individual images through super-resolution methods. Specifically, they aimed to employ deep convolutional networks for Single Image Super Resolution (SISR). Their motivation stemmed from the assertion, outlined in the paper's abstract, that traditional SISR methods based on examples could be viewed as akin to deep convolutional networks. Furthermore, their proposed Convolutional Neural Network (CNN) effectively managed patch extraction and fusion, a core aspect of conventional example-based methodologies.

Transformers, GANs and especially Diffusion models are notoriously computationally expensive to train. In most cases, when attempting to generate images from an unknown domain, the generated results are not satisfactory. However, both attempting fine-tuning or training separate models for each new data set is a computationally expensive task. Furthermore, performing fine-tuning does not always ensure that the newly learnt model has learnt enough of the new data distribution.

In the study by Hu et al. in 2021 [2], the concept of LoRA, a training strategy that uses rank decomposition to update weights based on the new training data. This reduces the computational time required to re-train the model on a new dataset. Until now, it has mostly been used on transformer based - large language tasks. We aim to apply this method on popular generative models in order to reduce the time required to fine-tune the model.

In this study, we explore the application of various attention and generative models like GANs, Transformers and Diffusion models that aim to achieve single image super resolution. As we explore the application, we identify its challenges and how well it is suited for the task. Further, each model is subjected to LoRA to evaluate its effectiveness. This effectiveness is also quanitified in how well it can adapt to newer domains.

II. Related Work

The use of generative models that can expand over large data sets and boast elite performance over traditional CNN architectures is not new to the research field. In this section, we detail specific works that employed models like Transformers, GANs and Diffusion models.

Transformer architectures for SISR. In the study by Lu et al. [3], the authors proposed the Efficient Super Resolution Transformer. The goal of this method is similarly to achieved SISR in the lowest computational complexity. The proposed model employs a combination of multiple transformer blocks that extract features on multiple patches of the input image with multi headed attention.

GAN architectures for SISR. In an attempt to employ GAN based architectures, Park et al [4] proposed a deep architecture that can efficiently include domain related features to improve realism in generation quality. By employing residual connections in the generator block, most of the significant latent features can be obtained. However, this study incorporated two discriminator architectures. One discriminator classically identified fake images from real ones as generated by the generator. The second discriminator focused on incorporating feature content from the ground truth to match the generated super resolution images to the current domain.

Diffusion architectures for SISR. Wang et al. in their study [5] introduced a novel framework for disentangling content and style in style transfer, s similar to the process of learning the style of the source domain in super-resolution

mapping. Unlike existing methods reliant on explicit or implicit techniques like Gram matrices or GANs, this approach extracted content explicitly and learns style implicitly, enhancing interpretability and control in disentangling content and style. By introducing a CLIP-based style disentanglement loss and leveraging diffusion models, the framework allowed the authors to control the features being picked up for the disentanglement. This could be helpful when trying to focus your diffusion model into generating realistic boundaries and features that exist in the source image, rather than generating non-existent ones.

Transfer Learning techniques. Large models require tremendous amount of compute to train on complex non-linear datasets. This becomes a challenging task for someone with limited compute resources. Instead of training a model from scratch, A pretrained model which has been trained on a similar task may be used as a baseline and further training is done on top of it. This results in lesser time taken for the new model to converge when compared to training it from scratch. Liu and Lang [14] proposed prefix-tuning as an alternative to fine-tuning. Instead of training all the layers in a model, they train only a prefix vector which holds domain specific knowledge.
Furthermore, Houlsby et al [15] propose Adapter Layers, which are independent layers that can be inserted in between existing set of layers. Once they are injected into the model structure and trained, they tend to learn data specific information which makes the model adapt to a new knowledge domain. However, these layers cannot be merged with existing layers, which leads to increase in parameters. Finally, Hu et al [2] present, Low Rank Adaptation(LoRA), a low rank matrix compute which is used to distinguish general and domain specific information. This results in an Low rank matrix which learns domain specific information and can be embedded into existing layers of a model. We discuss more about LoRA in section III-F. These approaches were initially proposed for Natural Language related tasks, but have been proven to be equally effective in Computer Vision as well.

Overall, the use of generative models has been of interest in recent computer vision works and its application towards image enhancement tasks has proven to be worth consideration. As the model becomes more efficient, and more computationally heavier, it opens up the need for research that facilitate easier transfer of knowledge and usability. This is where we beleive the concept of Low Rank Adaptation (LoRA) will help.

III. Methodology

All experiments are focused around the goal of acheiving 4x upscaled super resolution from low resolution inputs on a new domain. A good performing model is expected to take in low quality images from any new domain, able to fine tune to that domain in short time with low computational costs, and provide high quality super resolution outputs.

Fig. 1. Example Images from FloodNet, Set 5 and Set 14 dataset (clockwise)

A. Dataset Used

For this study we considered three data sets that facilitate our training and fine-tuning experiments. These datasets are:

1) FloodNet data set [6]: This data set was developed for a semantic segmentation and Visual Question Answering tasks. The images were captured from post-disaster scenarios, offers high-resolution UAS imagery along with detailed semantic annotations depicting damages. As discussed previously, the domain of aerial imagery is one that is highly affected by low resolution images. Applying super resolution can improve the usefulness of these images.
   In this study, we employed this data set with the same setting of high and low resolution pairs that we used for the DIV2K data set. Further, we retained the native split ratio of 80:20 for training and testing sets.
2) Set 5 data set [7]: a popular super resolution data set of just 5 images, used as a benchmark in many studies.
3) Set 14 data set [8]: a similarly popular benchmark data set for super resolution consisting of just 14 images.

B. Diffusion Models and Challenges

For this study, we explored the various diffusion models popularized by recent trends. Diffusion models were originally designed and are popularly used for text to image generation. However, workarounds have been developed that facilitate image to image generative tasks similar to super resolution. Simply put, the prompts are retained and the diffusion model learns the mapping between the low resolution and high resolution pairs during training.

However, the challenges lie in the quality of the generations. The diffusion models perform well in a limited domain, that is, the distribution of the data in its source domain. Any prompts to generate an image outside this source domain, ends up being a hallucinated, unrealistic image. In the case of super resolution this affects the quality of the generated higher resolution image. The features of the resultant high resolution image were generated, and did not exist in the original image.

In our experiments, we attempted to use the Latent Diffusion Upscaler Model (LDM Upscaler) proposed by Rombach et al. [9] in their 2022 study. Figure 2 depicts an inference example

Fig. 2. Ground truth and Inference of the LDM Upscaler depicted hallucinated outputs

of the LDM Upscaler model that was pretrained on popular real world image datasets like the COCO data set.

C. Transformer Models and Challenges

For our attempts with Transformer models, we employed a variation of the "Swin-Transformer" proposed by Liu et al. in their study [10]. The concept of the Swin-Transformer employs shifted windows to match the complexity of the self-attention scheme. This allows for the inclusion of lower scale domains like language and higher scale domains like imagery on a similar scale level. In addition, the swin transformer computes hierarchical feature maps across the depth of the layers, but applies the self attention only on the windows.

However, for the case of single image super resolution, the challenges we faced were with regards to the training of the Swin Transformer for super resolution. Reconstructing the original image in a higher resolution requires deconvolution steps for reconstruction. The overall architecture included feature extractors made up of multiple transformer blocks with residual connections. Since the resultant image is of a higher resolution, the training complexity increased significantly leading to longer training times. In addition, this lead to lossy reconstruction, resulting in low performance across metrics.

D. GAN Models and Challenges

GAN models are fundamentally adept for an image to image mapping task like super resolution. The architecture of GANs with generators to construct images from noise and discriminators to identify fake or low quality images can be varied to suit the task at hand. In this study we employ a variation of the popular Enhanced Super Resolution GAN (ESRGAN) introduced by Wang et al. in their 2018 study [11] and 2021 [12]. By controlling the amount of noise in the input image (the low resolution image of the pair), deep feature extraction, and upsampling to reconstruct the image in a higher resolution. As the discriminator learns the mapping between low resolution features and high resolution features, a general super resolution network is achieved that evaluated whether the generated image is relatively realistic as compared to the ground truth, rather than a binary decision of real or fake. The code we implemented is inspired by the work done

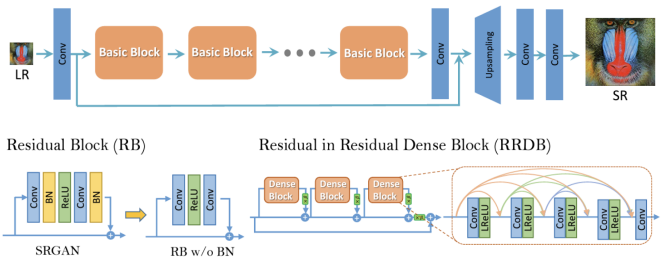

Fig. 3. Architecture of ESRGAN obtained from [12]

in [15, 16]. The architecture diagram of the ESRGAN can be seen in Figure 3, obtained from the works in [11].

E. Enhanced SRGAN

Ledig et al [13] proposed the first Generative approach to Super Resolution which led to further improvements to their structure. We analyse an enanced version of this architecture which was proposed by Wang et al [11] as shown in figure 3. Firstly, the authors realise that BatchNorm layers in the Residual in Residual Blocks lead to a significant decrease in computational complexity. Furthermore, the authos employ a Relativistic Discriminator instead of the standard discriminator which predicts the probability of an image being real or fake instead of solving a classification task. Finally, the authors use a combination of VGG based perceptual loss along with adversarial and L-1 Norm based content loss for the generator. Formally, the loss function for the generator is defined as

$$L_G = L_{perceptualloss} + \lambda \cdot L_{adversarialloss} + \eta \cdot L_{contentloss} \tag{1}$$

F. Applying LoRA

Fine-tuning has been a one of the most used methods for domain adaptation or transfer learning, due to its simplistic nature. LoRA [2] propose a low rank solution to disintegrate generalized weight representation from domain specific knowledge using the weight matrices of the model parameters. They propose an adapter based method to improve fine-tuning for LLMs. However, recent work has shown that these adapters work well with vision related tasks as well.

Weights are essentially huge matrices with ranks, which can be expressed as the result of some operation between 2 smaller matrices. The authors observe that data specific knowledge can be expressed using a two smaller low rank matrix,A and B, with minimal loss of information about the generalization of the model. This is underlying fact is leveraged to create a separate set of parameters. While performing fine-tuning, the pre-trained weights of the model are not subjected to any update, instead the new low-rank matrices are subjected to training. Once the training step is completed, these Low-rank matrices are embedded into the pre-trained weights. This results in an Adapter which holds rich information of the new trained task. Formally, given a set of pretrained weights, W_0, the weight update, h, can be expressed as

$$h = W_0 + \nabla W = W_0 \cdot x + BA \cdot x \tag{2}$$

Our approach to apply LoRA to our model is simple. The concept of LoRA is, when employing saved models for finetuning, the entire trainable parameters are not used. The initial parameters are frozen, while the newly added "LoRA parameters" are now trained and updated for the new domain. Thus, when deploying the model or during inference, we only need to use the LoRA parameters, instead of using all the parameters of the entire models. The code we implemented is inspired from the work done in [17].

G. Evaluation Metrics and Loss Functions

Throughout all experiments conducted in this study, we consider the Peak Signal to Noise Ratio (PSNR) in equation 3 and Structural Similarity Index (SSIM). The PSNR is a popular metric used to evaluate the effect of noise and the quality of a constructed image from the signal, while the SSIM is a full similarity metric quantifying how similar to images are to each other.

$$PSNR = 10 \cdot log_{10} \cdot \frac{MAX^2}{MSE} \tag{3}$$

IV. Experiments and Results

This study aims to evaluate and ablate on 2 major hypothesis. Firstly, we aim to evaluate the effect of LoRA on Generative Adversarial Network, more specifically the ESRGAN. The network is setup as suggested by Wang et al [11]. 23 Residual in Residual blocks are stacked together on top of a convolution and upsampling layer followed by a series of convolutions which return the predicted image.

The experiments conducted on the ESRGAN model included initial inference and fine-tuning tasks. To provide a representative comparison of performance, we detail the following section similar to an ablation study.

A. Pretrained Inference

The ESRGAN pretrained weights that we employ for our experiments is taken from the original implementation [11]. These weights are simply the ESRGAN trained a combination of real world images, including the DIV2K data set.

To demonstrate the effectiveness of the pretrained model, Table 1 displays the performance of the model on the FloodNet, Set5, and Set 14 datasets.

Experiment	PSNR	SSIM	Inf. Time
FloodNet inference	24.18	0.536	10.10s
Set5 inference	23.73	0.684	0.94s
Set14 inference	24.51	0.711	1.31s

TABLE I
COMPARISON OF ESRGAN INFERENCE RESULTS

B. Finetuning with LoRA

By employing these pretrained weights, the next step is to finetune the model on the FloodNet, Set5, and Set 14 datasets separately. In these experiments, we attempted to apply LoRA to the pretrained ESRGAN model. The model weights were loaded as a .pth file and LoRA parameters were initialized in it. The expectation was that the LoRA implemented model would have a lesser number of parameters and lesser training and inference time. This would facilitate easier fine-tuning on each of the datasets used: FloodNet, Set5, Set 14. Table 2 comparison the number of trainable parameters that will be updated during the finetuning process. Furthermore, we also evaluate the images produced by the LoRA adaptors.

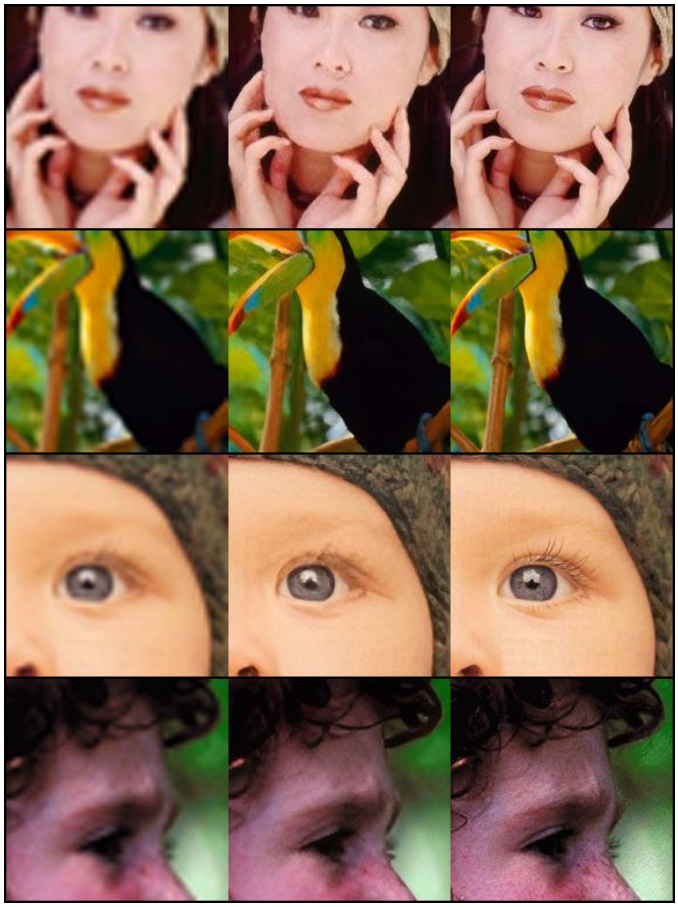

Fig. 4. Evaluation done on the Set4 Benchmark dataset. We compare the LR input, model output and ground truth HR images (left to right). The results suggest that the model is able to adapt the the dataset and produce visually valid images with a significant increase in the PSNR metric.

Model	Params w/o LoRA	Params w LoRA
ESRGAN	16,697,987	1,655,672

TABLE II
COMPARISON OF PARAMETERS BEFORE AND AFTER LORA FINETUNING.

C. Performance After LoRA

In the following table III, we can identify the improved performance for finetuning after applying LoRA. This can be quantified over model super resolution performance and its inference time.

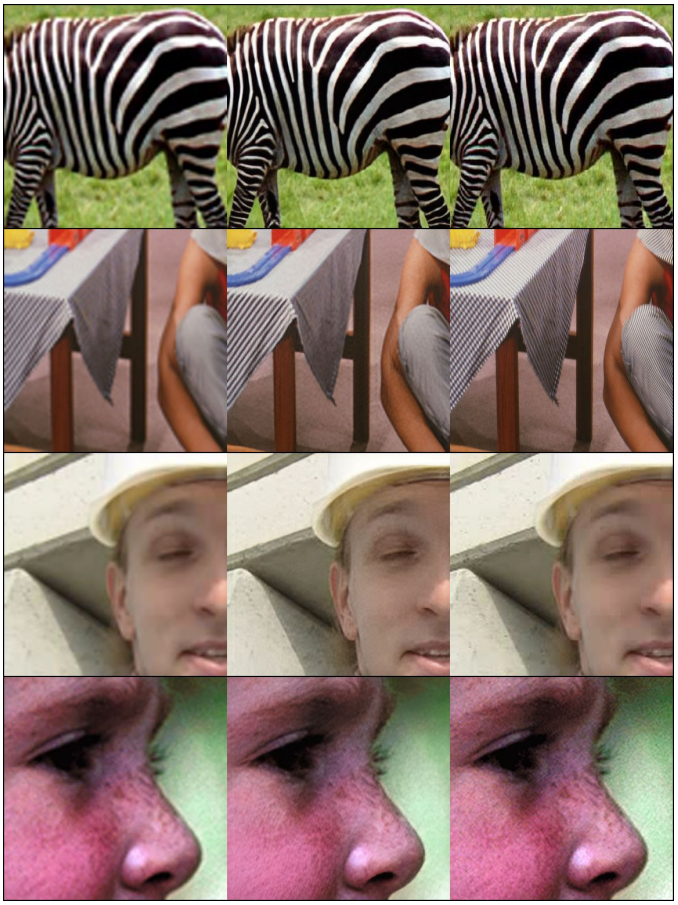

Fig. 5. Evaluation done on the Set14 Benchmark dataset. We compare the LR input, model output and ground truth HR images (left to right). The results suggest that the model is able to adapt the the dataset and produce visually valid images with a significant increase in the PSNR metric.

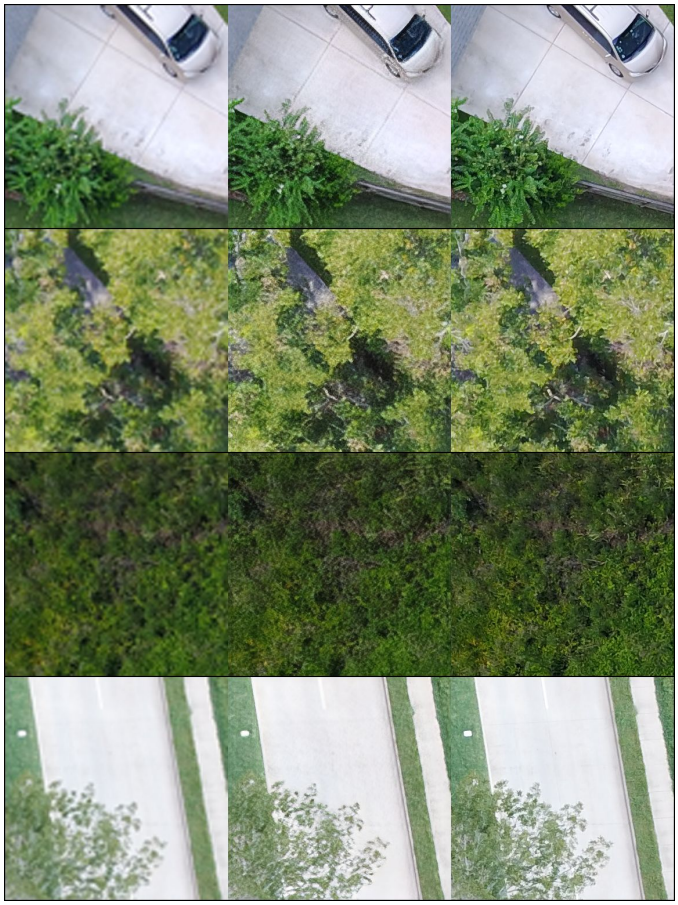

Fig. 6. Evaluation done on the FloodNet [6] Benchmark dataset. We compare the LR input, model output and ground truth HR images (left to right). The results suggest that the model is able to adapt the the dataset and produce visually valid images with a significant increase in the PSNR metric.

Experiment	PSNR	SSIM	Infer. Time
FloodNet finetuning	25.45	0.601	7.39s
Set5 finetuning	26.61	0.775	0.64s
Set14 finetuning	28.25	0.798	0.96s

TABLE III
PERFORMANCE COMPARISON AFTER FINETUNING WITH LORA.

V. SUMMARY & DISCUSSION

In this study, we attempt to the unique concept of Low Rank Adaptation to popular generative and attention based models withe the common goal of achieving single image super resolution. In addition, the effectiveness and benefits of using LoRA is quantified by finetuning the models on unseen domain datasets. Throughout the study we employ relatively moderate sized datasets, to focus our time and efforts on the effects of LoRA.

From our experiments we were able to identify the improvement in performance by applying LoRA to a popular GAN based model. Apart from the challenges we faced from Diffusion and Transformer, we can understand the application of LoRA: 1) reduces the number of trainable parameters, 2) reducing training and inference time and 3) improving the overall super resolution quality. From our experiments we observed an approximate 80 percent decrease in trainable parameters, with inference times reducing by around 30 percent. In addition to this, the PSNR has also shown a noticeable improvement across all three datasets.

In future attempts, we could consider expanding our data sets to larger domains, and complete the computations for diffusion and transformer networks.

REFERENCES

[1] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a Deep Convolutional Network for Image Super-Resolution," Computer Vision – ECCV 2014, pp. 184–199, 2014.

[2] E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," openreview.net, Oct. 06, 2021. https://openreview.net/forum?id=nZeVKeeFYf9

[3] "Transformer for Single Image Super-Resolution — IEEE Conference Publication — IEEE Xplore," ieeexplore.ieee.org. https://ieeexplore.ieee.org/document/9857219 (accessed Dec. 12, 2023).

[4] S.-J. Park, H. Son, S. Cho, K.-S. Hong, and S. Lee, "SRFeat: Single Image Super-Resolution with Feature Discrimination," Lecture Notes in Computer Science, pp. 455–471, Jan. 2018.

[5] Wang, Z., Zhao, L., & Xing, W. (2023). StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models. In Proceedings of

the IEEE/CVF International Conference on Computer Vision (pp. 7677-7689).
[6] “FloodNet: A High Resolution Aerial Imagery Dataset for Post Flood Scene Understanding — IEEE Journals & Magazine — IEEE Xplore,” ieeexplore.ieee.org. https://ieeexplore.ieee.org/document/9460988 (accessed Dec. 12, 2023).
[7] Bevilacqua, M., Roumy, A., Guillemot, C., & Alberi-Morel, M. L. (2012). Low-complexity single-image super-resolution based on non-negative neighbor embedding.
[8] Zeyde, R., Elad, M., & Protter, M. (2012). On single image scale-up using sparse-representations. In Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7 (pp. 711-730). Springer Berlin Heidelberg.
[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022, doi: https://doi.org/10.1109/cvpr52688.2022.01042.
[10] Z. Liu et al., “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2021, doi: https://doi.org/10.1109/iccv48922.2021.00986.
[11] X. Wang et al., “ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks,” Lecture Notes in Computer Science, pp. 63–79, 2019.
[12] “Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data — IEEE Conference Publication — IEEE Xplore,” ieeexplore.ieee.org. https://ieeexplore.ieee.org/document/9607421 (accessed Dec. 13, 2023).
[13] Ledig, Christian, et al. ”Photo-realistic single image super-resolution using a generative adversarial network.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[14] Li, Xiang Lisa, and Percy Liang. ”Prefix-tuning: Optimizing continuous prompts for generation.” arXiv preprint arXiv:2101.00190 (2021).
[15] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M. and Gelly, S., Parameter-Efficient Transfer Learning for NLP.
[16] https://github.com/ai-forever/Real-ESRGAN
[17] https://github.com/lizhuoq/Real-Esrgan
[18] https://github.com/cccntu/minLoRA