AI6126 Project 2
Blind Face Super-Resolution

Kozhimadam Shahin Shah, G2303660L

1 Introduction

This report presents a comprehensive solution to the mini-challenge of super-resolution of low-quality facial images using a dataset adapted from the FFHQ collection. The challenge centers around transforming 5000 high-quality images into their degraded counterparts in real-time using a provided degradation pipeline, then applying super-resolution techniques to recover image quality. The focus lies on meticulously evaluating various advanced neural network architectures and optimizing them to reconstruct high-quality images from these synthetically degraded inputs, within the framework of specified model constraints.

2 Methodology

2.1 Model Architectures

In the initial phase of the project, several lightweight model architectures were evaluated, including MSRResNet [3], RRDBNet [7], SwinIR [2], LatticeNet [4], and ShuffleMixer [5]. Each model was briefly trained for a few epochs to determine its effectiveness in super-resolution tasks. SwinIR was selected as the superior model due to its enhanced image quality and quicker convergence rates. This model incorporates shifted window-based self-attention mechanisms. Given the parameter constraints of the project, an embedding dimension of 78 was chosen to maintain an adequate model depth while keeping the total parameter count within limits. All other configurations were retained as per the standard SwinIR model, optimized for high performance and efficiency.

For the generative adversarial network (GAN) component, SwinIR was paired with a discriminator based on the U-Net architecture with spectral normalization [6]. To align with the lower capacity of the super-resolution (SR) network, the feature dimension of the discriminator was reduced to 24. This adjustment ensured that the discriminator did not overpower the generator, allowing for a balanced training process and enhancing the stability of the adversarial training. This configuration was employed only when the model was trained using GANs, resulting in two distinct setups by the project's conclusion: SwinIR alone (PSNR-based model) and SwinIR with a modified U-Net discriminator (GAN-based model).

2.2 Loss Functions

MSE loss was utilized when training the super-resolution (SR) network independently. This loss function calculates the mean squared error between each pixel in the predicted high-resolution image and the ground truth. This approach is directly related to maximizing the Peak Signal-to-Noise Ratio (PSNR), which is crucial for the project since the challenge evaluation is based on PSNR scores. A higher PSNR indicates a lower error rate and, thus, a higher quality of the reconstructed images compared to the original, minimizing the impact of noise on the signal's fidelity.

In the context of training with a generative adversarial network (GAN), the loss function strategy was adapted to include L1 pixel loss, perceptual loss, and GAN loss [1]. L1 pixel loss is employed here to promote more natural-looking results in the generated images, as it tends to preserve high-frequency details better than MSE, which can sometimes lead to blurrier results. Perceptual loss continues to enhance the visual quality by comparing the feature representations of the predicted and actual high-resolution images at different layers of a pretrained classification network, thus aligning the output more closely with human visual perception. Additionally, GAN loss was applied to enhance the adversarial training aspect, with the weight for GAN loss in the generator set at 0.1. This specific weighting helps balance the training process, ensuring that the generator not only produces realistic but also high-quality images that are true to the source material in both detail and texture.

2.3 Data Processing

To generate low-quality images from high-quality originals, a second-order degradation model was employed [6], closely mirroring practical degradation scenarios. This model sequentially applies classical degradation steps, encompassing blurring, resizing, noise introduction, and JPEG compression. Specific choices for each degradation type were as follows: Gaussian blurring (isotropic and anisotropic variants), downsampling (bicubic, bilinear, and area methods), additive noise (Gaussian, Poisson, color, and gray noise), and standard JPEG compression along with a 2D sinc filter to mimic common compression artifacts like ringing and overshoot. These processes were meticulously orchestrated to produce a varied set of degraded images, forming a comprehensive dataset for robust super-resolution model training and evaluation.

3 Training

3.1 Training Procedure

In the training of the super-resolution (SR) network, the degradation model was applied to each batch to simulate low-quality images, utilizing PyTorch with CUDA acceleration. This approach, while efficient, constrained the diversity of synthetic degradations within a batch. To address this limitation, a buffered dataloader was introduced. This dataloader featured a buffer that was replenished with batches from the primary training dataloader. Shuffling of the buffer occurred before the retrieval of each batch to ensure diversity, and the buffer was subsequently replenished from the training dataloader as long as data remained available. This mechanism facilitated independent processing of data from the model module while maintaining variation within the batch data.

For both the PSNR-based and GAN-based models, batch sizes of four were employed for both training and testing, with computations performed using BF16 precision. Gradient clipping was set at a maximum value of 1.0 to prevent exploding gradients. The Adam optimizer was configured with a learning rate of 2×10^{-4}, β_1 of 0.9, β_2 of 0.999, a weight decay of 0, and an epsilon of 1×10^{-8}. A cosine learning rate scheduler was used, with zero warm-up steps. Moreover, an exponential moving average of the model weights with a decay factor of 0.999 was utilized for inference to stabilize and improve generalization by reducing the variance in the model parameters.

The PSNR-based model underwent training for 700 epochs and the GAN-based model was initialized with weights from the PSNR-based model and trained for 300 epochs, with alternating updates between the SR network and the discriminator. To achieve a balance between the natural-looking results of the GAN-based model and the more faithful reproduction of the original low-quality images by the PSNR-based model, the weights of both models were interpolated using a factor of 0.25 [7], tilting towards the GAN-based model. This approach allowed for combining the strengths of both models, enhancing the visual naturalness while maintaining fidelity to the original. Peak Signal-to-Noise Ratio (PSNR) served as the metric for performance evaluation throughout the training process.

3.2 Training Machine Specifications

Training was conducted on a high-performance setup featuring two NVIDIA RTX 6000 Ada Generation GPUs, each with 48GB of memory. This configuration was selected for its advanced computational capabilities and ample memory, essential for efficiently managing the demands of the training process. The dual GPU setup significantly expedited training times and supported the utilization of larger batch sizes, enabling more extensive experimentation with model configurations and hyperparameters, thereby enhancing the overall model training and development process.

4 Results

4.1 Training Curves

Figure 1 displays the training dynamics of the two distinct models. Part (a) of the figure illustrates the loss curves for the PSNR-based model. The trend of both training and validation losses steadily decreasing and converging indicates a stable training process. It is observed that, apart from the initial training phase, the training loss generally remains above the validation loss. This pattern can be attributed to the dynamic nature of low-quality sample generation for training, as opposed to static validation samples.

In part (b) of Figure 1, the loss curves for the GAN-based model exhibit fluctuations, which is typical of the instability seen in adversarial training processes. These fluctuations reflect the ongoing competition between the generator and discriminator. An increasing trend in loss as training progresses suggests that the discriminator is

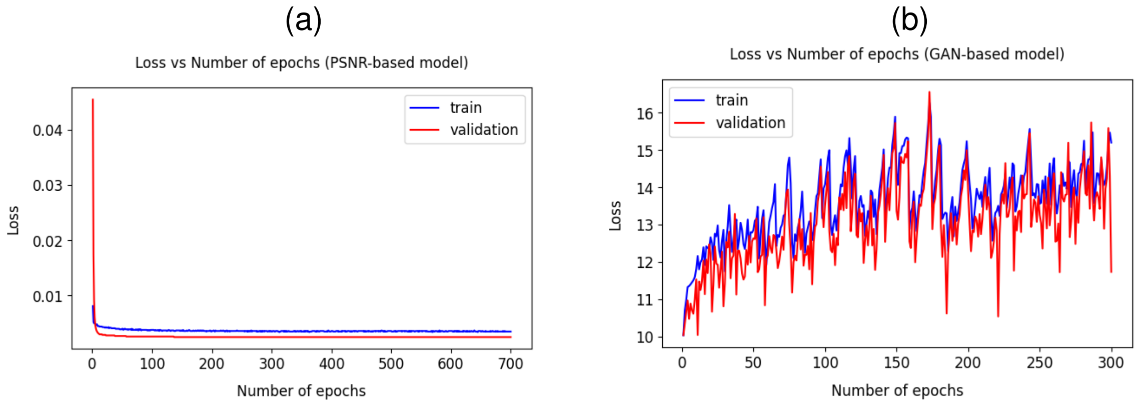

Figure 1: Training and validation loss curves for (a) the PSNR-based model and (b) the GAN-based model over the number of epochs.

gaining an advantage over the generator, underscoring the difficulty in balancing the adversarial components to achieve stable training.

4.2 Model Parameters

The SwinIR model comprises 2,251,758 parameters. This parameter count is within the limits set by the challenge, demonstrating the model's ability to effectively balance complexity and computational efficiency while adhering to specified constraints. In the adversarial training setup, the U-Net discriminator is configured with 616,057 parameters, strategically kept lower to ensure that it complements the generative capacity of the SwinIR model without overwhelming it during the adversarial learning phase.

4.3 Validation and Test Dataset Results

Refer to Table 1 for a comparative analysis of the PSNR values across the PSNR-based, GAN-based, and Interpolated models on validation and test datasets. The PSNR-based model achieves the highest PSNR, typically indicative of lower error rates in image reconstruction, though it tends to produce blurrier outcomes. In contrast, the GAN-based model, despite lower PSNR values, yields more natural-looking images. The Interpolated model, as anticipated, delivers PSNR values that lie between those of the PSNR-based and GAN-based models, offering a balance between natural appearance and faithfulness to the original low-quality images.

Table 1: PSNR values for the PSNR-based model, GAN-based model, and Interpolated Model on validation and test datasets.

Model Type	Validation PSNR (dB)	Test PSNR (dB)
PSNR-based Model	26.55	26.62
GAN-based Model	25.26	25.33
Interpolated Model	25.60	25.68

4.4 Results on Real-World LQ Images

The outcomes from applying the interpolated model to enhance real-world low-quality images are displayed in Figure 2. The results illustrate that the model effectively produces high-quality images that appear both natural and faithful to the original low-quality inputs.

5 Discussion

The chosen architectures and loss functions for this super-resolution challenge reflect a deliberate balance between achieving high resolution, maintaining natural image appearance, and adhering to computational constraints. The SwinIR model was identified as optimal not only for its capability to upscale low-resolution images to high-quality high-resolution outputs efficiently but also for its adaptability to the stringent parameter constraints of the challenge. Enhanced by a generative adversarial network approach, SwinIR demonstrated how architectural choices directly impact the quality of super-resolution outcomes, particularly in how these models manage the trade-offs between enhancing image details and preserving naturalness.

The deployment of various loss functions tailored to different aspects of the super-resolution process—ranging from MSE pixel loss for basic accuracy to L1, perceptual and GAN losses for improved visual fidelity—highlights the complex nature of modeling decisions that go beyond mere error minimization. These choices underscore the critical role of loss functions in steering model training towards desired outcomes, especially in scenarios where visual quality is subjective and multifaceted.

Furthermore, the introduction of a second-order degradation model and a buffered dataloader setup played crucial roles in simulating realistic training conditions. These methodologies ensured robustness and versatility in the trained models, preparing them for effective application on real-world data. This approach not only facilitated the practical training of these models but also highlighted the necessity for innovative data handling techniques in the advancement of super-resolution technologies.

6 Conclusion

The project successfully demonstrated the application of advanced neural network architectures to enhance low-quality facial images, significantly improving image resolution while maintaining naturalness. The use of the SwinIR model, augmented by generative adversarial techniques, addressed the challenges of super-resolution effectively, showcasing a balanced approach in preserving image authenticity and enhancing detail. The strategic implementation of varied loss functions and innovative data processing methodologies enabled robust performance across both standardized validation and complex real-world scenarios.

The project highlighted the potential of combining traditional and adversarial training methods to produce super-resolved images that are both high in quality and faithful to the original, even within the stringent constraints of the provided dataset and computational limits. These findings demonstrate how sophisticated modeling techniques can be effectively applied in practical scenarios.

References

[1] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[2] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[3] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[4] Radu Alexandru Rosu, Peer Schütt, Jan Quenzel, and Sven Behnke. Latticenet: fast spatio-temporal point cloud segmentation using permutohedral lattices. *Autonomous Robots*, 46(1):45–60, 2022.

[5] Long Sun, Jinshan Pan, and Jinhui Tang. Shufflemixer: An efficient convnet for image super-resolution. *Advances in Neural Information Processing Systems*, 35:17314–17326, 2022.

[6] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021.

[7] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

A Appendix

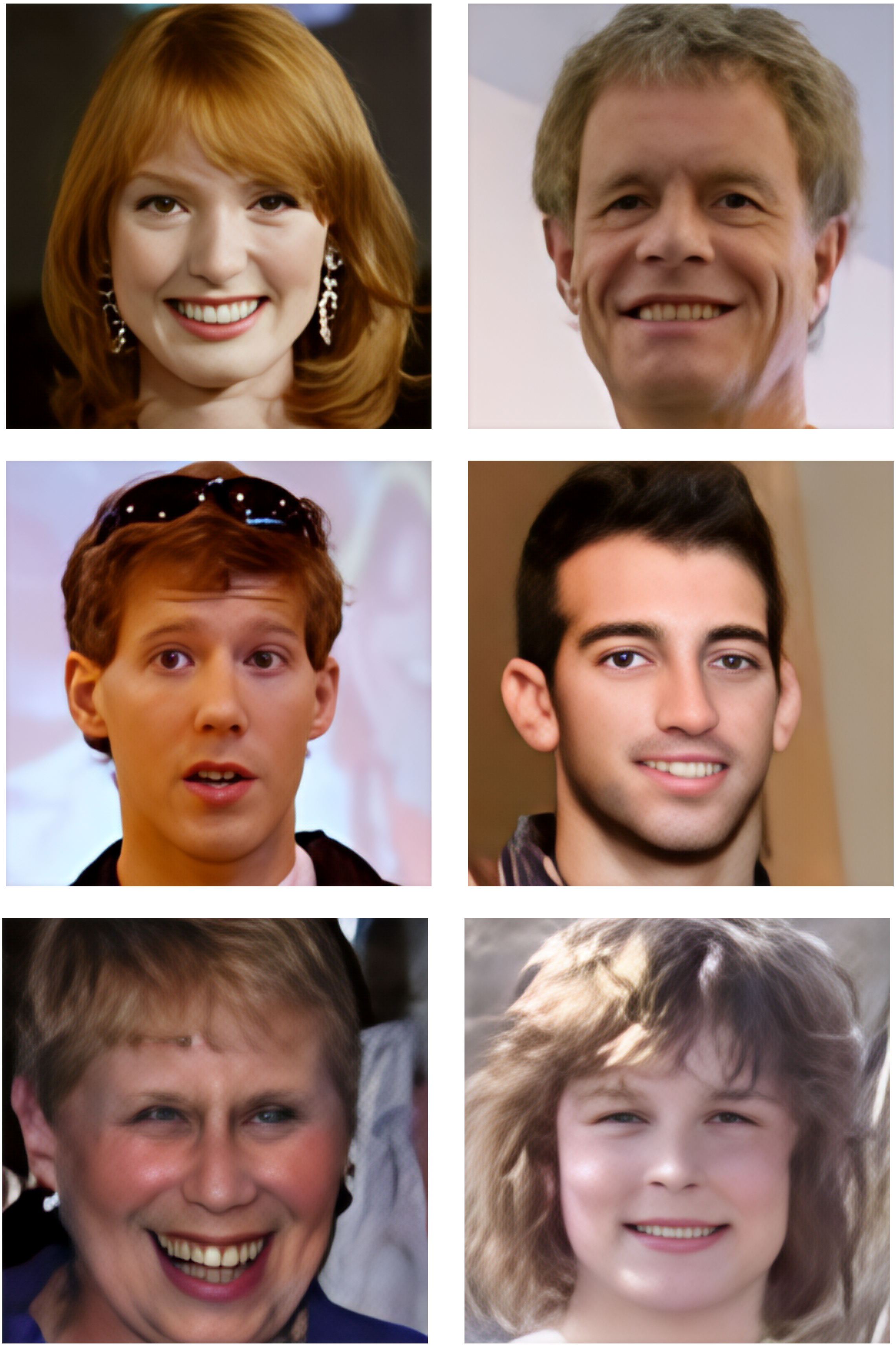

Figure 2: Predicted HQ images on 6 real-world LQ images