Face Image Super-Resolution Reconstruction Based on the Improved ESRGAN

Mingda Li
Computer Science, Xi'an Jiao Tong University, Xi'an, China

Keywords: Face Super-Resolution, ESRGAN, PieAPP Loss.

Abstract: Researchers have investigated face super-resolution (SR) reconstruction extensively because it can effectively improve biometric identification, surveillance, and image improvement. This study's primary objective is to enhance generative adversarial networks (GANs) in order to create powerful face super-resolution models. Specifically, this paper introduces the improved enhanced SR-GAN (ESRGAN) model as a baseline. In addition, this paper proposes the Perceptual Image-Error Assessment through Pairwise Preference (PieAPP) loss function to fuse and perceive the visual quality of facial images. Secondly, eliminating the batch normalization layer is proposed to improve the model structure while incorporating residual dense blocks (RRDB) in the generator. The introduction of PieAPP loss can refine the SR process and emphasize the quality of the generated image. This study is conducted on an extensive and diverse face dataset, which includes facial images at various resolutions. Experimental results show that the "L1+Visual Geometry Group (VGG)+PieAPP" loss function is always better than the original ESRGAN model in various quantitative indicators. These improvements result in superior perceived image quality and user satisfaction. The practical significance of this research lies in its contribution to the development of more realistic and aesthetically pleasing facial reconstructions.

1 INTRODUCTION

GANs are an excellent generative model (Hong et al 2019). Since the inception of GANs, there has been a remarkable acceleration in generating high-resolution (HR) images. GANs have found diverse applications across fields such as photo editing, video synthesis, domain translation, and augmenting data. One particularly important application is SR, a technique to enhance the spatial detail of digital images (Lin et al 2018). Currently, Super-Resolution Generative Adversarial Networks (SRGANs) can be used for face reconstruction. Facial images contain critical biological features unique to individuals, making them valuable for identity verification. However, challenges arise due to factors such as the quality of imaging hardware, shooting conditions, subject movement, and age-related changes. These challenges often result in low-resolution issues like blurriness, noise, and distortions in facial images, rendering them unsuitable for reliable face recognition (Cao et al 2021). So there is a pressing need for the development of super-resolution reconstruction methods tailored for facial images. These techniques seek to increase the image quality as well as the precision and efficacy of face recognition systems.

Karras and his team have found a significant problem in generator networks related to image quality. They propose a unique approach of treating network signals as continuous to fix this issue effectively. Their generated networks are impressive in handling image translation and rotation, even for small details. These networks have distinct internal structures compared to Style Generative Adversarial Networks (StyleGAN2) but maintain comparable image quality. This discovery holds promise for improved generative models, particularly in video and animation applications. Yuan and colleagues address the challenges of low-resolution images and the absence of effective downsampling techniques through a three-phase approach. They introduce a network architecture called "cycle-in-cycle." Initially, the input is transformed into a low-resolution, noise-free space. Then, an already-trained deep model is applied for upsampling. Extensive tuning is performed on both modules to obtain high-resolution results. Their unsupervised method, tested on NTIRE2018 datasets, yields results comparable to advanced supervised models with peak

Li, M.
Face Image Super-Resolution Reconstruction Based on the Improved ESRGAN.
DOI: 10.5220/0012820100003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 471-476
ISBN: 978-989-758-705-4

signal-to-noise ratio (PSNR) and structural similarity index (SSIM) values of 24.33 and 0.69, respectively (Karras et al 2021). YUN and their team have developed a method for creating high-resolution (HR) facial photos that are visually impressive and free of blurriness. They start by extracting feature maps from low-resolution (LR) facial photos using a five-layer convolutional neural network (CNN). Following this, clear and blurry HR face images are produced using two-branch encoder-decoder networks. The technique improves the reconstruction of HR face structures by using both local and global discriminators. The success of this approach in creating lifelike HR face photographs from a wide range of LR inputs is supported by both qualitative and quantitative evaluations. Furthermore, a practical use-case scenario validates the hypothesis that this method has the potential to advance the field of face recognition beyond current techniques (Zhang et al 2020).

Currently, face image super-resolution reconstruction algorithms can be broadly categorized into three groups: techniques that rely on interpolation, reconstruction, and machine learning, where machine learning-based methods have better performance (Parker et al 1983, Tong et al 2017 & Ledig et al 2017). The main objective of this study is to explore an improved ESRGAN model for face image super-resolution reconstruction (Choi and Park 2023). Specifically, first, to further elevate the visual quality, this study thoroughly analyzes the network design, adversarial loss, and perceptual loss—three important facets of SRGANs—and improve each of these elements. Second, the author uses additional perceptual loss (computed using the pre-trained PieAPP web network computation) to train the generator, adding jump connections to the discriminator to use different combinations of features at different scales, and replacing the Leaky ReLU activation function in the discriminator with a ReLU activation function to solve the problem of not being able to recover the local details, which leads to blurred or unnatural visual effects. Third, the predictive performance of the different models is analyzed and compared. The experimental results demonstrate that improved ESRGAN can produce perceptually better SR images for face reconstruction than the original model within six Evaluation indices.

2 METHODOLOGY

2.1 Dataset Description and Preprocessing

The dataset used in this study, called human faces, is sourced from Kaggle (Dataset). It is a collection of 7219 high-resolution images with different resolutions useful for multiple use cases such as image identifiers, and classifier algorithms. The dataset carefully blends all prevalent racial and ethnic groupings, age ranges, and profile types in an effort to produce an objective dataset with a few GAN-generated photos to boost diversity and authenticity. The generator and discriminator were trained using the first 70% of the dataset as a training set and the second 30% as a validation set. Since each image in the dataset has a different resolution, the high-resolution images are randomly cropped to 128×128 pixels while the low-resolution are 32×32 pixels. The author similarly used a 50% probability of a horizontal flip and a 50% probability of a random 90-degree rotation to add variety to the data. A selection of the dataset's photos are shown in Fig. 1.

2.2 Proposed Approach

The focus of this proposed method for face restoration revolves around the ESRGAN model and innovative additional perceptual loss. This method is built upon the solid basis of GAN-based super-resolution, a well-known convolutional neural network, combined with

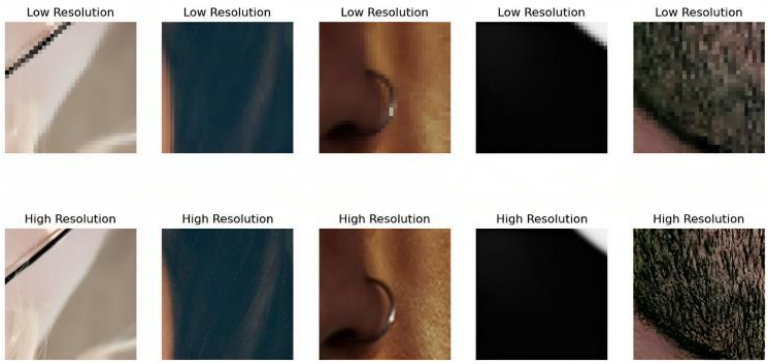

Figure 1: Images from the Human Faces dataset (Picture credit: Original).

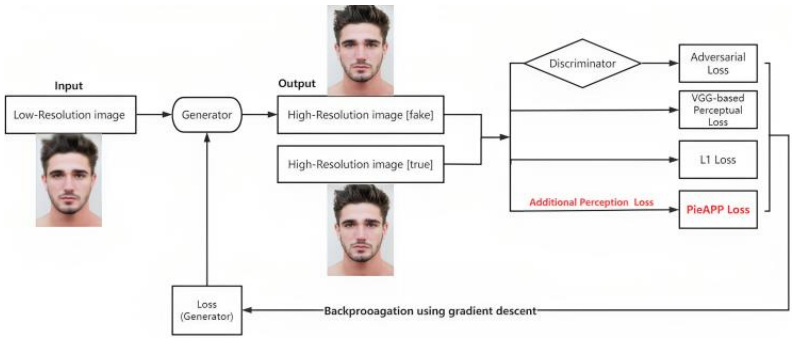

Figure 2: The pipeline of the model (Picture credit: Original).

the additional PieAPP loss. These technologies, when combined, will be able to make the local details of the image more realistic and the visual effect clearer and more natural in terms of human senses. It leads to better performance in the face resolution task. Fig. 2 below illustrates the structure of the system.

2.2.1 Network Architecture

The author has preserved the innovative modifications of ESRGAN (Wang et al 2018). To enhance the quality of the generated images in SRGAN, the architecture of the generator has introduced two key modifications. First, all Batch Normalization (BN) layers must be removed. The second involves replacing the traditional basic block with the recently proposed RRDB, which is an alternative. A multi-level residual network topology with numerous dense connections is integrated into the RRDB. When doing SR and deblurring, two PSNR-focused activities, it is better to omit BN layers. BN layers use the training dataset's estimated median and variance for testing, but they use the mean and variance within a batch for feature normalization during training. The problem emerges when there are large statistical differences between the training and testing datasets, as BN layers may create unfavorable artifacts and impair the model's capacity to generalize. BN layers are more likely to introduce these artifacts, according to actual studies, especially when the network is deep and working within a GAN framework. These artifacts may appear erratically during training iterations and in various situations, upsetting the goal of stable performance during training. Therefore, in order to provide stable training and constant performance, BN layers have been excluded from this investigation. Additionally, the elimination of BN layers improves generalization capacities, lowers computational complexity, and saves memory.

LeakyReLU functions are replaced by ReLU functions in each discriminator convolution layer. This research makes this change because ReLU functions lead to perceptually superior super-resolution results.

2.2.2 Loss Function

In the original ESRGAN model, the loss function comprises three distinct components: adversarial loss, L1 loss, and VGG based perception loss.

A competitive process between the Generator and the Discriminator is involved in adversarial loss. The Discriminator's job is to distinguish between created and actual images, while the Generator seeks to produce high-resolution images that closely resemble real ones. In an effort to reduce this loss and ultimately improve image quality, the Generator makes it difficult for the Discriminator to discriminate between the two.

The created high-resolution image's difference from the actual high-resolution image at the pixel level is measured as L1 loss. It determines and adds the absolute difference between each pixel in the two photos. L1 loss helps to maintain the detail of the generated image and reduces the noise in the generated image.

By feeding the produced image and the real image into a pre-trained deep convolutional neural network (a VGG network), the VGG-based perception loss is determined. It measures the difference between the feature representations of the two images in the network. Since the VGG network has been trained on large-scale image classification tasks, it captures the perceptual quality of the image. This perceptual loss helps to generate more realistic and visually pleasing images.

These three loss components work together to motivate the generator to produce high-quality super-resolution images. The generator learns to produce more realistic images with the adversarial loss; with the L1 loss, the details of the image are preserved; and with the perceptual loss, the perceptual quality of the image is improved. These three losses together help ESRGAN to generate higher-quality images.

This combination of loss functions proved to be effective, resulting in visually superior super-resolution images compared to using only L1 or L2 (MSE) loss, as seen in previous super-resolution methods. However, it's worth noting that the VGG-based perceptual loss may not be the optimal choice for super-resolution tasks, as the VGG network is originally trained for image classification purposes. Additionally, despite the incorporation of L1 loss, ESRGAN still faced challenges such as low PSNR and the presence of unpleasant visual artifacts, which negatively impacted the overall perceptual quality of the super-resolved images.

Therefore, the author tries to incorporate an extra mechanism that predicts perceptual image errors akin to human observers. This predicted error will then be employed to refine the generation process of our high-resolution images. Thus, a comparison of our improved loss function and the original loss function is as follows.

$$L_{original} = \alpha_1 L_{adv} + \alpha_2 L_{VGG} + \alpha_3 L_1 \quad (1)$$

$$L_{improved} = \alpha_1 L_{adv} + \alpha_2 L_{VGG} + \alpha_3 L_1 + \alpha_4 L_{PieAPP} \quad (2)$$

The PieAPP loss is used in the context of image super-resolution to assess how closely the created high-resolution image resembles the original high-resolution image. It does so by analyzing pairs of images and determining which one is preferred by human observers in terms of visual quality.

This loss function takes into account various aspects of image quality, including sharpness, clarity, and overall visual appeal, to guide the training of super-resolution models. By using PieAPP loss, models can learn to generate high-resolution images that are not only mathematically accurate but also visually pleasing to human observers. This enhances the perceptual quality of the super-resolved images, making them more realistic and satisfying to the human eye. Fig. 3 below illustrates the pre-trained PieAPP model. Function "F" is trained to map an image to its perceptual error concerning the reference.

2.3 Implementation Details

In the execution of the suggested model, several important aspects are underscored. For hyperparameters, the learning rate is established at 0.0001. A smaller learning rate ensures stable convergence but may require longer training time. The model runs for a total of 100 epochs in a batch size of 8. It is assumed that the model learns to focus on the facial features for the emotion classification task.

3 RESULTS AND DISCUSSION

In the conducted study, the original ESRGAN model and the model with PieAPP added as a loss function are trained with the same human faces dataset, other parameters remain the same. Face restoration images generated by the model are evaluated for six evaluation indexes. The Peak Signal-to-Noise Ratio (PSNR), one of these evaluation indices, measures image fidelity at the pixel level. SSIM focuses on luminance, contrast, and structure. Learned Perceptual Image Patch Similarity (LPIPS) represents Learned Perceptual Image Patch Similarity. Differential Mean Opinion Score (DMOS) quantifies subjective differences.

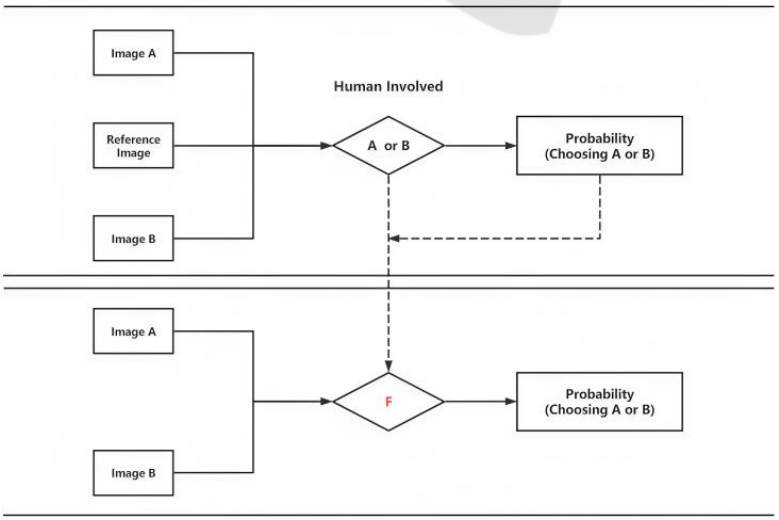

Figure 3: The pre-trained PieAPP model (Picture credit: Original).

Perceptual Index (PI) denotes Perceptual Index, while PieAPP signifies Perceptual Image-Error Assessment through Pairwise Preference. Notably, an asterisk symbol indicates that smaller values indicate superior performance. Comparative results of the two models are provided in Table 1.

Table 1: Comparative Results of Two Models.

	PSNR	SSIM	LPIPS*	DMOS*	PI*	PieAPP*
L1+VGG	15.99	0.8283	0.6867	76.3268	29.3040	133.3439
L1+VGG+PieAPP	24.67	0.9307	0.3614	41.4406	31.9565	67.8631

In terms of quantitative quality (PSNR and SSIM), the "L1+VGG+PieAPP" loss produced better results. This means clearer, sharper images with higher contrast and better structural similarity. Also, the improved model performs better on LPIPS DMOS and PieAPP metrics. An image with a low score is more likely to be accurate, harder for a human to tell apart, and boost user happiness. In addition, low scores help reduce storage and bandwidth requirements for image transmission. In addition, it creates a reconstruction image with more accurate texture features based on the results of subjective human eye observation. Different metrics are more applicable in different contexts, so it is often necessary to consider multiple metrics to fully assess image quality. The performance of the improved model is still inferior to the original model in terms of PI, and how to improve the performance in terms of PI will be a goal for the future. In summary, the improved ESRGAN model with an additional PieAPP loss has better performance for face restoration than the original ESRGAN model.

4 CONCLUSION

This study explored the fascinating field of face picture super-resolution reconstruction with the goal of improving the visual appeal and perceptual accuracy of facial images. The proposed method, an improved ESRGAN model integrated with an innovative PieAPP loss function, has been thoroughly examined and showcased as a powerful tool for face restoration. The loss function is proposed to analyze and refine the super-resolution process. Leveraging the removal of Batch Normalization layers and the incorporation of the RRDB in the generator's architecture, as well as introducing the PieAPP loss, the model is designed to produce clearer and more natural high-resolution facial images. The effectiveness of the suggested strategy is assessed by extensive experiments using a variety of evaluation indices. The results consistently demonstrated that our "L1+VGG+PieAPP" loss function outperforms the original ESRGAN model across various quantitative metrics, including PSNR, SSIM, LPIPS, DMOS, and PieAPP, resulting in superior perceptual image quality and user satisfaction. In the future, this research will turn its attention to optimizing loss functions as the primary research objective for the next stage. In an effort to improve the field of face image super-resolution and contribute to the creation of more accurate and aesthetically acceptable facial reconstructions, the emphasis will be on the in-depth investigation of alternative loss functions designed expressly for face image super-resolution.

REFERENCES

Y. Hong, U. Hwang, J. Yoo, et al. “How generative adversarial networks and their variants work: An overview,” ACM Computing Surveys (CSUR), vol. 52, 2019, pp. 1-43.

J. Lin, T. Zhou, Z. Chen, “Multi-scale face restoration with sequential gating ensemble network,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

M. Cao, Z. Liu, X. Huang, et al. “Research for face image super-resolution reconstruction based on wavelet transform and SRGAN,” 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, vol. 5, 2021, pp. 448-451.

T. Karras, M. Aittala, S. Laine, et al. “Alias-free generative adversarial networks,” Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 852-863.

N. Zhang, Y. Wang, X. Zhang, et al. “An unsupervised remote sensing single-image super-resolution method based on generative adversarial network,” IEEE Access, vol. 8, 2020, pp. 29027-29039.

J. A. Parker, R. V. Kenyon, D. E. Troxel, “Comparison of interpolating methods for image resampling,” IEEE Transactions on medical imaging, vol. 2, 1983, pp. 31-39.

T. Tong, G. Li, X. Liu, et al. “Image super-resolution using dense skip connections,” Proceedings of the IEEE international conference on computer vision, 2017, pp. 4799-4807.

C. Ledig, L. Theis, F. Huszár, 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, pp. 4681-4690.

Y. Choi, H. Park, “Improving ESRGAN with an additional image quality loss,” Multimedia Tools and Applications, vol. 82, 2023, pp. 3123-3137.

Dataset
https://www.kaggle.com/datasets/ashwingupta3012/human-faces.
X. Wang, K. Yu, S. Wu, et al. “Esrgan: Enhanced super-resolution generative adversarial networks,” Proceedings of the European conference on computer vision (ECCV) workshops, 2018.