Image super resolution of Human face images using GAN

Zihuan Jiang, Shenxiaozhu Xu, Lucia Sun

Abstract

This paper investigates the application of Generative Adversarial Networks (GANs) for enhancing the resolution of human facial images, aiming to improve image quality. Utilizing the Flickr-Faces-HQ dataset, our GAN model upscales low-resolution images (512x512 pixels) to higher resolutions (1024x1024 pixels). We assess model performance through qualitative evaluations and quantitative measures, particularly the Peak Signal-to-Noise Ratio (PSNR), where our model achieved a PSNR of 25.66 dB. This indicates a satisfactory level of enhancement, though constrained by limited GPU resources, which impacted the model's depth and training duration. The findings highlight the effectiveness of GANs in image super-resolution tasks and underscore the need for enhanced computational resources to achieve greater fidelity in super-resolved images.

1. Introduction

Image super-resolution, the process of enhancing the detail and quality of images, has always been a critical challenge in the field of digital imaging and computer vision. Its importance spans across various applications, from satellite imaging and medical diagnostics to digital archiving and law enforcement. Historically, the resolution of an image could mean the difference between identifying a critical geographical feature in satellite imagery or decoding microscopic details in a medical scan. For instance, during the Cuban Missile Crisis in 1962, the ability to discern missile sites from high-altitude reconnaissance photographs played a pivotal role in international negotiations [1]. Similarly, in the medical field, the development of high-resolution imaging techniques like MRI and CT scans has revolutionized diagnostics, allowing for early detection and treatment planning for countless conditions.

Prior to the advent of deep learning and Generative Adversarial Networks (GANs), traditional methods for image upscaling included bicubic interpolation and other algorithmic approaches that essentially estimated and filled in missing pixel information based on surrounding pixels [2]. While these techniques provided a basic framework for image enhancement, they often resulted in loss of details with significant levels of degradation [2].

The introduction of GANs has marked a significant turning point in the field of image super-resolution, especially concerning human faces. GANs, through their adversarial training process, learn to generate high-resolution images that are increasingly difficult to distinguish from real, high-resolution images. This capability is especially useful in applications involving human faces, where clarity and detail are key for tasks such as facial recognition in security systems, enhancing low-resolution footage in law enforcement investigations, and improving the quality of video calls in telecommunications [3]. Furthermore, in the entertainment industry, super-resolution can breathe new life into historical footage, allowing for the restoration and preservation of cinematic history. A notable example that illustrates this is the project "Bringing the Past Back to Life" by Denis Shiryaev. Utilizing GANs, Shiryaev spectacularly enhanced and colorized the Lumière brothers' iconic 1896 film, "Arrival of a Train at La Ciotat," transforming it from a grainy, black-and-white short into a vibrant, high-definition experience [4].

However, the challenge of enhancing human face images is compounded by the complexity of facial features and the importance of maintaining individual identity characteristics. Low-quality facial images can lead to misidentification, false positives in security systems, and loss of critical information in digital forensics. The balance between enhancing image quality and preserving identity integrity is a nuanced field of study within image super-resolution, pushing the boundaries of GAN capabilities and exploring new frontiers in artificial intelligence and machine learning.

This paper aims to leverage GANs for the super-resolution of human faces, navigating the balance between enhancing image quality and maintaining the essence of individual identity. By delving into this specialized application of GANs, we seek to address the challenges posed by low-quality facial images. Our exploration not only contributes to the advancement of super-resolution techniques but also highlights the transformative impact of GANs in refining our visual world.

2. Literature Reviews

2.1. Face Image Super-resolution Based On Relative Average Generative Adversarial Networks [5]

Ying Liu and Li Zhu's paper presents a novel approach for facial image super-resolution using Relative Average Generative Adversarial Networks (RaGAN), which surpass traditional methods by focusing on the relative realism of images and integrating a dual-focused loss function for enhanced detail and integrity.

2.2. Transfer-Gan: Multimodal Ct Image Super-Resolution Via Transfer Generative Adversarial Networks [6]

Xiao et al.'s study introduces TransferGAN, a novel approach that combines Generative Adversarial Networks (GANs) with transfer learning to enhance the resolution of multimodal computed tomography (CT) images. By leveraging shared information across different imaging modalities, TransferGAN significantly improves visualization and quantitative image quality.

2.3. ID Preserving Face Super-Resolution Generative Adversarial Networks[3]

Li et al. introduce IP-FSRGAN, a novel Generative Adversarial Network-based framework designed for enhancing low-resolution facial images to high resolution while preserving the subject's identity through an innovative ID preserving module. IP-FSRGAN's integration of an ID preserving loss marks an advancement in super-resolution imaging, effectively maintaining identity features in super-resolved images.

2.4. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network [7]

Ledig et al.'s SRGAN, a Generative Adversarial Network for image super-resolution, enhances the perceptual quality of upscaled images by focusing on fine texture recovery through a perceptual loss function. This function combines adversarial and content losses, leading to photo-realistic results at 4x upscaling, surpassing traditional methods focused on minimizing mean squared error.

2.5. CA-ESRGAN: Super-Resolution Image Synthesis Using Channel Attention-Based ESRGAN [8]

This study introduces enhancements to the ESRGAN algorithm by incorporating SENet and ECA-Net for improved feature interdependence and model efficiency, and by using LPIPS in the discriminator for advanced perceptual quality assessment.

2.6. Edge-Enhanced GAN for Remote Sensing Image Superresolution [9]

Jiang et al.'s study introduces the Edge-Enhanced GAN (EEGAN), which combines an ultradense subnetwork (UDSN) for reconstructing sharp super-resolution images from low-resolution inputs, and an edge-enhancement subnetwork (EESN) that purifies noise-contaminated edges.

2.7. RankSRGAN: Generative Adversarial Networks With Ranker for Image Super-Resolution [10]

RankSRGAN addresses the challenge of evaluating perceptual-based super-resolution (SR) techniques in the absence of objective metrics like PSNR and SSIM, which complicates direct comparison of different algorithms. By generating super-resolved images using various SR methods and ranking them with a non-differentiable perceptual metric (e.g., NIQE), RankSRGAN trains a Ranker with a Siamese architecture to mimic the behavior of these perceptual metrics using margin-ranking loss.

2.8. Frequency Domain-Based Perceptual Loss for Super Resolution [11]

S. Sims introduces the Frequency Domain Perceptual Loss (FDPL) for single-image super-resolution, a novel approach that operates in the frequency domain using the Discrete Cosine Transform (DCT) and JPEG luminance quantization table to enhance perceptual quality by focusing on frequencies vital to human perception.

2.9. Edge-Based Loss Function for Single Image Super-Resolution [12]

The paper introduces an edge-based loss function for super-resolution in CNNs, addressing the blurriness caused by MSE loss by focusing on salient edge reconstruction for sharper images. Demonstrating superior quantitative and qualitative results over MSE, especially in images with prominent edges, this approach emphasizes the importance of edges for improved human perception and computer vision tasks.

2.10. MSG-CapsGAN: Multi-Scale Gradient Capsule GAN for Face Super Resolution [13]

The MSG-CapsGAN introduces a novel approach for facial super-resolution (SR), tackling the challenge of distortion in high-resolution results from extremely low-resolution images by employing a Multi-Scale Gradient GAN combined with a Capsule Network as its discriminator. This model simplifies the SR process while maintaining high precision and pose invariance.

3. Description of the dataset

The FFHQ dataset (https://github.com/NVlabs/ffhq-dataset) contains 70,000 high-resolution, unlabelled PNG images, each 1024 by 1024 pixels, featuring a single face. From this dataset, we randomly selected 500 images and divided them into a training set (50 percent), a testing set (30 percent), and a validation set (20 percent).

4. Problem Statement

The goal of this paper is to train a Generative Adversarial Network for image super-resolution based on the Flickr-Faces-HQ(FFHQ) dataset. Given a low resolution image of dimensions 512 by 512 pixels, generate the corresponding high resolution image of dimensions 1024 by 1024 pixels.

5. Model Description

5.1. Data Preprocessing

The preprocessing of input data is a crucial step in the preparation for the training, validation, and testing phases of both

low-resolution and high-resolution images. For low-resolution images, a comprehensive transformation protocol is applied across all data subsets. This includes the application of a Gaussian blur with a kernel size of (7,7) and a sigma of 5 to simulate the low-resolution effect and potentially mitigate noise, resizing of images to a consistent dimension of 512 by 512 pixels to ensure uniform input sizes, and normalization of the RGB color channels based on predetermined mean and standard deviation values. This normalization process is designed to align the data distribution closely with that of pre-trained networks, thereby enhancing the efficacy of transfer learning strategies.

In contrast, the preprocessing process for high-resolution images is adjusted to preserve intrinsic details, omitting the Gaussian blur. These images are resized to 1024 x1024 pixels to maintain a high level of detail and undergo the same normalization process as their low-resolution counterparts.

5.2. Generator

The generator takes in the transformed images (low resolution) and recreates details to produce the corresponding high-resolution images. The architecture involves convolutional layers, pixel shuffling, and a distinctive mechanism for integrating residual scaling, thus forging a path toward the synthesis of realistic images.

The generator begins with a convolutional layer that downsamples the input image by a factor of 2. This initial step, through a convolution operation characterized by a kernel size of 2 and a stride of 2, reduces the spatial dimensions of the input while expanding its channel depth to 64. This layer lays the groundwork for a sequence of following convolutional layers, which preserve the spatial dimensions of their inputs through a chosen stride and padding configuration. These layers, while varying the depth of the feature maps across a spectrum of channels, are interspersed with the application of the Tanh activation function.

A distinctive feature of the generator is the incorporation of the pixel shuffle operation, strategically deployed at key points within the model to facilitate image upsampling. This technique not only enhances the resolution of the feature maps but also allows for modulation of the channel depth. As a result, the generator is equipped with a refined ability to improve the detail in the images it creates.

Further distinguishing this model is the incorporation of a residual connection, which leverages bilinear interpolation to upsample the input to align with the spatial dimensions of the model's final output. This residual output is combined with the output from the main processing pathway, modulated by a learned scale factor. The scale factor is constrained to a minimum of 0.2 to preserve the original image's style and a maximum of 0.8 to prevent the bilinear residual from becoming overly dominant. Such an approach significantly enhances the detail of the generated images.

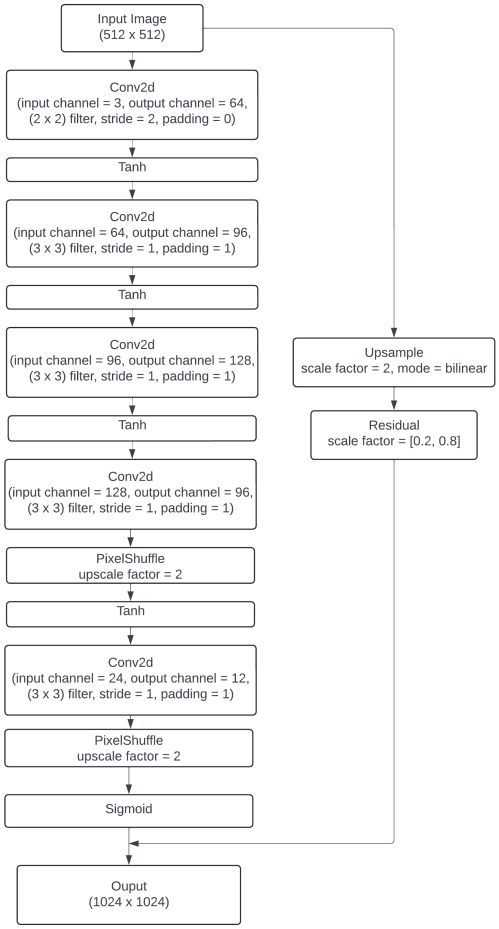

Figure 1: Generator Architecture

5.3. Discriminator

The discriminator's architecture is designed to process input images generated by the Generator through a series of layers, with the ultimate goal of efficiently distinguishing between the generated high-resolution images and the original high-resolution datasets. The core of the discriminator is constructed from sequential discriminator blocks, each designed to perform specific functions: convolution layer, leaky ReLU activation, dropout layer, batch normalization (conditional), and adversarial layer.

This discriminator begins with a two-dimensional convolutional layer, which accepts an input comprising three channels (representing RGB images) and transforms it into a tensor with 12 feature maps. This transformation is facilitated by the application of a 3x3 kernel, employing a stride of 2 for downsampling, while a padding of 1 is maintained to approximate the original input dimensions closely. The inclusion of the leaky ReLU activation function introduces non-linearity into the model, ensuring the propagation of gradients even for negative input values. To mitigate the risk of overfitting, a dropout layer is applied, randomly nullifying a portion of the feature

maps.

The discriminator's architecture is further enriched with five additional convolutional blocks, progressively modifying the depth of the feature maps from 12 to 48, then to 72, subsequently reducing it to 48, and ultimately scaling it down to 12 and finally 3 channels. Each convolutional block is succeeded by the application of leaky ReLU and dropout layers, employing a consistent approach to introduce non-linearity and prevent overfitting. Importantly, the architecture incorporates BatchNorm2d layers following specific convolutional layers, which standardize the output by normalizing it.

The culmination of the convolutional sequence leads to an advancement layer, performing a linear transformation that condenses the flattened output of the preceding convolutional layer into a singular value. This value represents the discriminator's judgment regarding the authenticity of the input image. Following it, a sigmoid activation function is applied, constraining the output to a range between zero and one, thereby rendering it interpretable as the probability of the image being real.

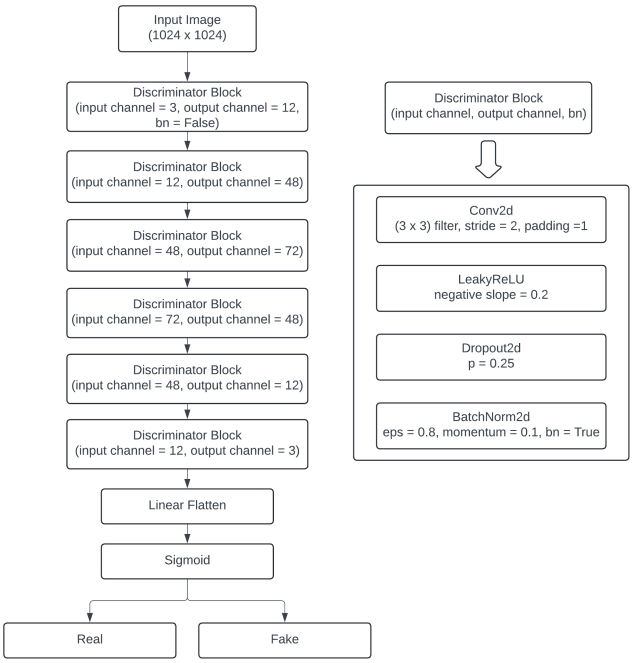

Figure 2: Discriminator Architecture

5.4. Training

Throughout this training period, we employed the binary cross-entropy loss function and used the Stochastic Gradient Descent (SGD) optimizer for optimizing both the generator and the discriminator.

The discriminator's goal is to distinguish between real and generated images. It computes losses separately for misclassified real images and for misclassified fake images. The total loss is the sum of these two losses. The generator is trained to fool the discriminator. It generates high-resolution images from low-resolution ones, and these generated images are fed to the discriminator. The loss is computed based on how well the discriminator is fooled into thinking the generated images are real.

At the end of each epoch, the generator is evaluated on a validation set. This involves generating high-resolution images from low-resolution images in the validation dataset and comparing these generated images with the real high-resolution images using the PSNR (Peak Signal-to-Noise Ratio) metric to quantify image quality improvements.

Our training process spanned 80 epochs, with the initial 60 epochs utilizing a learning rate of 0.02, followed by a learning rate of 0.005 for the remaining 20 epochs. The decrease in the learning rate was prompted by the observation that the validation accuracy had begun to plateau (Figure 3).

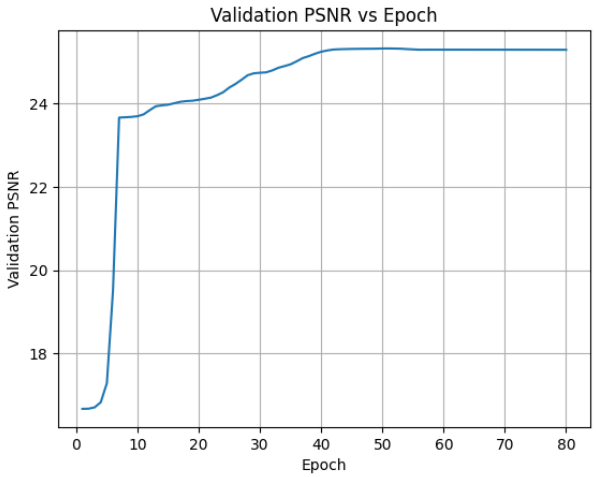

Figure 3: Graph showing validation accuracy begins to plateau

6. Results

6.1. Quantitative Results

In the evaluation of our Generative Adversarial Network (GAN) model's performance on super-resolution tasks, we utilize the Peak Signal-to-Noise Ratio (PSNR) as a critical metric. The PSNR, which measures the quality of reconstructed images compared to their original high-resolution counterparts, has proven useful for assessing the fidelity of image enhancements. Achieving an average PSNR of 25.66 dB on the test set, our GAN model demonstrates satisfactory performance.

Historically, PSNR values have varied across different GAN implementations. Early or simpler models often reported PSNRs in the range of 20 dB to 24 dB. These models likely lacked advanced architectural enhancements or sophisticated training protocols that later iterations, like ours, incorporated. Conversely, state-of-the-art models, employing more complex

layers and extensive training, have achieved PSNRs exceeding 28 dB. These models benefit significantly from deep networks, which are instrumental in capturing high-frequency details essential for superior PSNR scores.

Our model's performance, while superior to many earlier implementations, does not reach the peak values reported by the most advanced frameworks. The architecture and training of our Generative Adversarial Network (GAN) model for super-resolution were constrained due to limited GPU access, impacting both the model's depth and its training regimen. Specifically, the number of convolution layers and feature maps was reduced to manage computational demands within the available resources.

6.2. Qualitative Results

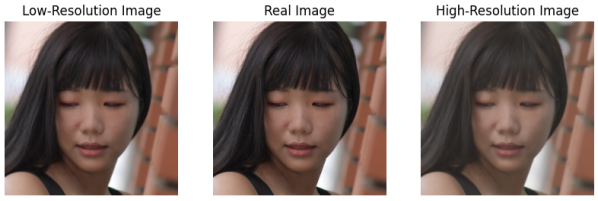

Figure 4: Output result example 1

Figure 5: Output result example 2

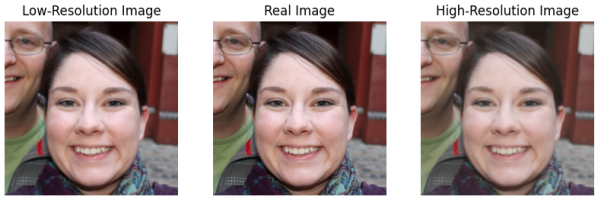

Figure 6: Output result example 3

From the qualitative results, it is evident that our GAN model successfully generates high-resolution images from low-resolution inputs that reasonably resemble the real images, with slight variations in color style. This observation indicates that while our model effectively captures the overall structure and details of the original images, there is room for improvement in accurately reproducing the color dynamics. Such discrepancies in color reproduction might stem from the model's current configuration or training limitations, suggesting that further tuning of the network's layers or an extension of the training process could enhance its fidelity in color matching.

7. Conclusion

To summarize, this research has leveraged Generative Adversarial Networks (GANs) to advance the field of image super-resolution, specifically focusing on enhancing human facial images. Throughout this study, our GAN model demonstrated a satisfactory ability to upscale low-resolution images to higher resolutions with a reasonable degree of accuracy, achieving an average Peak Signal-to-Noise Ratio (PSNR) of 25.66 dB. This performance situates our model above many prior implementations yet below the capabilities of the most advanced current models, primarily due to constraints imposed by limited access to computational resources, such as GPU availability. This limitation affected both the depth of our model's architecture and the extent of its training, thereby restricting the potential for achieving higher PSNR values which are indicative of superior image quality.

Qualitatively, our model successfully generated images that closely resemble real high-resolution photographs, although some discrepancies in color style were noted. This suggests that while the model effectively captures structural details, improvements are needed in color accuracy, potentially through further tuning of the network's layers or extended training periods.

By pushing the boundaries of GAN capabilities and exploring new frontiers in artificial intelligence for image processing, this research not only enhances our understanding of super-resolution techniques but also highlights the transformative potential of advanced machine learning technologies in refining visual data. Future work will focus on overcoming current limitations by enhancing computational capacity, refining model architectures, and extending training protocols.

References

[1] National Geospatial-Intelligence Agency, 13 days over cuba: The role of the intelligence community in the cuban missile crisis, 2023, accessed: 2024-02-24.

[2] Y. Li, F. Qi, Y. Wan, Improvements on bicubic image interpolation, in: 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Vol. 1, 2019, pp. 1316–1320. doi:10.1109/IAEAC47372.2019.8997600.

[3] J. Li, Y. Zhou, J. Ding, C. Chen, X. Yang, Id preserving face super-resolution generative adversarial networks, IEEE Access 8 (2020) 138373–138381. doi:10.1109/ACCESS.2020.3011699.

[4] A. Liszewski, Neural networks upscale film from 1896 to 4k, make it look like it was shot on a modern smartphone, 2020, accessed: 2023-02-26.

[5] Y. Liu, L. Zhu, Face image super-resolution based on relative average generative adversarial networks, in: 2021 2nd Asia Symposium on Signal Processing (ASSP), 2021, pp. 38–43. doi:10.1109/ASSP54407.2021.00014.

[6] Y. Xiao, K. R. Peters, W. C. Fox, J. H. Rees, D. A. Rajderkar, M. M. Arreola, I. Barreto, W. E. Bolch, R. Fang, Transfer-gan: Multimodal ct image super-resolution via transfer generative adversarial networks, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 195–198. doi:10.1109/ISBI45749.2020.9098322.

[7] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic single image super-resolution using a generative adversarial network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 105–114. doi:10.1109/CVPR.2017.19.

[8] J. Liu, N. P. Chandrasiri, Ca-esrgan: Super-resolution image synthesis using channel attention-based esrgan, IEEE Access 12 (2024) 25740–25748. doi:10.1109/ACCESS.2024.3363172.

[9] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, J. Jiang, Edge-enhanced gan for remote sensing image superresolution, IEEE Transactions on Geoscience and Remote Sensing 57 (8) (2019) 5799–5812. doi:10.1109/TGRS.2019.2902431.

[10] W. Zhang, Y. Liu, C. Dong, Y. Qiao, Ranksrgan: Generative adversarial networks with ranker for image super-resolution, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 3096–3105. doi:10.1109/ICCV.2019.00319.

[11] S. D. Sims, Frequency domain-based perceptual loss for super resolution, in: 2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP), 2020, pp. 1–6. doi:10.1109/MLSP49062.2020.9231718.

[12] G. Seif, D. Androutsos, Edge-based loss function for single image super-resolution, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 1468–1472. doi:10.1109/ICASSP.2018.8461664.

[13] M. M. Majdabadi, S.-B. Ko, Msg-capsgan: Multi-scale gradient capsule gan for face super resolution, in: 2020 International Conference on Electronics, Information, and Communication (ICEIC), 2020, pp. 1–3. doi:10.1109/ICEIC49074.2020.9051244.