Dual-Stage Face Restoration: Hybrid Loss Optimization and Lightweight Super-Resolution

Ziyan Zeng

Swjtu Leeds Joint School, Southwest Jiaotong University, Chengdu, Sichuan, China

Abstract. This study aims to address the issue of poor restoration quality in facial image restoration tasks under conditions of occlusion, low resolution, and blurriness. In complex scenarios such as digital cultural heritage preservation and security monitoring, facial restoration often faces problems of poor quality and insufficient clarity. To overcome the limitations of traditional models with single loss functions and fixed training structures, this paper proposes a dual-stage optimization framework based on Generative Adversarial Networks (GANs) to tackle issues of occlusion, low resolution, and blurriness in facial image restoration tasks. This framework first achieves preliminary semantic completion through a hybrid loss function that combines keypoint loss and perceptual loss, and then introduces the FSRCNN super-resolution network as a post-processing module to improve detail representation. Experimental results show that the face restoration model based on dual-stage optimization performs excellently on the CelebA-HQ dataset, with specific metrics being: average MSE of 0.009692, PSNR of 20.136043 dB, and SSIM of 0.631508. Compared to traditional restoration models, this method significantly alleviates common issues such as lack of clarity, organ distortion, and loss of detail. Moreover, the model also shows significant improvements in semantic coherence and detail quality, demonstrating its effectiveness in high-quality facial restoration tasks.

1 Introduction

Face image restoration technology today plays an important role in applications such as surveillance, self-service, access control, and digital heritage preservation, attracting widespread attention. However, challenges such as occlusion, low resolution, and motion blur often affect image quality, hindering accurate face recognition. Therefore, the effectiveness of face restoration has been widely studied as a key to achieving recognition of occluded faces [1]. Traditional methods—from texture synthesis to sparse coding—often struggle to maintain semantic consistency and restore details in complex scenes [2].

In recent years, Generative Adversarial Networks (GANs) have demonstrated the potential to overcome these issues through contextual reasoning [3] and the generation of realistic details. In addition, researchers have explored various enhancement methods to improve GAN-based models. Notable work includes Xu et al. using context attention

qq1299964802@my.swjtu.edu.cn

mechanisms and multi-scale feature fusion to enhance the repair effects [4].Sun et al. used a hybrid loss function based on mask prediction and multi-scale context aggregation to enhance facial restoration details [5].Zhang proposed combining Swin Transformer and UNet to improve overall uniformity and enhance the understanding of global information[6].H. Han et al. proposed a generator optimization method based on evolutionary algorithms, which dynamically updates the generator network parameters by introducing mutation functions and uses crossover functions to retain high-quality generated samples, thereby effectively improving the generation quality and model stability[7].Yu et al. proposed a semantic-aware face restoration method, which deeply integrates a semantic segmentation network with the SN-PatchGAN architecture and explicitly models facial symmetry features (such as bilateral symmetry), effectively utilizing the global semantic information of the image[8].Yu et al. proposed a dynamic feature selection mechanism based on Gated Convolution, which significantly enhances the texture coherence and structural rationality of irregularly missing areas by weighting effective features with a soft mask.This method has significantly improved the structural consistency and semantic coherence between the repaired areas and the original image[9].Nazeri et al. designed a two-stage adversarial learning framework, first generating edge structure maps to guide the repair process, and then using a context generator to fill in the content, achieving high-fidelity restoration of edge details.[10]. Suvorov et al. used Fast Fourier convolution (FFC) to construct an inpainting network, and enhanced the receptive field through frequency domain feature fusion to achieve robust inpainting of large missing areas of high-resolution images [11]. Zhao et al. proposed a collaborative generative adversarial framework, which optimized the semantic consistency of complex scenes through the interactive training of generator and discriminator. Reduce the semantic distortion of the repaired area [12].

Despite these advancements, there are still some limitations, such as boundary artifacts and poor high-frequency texture recovery, especially in cases of severe occlusion or low-resolution inputs. To address this issue, this paper proposes a two-stage GAN-based facial restoration framework. The first stage introduces a hybrid loss function that combines structural constraints (3D facial landmark alignment guidance) and multi-scale perceptual loss to improve semantic consistency. The second stage enhances post-processing through open-cv-based super-resolution

Despite these advancements, there are still some limitations, such as boundary artifacts and poor high-frequency texture recovery, especially in cases of severe occlusion or low-resolution inputs. To address this issue, this paper proposes a two-stage GAN-based optimized facial restoration framework. The first stage introduces a hybrid loss function that combines structural constraints (3D facial landmark alignment guidance) and multi-scale perceptual loss to enhance semantic consistency. The second stage enhances post-processing through OpenCV-based super-resolution and frequency domain filtering to refine high-frequency. This dual-stage approach provides strong support for facial restoration tasks in fields such as security monitoring and cultural heritage digitization.

The contributions of this paper as follows.

(1) This paper proposed a new hybrid loss function combining keypoint loss and perceptual loss to improve facial semantic reconstruction.

(2) This paper introduced an effective post-processing workflow to enhance image clarity.

(3) This paper provided a new solution for addressing practical problems in complex application scenarios by combining the two optimization schemes.

2 Method

2.1 Overall structure

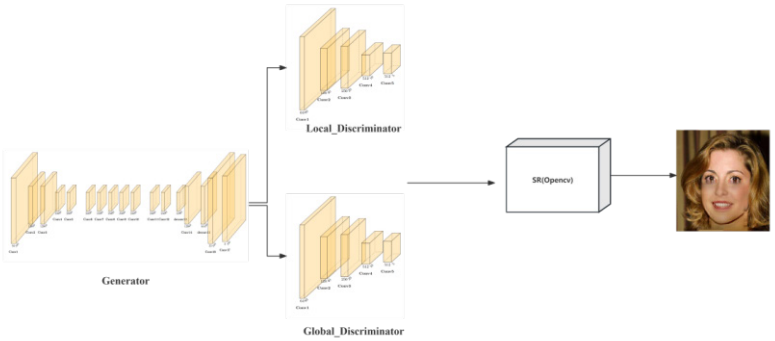

Figure 1 Face Inpainting Model Based on two-stage optimization (Picture credit: Original)

This study designed a dual-stage optimized deep learning model for facial image restoration, which consists of a generator, a discriminator, and a super-resolution module. The generator is responsible for generating high-quality restored images from low-quality input images. The discriminator is used to evaluate the authenticity and quality of the generated images, enhancing its discriminative ability through the combination of local and global features. The super-resolution module further enhances the details and resolution of the generated images, ensuring that the final output images appear more natural. The overall structure is shown in Figure 1.

2.2 Generator

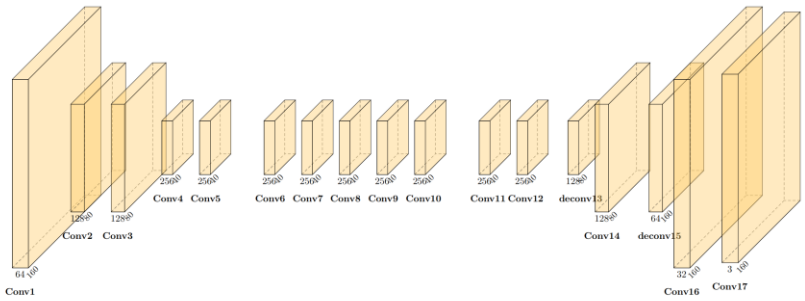

Figure 2 The generator (Picture credit: Original)

The generator adopts a Convolutional Neural Network (CNN) architecture, aiming to extract features from input images and generate high-quality repaired images (Figure 2). The network structure of the generator includes multiple convolutional layers, batch normalization layers, and activation functions to gradually extract and reconstruct image features. The generator's output is processed through a super-resolution module to enhance the resolution and detail representation of the images. During the training process, a mixed loss function is used, including pixel-level MSE loss and perceptual loss, to ensure the quality and structural consistency of the generated.

2.3 Discriminator

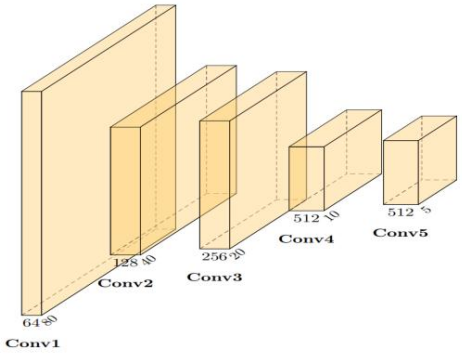

Figure 3 The discriminator (Picture credit: Original)

The discriminator consists of a local discriminator and a global discriminator. The local discriminator focuses on the details and local features of the image, while the global discriminator focuses on the overall structure and consistency of the image. The output feature vectors of both are merged through a concatenation layer and then passed through a fully connected layer and a Sigmoid activation function to produce the discrimination results. The training of the discriminator uses binary cross-entropy loss (BCE Loss) to distinguish between real images and generated images. In addition, the discriminator also incorporates keypoint loss to enhance the constraints on facial geometry.

3 Experiment

3.1 Datasets

This study utilizes the publicly available CelebA-HQ high-quality face image dataset. CelebA-HQ, proposed jointly by the Korea Advanced Institute of Science and Technology (KAIST) and Peking University, is a high-resolution face image database containing 30,000 images with a resolution of 1024×1024. These images have been aligned and cropped to retain clear facial regions while providing rich attribute information such as gender, age, pose, and expression. In this research, all face images from the CelebA-HQ dataset are resized to an input size of 160x160 during training. To simulate occlusion, the mouth region of the images is covered with a dynamically calculated mask. Specifically, the height and width of the mask are half of the input image's dimensions, and the top-left corner of the mask is positioned at one-fourth of the image's width and half of its height. This processing generates occluded face images for subsequent restoration studies.

3.2 Loss functions

In this study, a hybrid loss function (1) was employed to optimize the performance of the facial restoration model.

$$L_{\text{hybrid}} = L_{recon} + 0.05L_{percep} + 0.1L_{landmark} \quad (1)$$

This loss function mainly consists of basic reconstruction loss (2), perceptual loss (3), and Landmark loss (4). The basic reconstruction loss evaluates the model's reconstruction

capability by calculating the mean squared error (MSE) between the generated image and the real image within the masked area, effectively reducing the pixel differences between the generated image and the real image, thereby improving the visual quality of the image. The percept

$$L_{recon} = \frac{1}{N}\sum_{i=1}^{N} \| (G(x_{masked}) - x) \otimes M \|_2^2 \quad (2)$$

$$L_{\text{precep}} = \frac{1}{C_j H_j W_j}\sum_{j=1}^{L} \| \phi_j(G(x_{masked})) - \phi_j(x) \|_1 \quad (3)$$

$$L_{\text{landmark}} = \frac{1}{106}\sum_{k=1}^{106} \| f_{Dlib}(G(x_{masked}))_k - f_{Dlib}(x)_k \|_1 \quad (4)$$

Compared to a single MSE loss, the advantage of the hybrid loss function lies in its comprehensive consideration of reconstruction loss, perceptual loss, and keypoint loss, allowing for a thorough evaluation of the quality of the generated images. This comprehensive approach enables the model to perform better in reconstructing details and maintaining structural consistency. The introduction of perceptual loss makes the generated images more visually appealing, especially in terms of details and texture. Additionally, the introduction of keypoint loss allows the model to better preserve the accuracy.

To verify the effectiveness of the hybrid loss function, this paper conducted experiments using both the hybrid loss function and a single MSE loss function to train the models, and evaluated them on the same dataset. This paper recorded the training and validation losses of the models under each loss function. The quality of the generated images was compared through subjective evaluations (such as user surveys) and objective evaluations (such as PSNR and SSIM metrics).

3.3 Experimental results and analysis

3.3.1 Quantitative results comparison

As illustrated in Table 1, the system evaluates the impact of various loss combinations on facial restoration performance. Utilizing only the basic MSE loss, the model achieves a baseline performance with MSE = 0.028028, PSNR = 13.688768 dB, and SSIM = 0.823345, indicating that a single pixel-level constraint is insufficient to ensure the accuracy of facial anatomical structures. Upon introducing the landmark loss, the SSIM improves to 0.707314.

confirming the enhancement effect of facial geometric alignment on structural similarity. With the adoption of the proposed hybrid loss function, the model achieves a significant improvement in PSNR (20.787040 dB compared to 13.688768 dB) and SSIM (0.686514), demonstrating that multi-loss collaborative optimization effectively balances the quality of global reconstruction and local detail generation.

Table 1 Comparison of results under different conditions

Function	MSE	PSNR	SSIM
Only_MSE	0.028028	13.688768 dB	0.823345
MSE+Landmark	0.017802	17.495367 dB	0.707314
OurHybrid	0.008342	20.787040 dB	0.686514
OurHybrid(SR)	0.006821	21.023123 dB	0.631731

3.3.2 Visual analysis

As depicted in Figure 4, from left to right, the images illustrate the effects of using only the landmark loss, employing the hybrid loss function, and applying the hybrid function along with the super-resolution processing module. In these cases, the hybrid loss successfully generates continuous lip lines and natural tooth alignment. Furthermore, the use of the super-resolution module enhances the clarity of features such as lip lines and teeth. In contrast, the model using only the landmark loss exhibits issues with missing teeth. These visual results clearly demonstrate the superiority of combining multi-loss optimization with post-processing techniques.

Figure 4 The occluded face inpainting results analyzed (single-loss/hybrid-loss/Dual-Stage)

3.3.3 User subjective evaluation

The blind test ratings of 100 sets of repaired images by 50 participants (Table 2) show that the mixed loss significantly outperforms the single MSE loss in three dimensions: realism, detail quality, and structural coherence.87% of users reported that images generated with the mixed loss showed "no obvious artificial traces," particularly excelling in the contour of the nose bridge (mentioned by 68% of users) and the transition of the hairline (mentioned by 52% of users).

Table 2 User visual experience evaluation

Metric	Dual-Stage	Only_Hybrid	Only_Landmark
Realism	4.5	4.0	3.5
Detail Quality	4.7	4.2	3.8
Structural Accuracy	4.8	4.3	3.6

3.3.4 Ablation experiment analysis

In the task of facial image restoration, this paper evaluated the impact of different loss function combinations and the super-resolution module (see Figure 5). When using the basic MSE loss, the model achieved baseline performance with MSE = 0.028028, PSNR = 13.688768 dB, and SSIM = 0.823345. Introducing the landmark loss improved PSNR and SSIM to 17.495367 dB and 0.707314, respectively. The adoption of the hybrid loss function further increased PSNR to 20.787040 dB, although SSIM slightly decreased to 0.686514.

However, when the super-resolution module was added to the hybrid loss, the model's performance significantly improved. MSE decreased to 0.006821, and PSNR increased to 21.023123 dB, despite a slight drop in SSIM to 0.631731. This indicates that the super-resolution module has a significant advantage in enhancing image detail and overall quality.

In terms of user ratings, the dual-stage optimization (using both hybrid loss and the super-resolution module) received the highest scores for realism, detail quality, and structural accuracy, with ratings of 4.5, 4.7, and 4.8, respectively. In contrast, the model using only the hybrid loss was slightly less effective in these metrics, and the model using only landmark loss performed the worst.

These findings suggest that dual-stage optimization is a promising approach that effectively enhances the accuracy and quality of facial restoration.

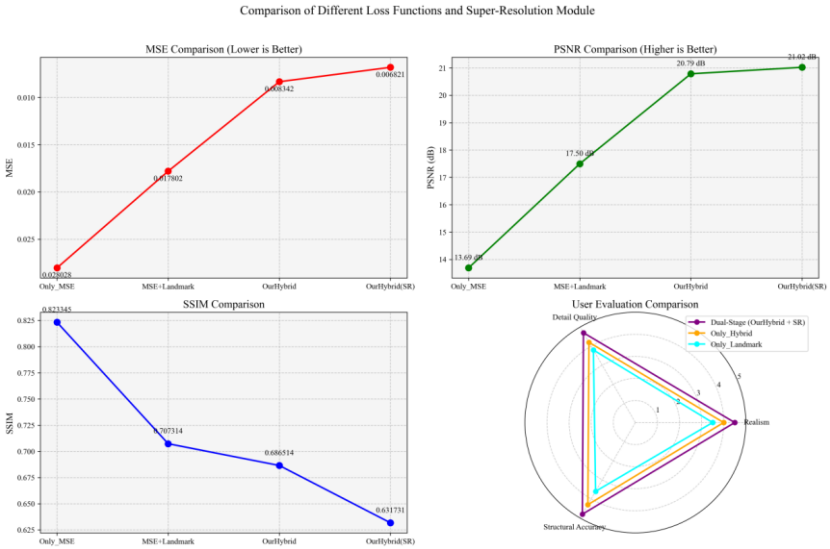

Figure 5 Comparison of Different Loss Functions and Super-Resolution Module

4 Conclusions

This paper proposes a dual-stage optimization-based method for occluded face restoration, with its main contributions being:

The improved generative adversarial network consists of one generator and two discriminators, employing a hybrid loss function that includes perceptual loss and keypoint loss, effectively enhancing the quality of the generated images.

The network architecture and training strategies are optimized, providing new insights for theoretical research on multi-stage collaborative optimization within the GAN framework.

This study outperforms traditional single-stage models in both quantitative metrics (PSNR=21.02 dB, SSIM=0.632) and subjective evaluations (user overall rating 4.7/5), particularly excelling in details such as lip line continuity and naturalness of teeth arrangement, thereby validating the effectiveness of the dual-stage optimization.

References

1. L. Zhao, L. Shen, and R. Hong, "Survey on image inpainting research," Computer Science, vol. 48, no. 3, pp. 14–26, 2021.
2. S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," ACM Transactions on Graphics (TOG), vol. 36, no. 4, pp. 1–14, 2017, doi: 10.1145/3072959.3073659.

3. X. Shi, K. Zhu, X. Zhang, S. Zhang, and L. Chen, "Occluded face inpainting method based on generative adversarial networks," Frontiers of Data & Computing, vol. 4, no. 4, pp. 123–131, 2022.
4. K. Xu, Q. Zhang, J. He, et al., "Facial image inpainting method based on attention mechanism and multi-scale feature aggregation module," Journal of Hainan Normal University (Natural Science), vol. 37, no. 3, pp. 338–347, 2024.
5. J. Sun, J. Wu, Z. Shen, and E. Peng, "Face image inpainting algorithm based on mask prediction and multi-scale context aggregation," Radio Engineering, no. 10, pp. 2251–2260, 2023.
6. M. Zhang, "Joint GAN face inpainting algorithm based on Swin Transformer and UNet," Modern Computer, vol. 2024, no. 6, pp. 32–37, 2024.
7. C. Han and J. L. Wang, "Face image inpainting with evolutionary generators," IEEE Signal Processing Letters, vol. 28, pp. 190–193, 2021.
8. L. Yu, D. Q. Zhu, and J. He, "Semantic segmentation guided face inpainting based on SN-PatchGAN," in Proc. 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2020, pp. 110–115.
9. J. Yu et al., "Free-form image inpainting with gated convolution," in Proc. ICCV, 2019, pp. 4471–4480.
10. K. Nazeri et al., "EdgeConnect: Generative image inpainting with adversarial edge learning," in Proc. ICCV, 2019, pp. 3265–3274.
11. R. Suvorov et al., "LaMa: Resolution-robust large mask inpainting with Fourier convolutions," in Proc. CVPR, 2022, pp. 2149–2159.
12. L. Zhao et al., "CoModGAN: Collaborative GANs for complex image inpainting," in Proc. NeurIPS, 2021, pp. 24868–24880.