SCIENTIFIC Publishing Limited

Digital Technologies Research and Applications

http://ojs.ukscip.com/index.php/dtra

Article

Generative Adversarial Lightweight Classroom Face Recognition and Hierarchical Reshaping Optimization Model

Xuliang Yang [1,2,*], and Rodolfo C. Raga Jr. [3]

[1] University and Urban Integration Development Research Center, Dongguan City University, Dongguan 523000, China

[2] College of Computing & Information Technologies, National University, Manila 0900, Philippines

[3] College of Computer Studies and Engineering, Jose Rizal University, Mandaluyong 0900, Philippines

[*] Correspondence: yangxuliang@dgcu.edu.cn

Received: 30 October 2025; **Revised:** 10 December 2025; **Accepted:** 27 January 2026; **Published:** 10 February 2026

Abstract: To address the significant decline in face recognition performance caused by low resolution, high noise, and complex degradation factors in security surveillance scenarios, this paper proposes a joint optimization framework that integrates a Transformer and a Generative Adversarial Network (GAN). The innovation of this framework lies in: (1) designing the Face Reconstruction Transformer (FRFormer), which integrates a hierarchical window attention mechanism and a multi-level feature pyramid structure, enhancing the ability to retain identity features through local-global collaborative modeling; (2) constructing the GFP-GAN reconstruction model, which combines pre-trained face priors and degradation removal modules, and utilizes adversarial training to improve image authenticity and detail restoration. Experiments show that when the input is 32 × 32 pixels, the PSNR of GFP-GAN is increased by more than 8 dB, and the SSIM reaches 0.953; FRFormer achieves recognition accuracies of 99.58% and 96.31% on the LFW and AgeDB-30 benchmarks, respectively, which are 0.08 and 0.13 percentage points higher than those of Swin Transformer. Ablation experiments verify the effectiveness of the window attention mechanism and hierarchical reconstruction strategy, especially in noise suppression and cross-pose recognition tasks. This framework has broad application potential in degraded visual conditions, such as biometric recognition and medical image analysis, and provides an end-to-end solution for low-quality face recognition.

Keywords: Generative Adversarial Networks (GANs); Transformer Architecture; Face Reconstruction; Window Attention Mechanism

1. Introduction

Facial recognition has become a core biometric technology in applications such as public safety, financial services, and human-computer interaction [1–3]. Although substantial progress has been made under controlled imaging conditions, recognition performance often deteriorates significantly in real-world scenarios where facial images are affected by low resolution and various forms of noise [4]. In the monitoring environment, factors such as long capture distance, suboptimal lighting, and compression related artifacts further complicate reliable identity feature extraction, thereby limiting the effectiveness of traditional recognition systems [4,5].

The adoption of deep learning has greatly advanced facial recognition research, enabling models based on convolutional neural networks to achieve performance close to that of humans when high-quality facial images are available. However, existing research has largely focused on recognition tasks under ideal imaging conditions,

and there are still significant technical bottlenecks in robust face recognition for low-quality facial images. Traditional super-resolution reconstruction methods can improve image resolution to a certain extent but struggle to effectively remove complex noise and restore identity-discriminative features [6]. Meanwhile, deep learning-based methods still have limitations in noise modeling and feature preservation, often leading to distorted details and loss of identity information in reconstructed images [6].

With the breakthrough progress of the Transformer architecture in the field of computer vision, its powerful global modeling capabilities have provided new insights for low-quality image processing. Vision Transformers, through the self-attention mechanism, can effectively capture long-distance dependencies, demonstrating superior performance to traditional convolutional neural networks in image restoration tasks. However, existing research has largely focused on natural image reconstruction, and dedicated restoration models for highly structured facial images remain to be explored. Moreover, how to organically integrate image reconstruction with face recognition tasks to achieve end-to-end identity feature enhancement has become a critical issue that urgently needs to be addressed.

This study addresses the challenge of recognizing low-resolution and high-noise face images by proposing an innovative solution that integrates Transformers with Generative Adversarial Networks (GANs). By constructing the Face-Reconstructor Transformer (FRFormer) network architecture, it innovatively combines the hierarchical window attention mechanism of Swin Transformer with adversarial training strategies to achieve enhanced reconstruction of the identity features in degraded face images. Based on the above observations, GFP-GAN is integrated into a restoration module to bridge the image space domain and identity feature space, thereby achieving a unified pipeline that connects facial image restoration with subsequent feature enhancement. Unlike existing methods that treat reconstruction and recognition as loosely coupled tasks, the proposed framework emphasizes their joint optimization. In particular, a window-based hierarchical attention mechanism is used to combine local detail preservation with global structure modeling, while a perceptual degradation adversarial training strategy is employed to improve robustness under complex and mixed noise conditions. In addition, identity preservation constraints were explicitly introduced in the optimization objectives, thereby improving reconstruction quality and recognition performance in a coordinated manner.

From a methodological perspective, this study provides empirical evidence that Transformer-based architectures can effectively adapt to degraded facial images when combined with appropriate attention design and adversarial supervision. From an application perspective, the proposed framework addresses the practical challenges associated with low-quality facial input, making it relevant to deployment scenarios such as intelligent security systems and mobile identity verification. Extensive experiments conducted on multiple public benchmarks have shown consistent improvements in image restoration metrics (such as PSNR and SSIM) and recognition accuracy compared to representative baselines based on CNN and Transformer.

2. Related Work

2.1. Traditional Facial Image Processing Methods

With the advancement of deep learning, research on low-resolution and high-noise face processing is constantly expanding, resulting in various methods to address degraded visual conditions. The existing methods can be roughly divided into three categories: super-resolution-based reconstruction, robustness-oriented noise modeling, and cross-modal feature learning, each emphasizing different architecture designs and training strategies.

2.2. Joint Restoration Method Based on Generative Adversarial Network

VQFR [7] is a face restoration method that utilizes codebook priors. By representing facial structures in a discrete latent space, this method reduces the impact of noise while preserving identity-related information. Specifically, a two-stage training strategy is adopted: In the first stage, the VQ-VAE model is trained to learn the face structure codebook. In the second stage, a deformable attention module is introduced to align the degraded features with the codebook features. Experiments show that the PSNR of this method reaches 28.7dB on the CelebA-Test and WebPhoto-Test datasets, which is 1.2dB higher than that of GFP-GAN. PSFR-GAN [8] innovation in introducing facial image resolution as guide information. The network architecture consists of three key modules: the parsing and prediction module generates the facial region segmentation map, the texture transfer module injects parsing

information through spatial adaptive normalization (SPADE), and the multi-scale discriminator ensures the consistency of details. Experiments on the LFW low-score dataset (16 × 16 pixels) show that this method improves the recognition accuracy from 58.3% of the traditional method to 82.6%. HiFaceGAN [9] proposed a hierarchical feature fusion mechanism, a design generated by coarse to fine three phase network: global structure generator (128 × 128), the local details enhancer (256 × 256), and grain refinement (512 × 512). A gated attention module is introduced at each stage to dynamically fuse multi-scale features. Tests on the synthetic noise dataset show that this method achieves 0.913 in the SSIM index, which is 9% higher than that of the common cascading network.

2.3. Robust Feature Learning Based on Transformer

TransFace [10] applied the Vision Transformer to low-quality face recognition for the first time. Its core innovation lies in deformable Position Coding (DPE) and Locally Enhanced Attention (LEA): DPE predicts the position offset through deformable convolution and dynamically ADAPTS to the facial geometry structure; LEA incorporates local gradient features when calculating attention to enhance robustness against noise. Experiments on the CFP-FP cross-pose dataset show that this method achieves a recognition rate of 92.4% under the 1/8 downsampling condition, which is 14.6% higher than ResNet-50. Noise-Aware ViT [11] training strategy is put forward. The network contains a dual-path architecture: The main path uses standard ViT to process degraded images, and the auxiliary path predicts the noise distribution map through the noise estimation module. The two paths interact through feature gating in the multi-layer Transformer Block and dynamically adjust the attention weights. Under the condition of synthetic Gaussian noise ($\sigma = 25$), this method achieved a Rank-1 accuracy rate of 76.8% in the MegaFace Challenge, demonstrating excellent noise robustness. SwinFIR [12] will teach Transformer combined with frequency domain to learn. Design a frequency-aware shift window mechanism to calculate the attention weight in the frequency domain space: Decompose the features into low-frequency components (identity information) and high-frequency components (noise/detail) through DCT transformation, and use dynamic filters for frequency band selection. In the NTIRE2020 Real-world Super Resolution Challenge, the method achieved a PSNR of 31.2 dB, which is 2.4 dB higher than the conventional CNN method.

2.4. Cross-Modal Joint Learning Method

DualPath-RCNN [13] constructed a dual-path feature interaction network. The visible light path adopts EfficientNet to extract texture features, and the near-infrared path uses lightweight MobileNet to extract illumination invariant features. By constraining the feature space alignment through cross-modal contrastive learning loss (CMCL), the recognition accuracy on the low-illumination face dataset DarkFace reaches 78.3%, which is 21.5% higher than that of the single-modal method. CycleTransGAN [14] proposed a cycle Transformer-based consistency GAN framework. The multi-head cross-attention mechanism is introduced in the generator design to achieve feature mapping from the noise domain to the clear domain. The discriminator adopts a multi-scale ViT structure to simultaneously evaluate image quality and identity consistency. In the cross-resolution face matching task, the method reached 89.7% on the IJB-C dataset TAR@FAR = 0.1%, which is 12.3% higher than ArcFace.

2.5. Self-Supervised Pre-Training Method

Mae-face [15] introduced the mask autoencoder (MAE) into Face pre-training. An asymmetric codec architecture is adopted: the encoder processes 25% of the visible blocks (including key facial areas), and the decoder reconstructs the complete face. After pre-training, in the low-score face recognition task with only 10% of labeled data, this method achieved an accuracy rate of 91.2% on AgeDB-30, which was 23.6% higher than the baseline of supervised learning. This study presents a design comparison and measurement of a combined optimization framework for ContraFace [16]. A dynamic memory bank is built to store feature prototypes and enhance feature discriminability through Hybrid Negative Sample Generation (MNSG). Under the condition of synthetic noise (Gaussian mixture salt-and-pepper noise), this method maintains a recognition rate of 98.1% on the LFW dataset, and the noise robustness is improved by 17.8% compared with the traditional contrastive learning.

2.6. Comparative Analysis

It can be known from **Table 1** that Bicubic interpolation achieves a PSNR of only 23.2 dB on the CelebA-HQ dataset, while SRGAN [17] reaches 26.7 dB. In the LFW low-resolution subset test (32 × 32 pixels), traditional LBP methods have an identification rate of 58.3%, whereas ResNet-50 [18] improves this to 89.6%. Specifically, GAN-based methods reduce the EER from 12.4% to 5.8% in cross-modal face recognition tasks through adversarial training mechanisms [19,20].

Table 1. Compare the performance of different image super-resolution methods.

Method Type	PSNR(dB)	SSIM	Recognition Rate(%)
Bicubic interpolation	23.2	0.72	58.3
SRCNN	24.8	0.81	82.1
SRGAN	26.7	0.89	91.4
Transformer	27.9	0.92	95.6

The existing methods still face the following challenges in low-resolution and high-noise face processing: 1) The existing GAN-based methods are prone to artifacts under complex noise conditions, and the identity retention ability needs to be improved [1]; 2) The computational complexity of the Transformer model limits its application in mobile terminals [21]; 3) Cross-modal methods have strict requirements for data alignment and are limited in practical scenarios. The FRFormer proposed in this paper reduces the computational cost to 42% of that of the Swin Transformer while ensuring identity consistency through the local-global attention fusion mechanism and lightweight design, providing a new idea for real-time low-quality face processing.

Although existing methods have made progress, there are still some challenges in low-quality facial recognition. Many methods heavily rely on high-resolution pre-trained models, which often lead to significant performance degradation when the input resolution drops to extreme levels (e.g., below 16 × 16 pixels). In addition, current global local feature fusion strategies are often insufficient to maintain identity consistency in severely degraded situations, and limited dynamic noise modeling further reduces robustness in motion blur and mixed noise modes [1]. To address these issues, FRFormer introduces window based self attention in the hierarchical feature pyramid, providing more effective multi-scale representation learning for severely degraded facial inputs. Therefore, the proposed network achieved measurable improvement in recognition accuracy under extremely low resolution conditions while maintaining identity consistency [22].

Based on this design, the entire framework integrates the image restoration module with a dedicated facial recognition backbone. By combining global local feature modeling with adversarial supervision, this method aims to improve recognition reliability in complex and unconstrained imaging scenes.

3. Materials and Methods

3.1. Image Restoration via GFP-GAN

3.1.1. Degradation Removal Module

Using multi-scale convolutional networks to model noise characteristics and extract unique features, and gradually reducing degradation effects through cascaded residual blocks, this design helps to transform low-quality inputs into a latent feature space where noise components are minimized while semantic facial information is more effectively preserved.

3.1.2. Generative Adversarial Network

A pre-trained facial GAN is incorporated as a structural prior to guide the restoration process. The generator follows a U-Net–like design that supports multi-resolution feature fusion across different scales. Restoration is performed in a hierarchical manner through four successive up-sampling stages, each combining a Swin Transformer block with a PixelShuffle operation, while skip connections are preserved to maintain low-frequency structural information.

During decoding, a channel and spatial attention mechanism based on CBAM is applied to emphasize visually critical facial regions, including the eyes, nose, and mouth. The discriminator adopts a multi-scale PatchGAN architecture, enabling local texture authenticity to be assessed at different spatial resolutions through a Markovian discrimination strategy.

The adversarial loss function is designed as Equation (1):

$$\mathcal{L}_{adv} = \mathbb{E}[\log \mathrm{D}(\mathrm{I_{HR}})] + \mathbb{E}[\log(1 - \mathrm{D}(\mathrm{G}(\mathrm{I_{LR}})))] \tag{1}$$

3.1.3. Composite Loss Function

Identity Preservation Loss: A pre-trained ArcFace model is employed to encourage consistency between the identity representations of the generated images and their corresponding high-quality references, as defined in Equation (2).

$$\mathcal{L}_{id} = ||\phi(\mathrm{G}(\mathrm{I_{LR}})) - \phi(\mathrm{I_{HR}})||_2^2 \tag{2}$$

Structural similarity loss: Measuring perceptual quality through the LPIPS metric in Equation (3):

$$\mathcal{L}_{perc} = \sum_{\mathrm{l}} \lambda_{\mathrm{l}} ||\psi_{\mathrm{l}}(\mathrm{G}(\mathrm{I_{LR}})) - \psi_{\mathrm{l}}(\mathrm{I_{HR}})||_1 \tag{3}$$

This indicates the feature extractor of the VGG19's 19th layer.

3.2. Face Recognition Method Based on FRFormer

3.2.1. Enhanced Transformer Architecture

The FRFormer, an improvement on the Swin Transformer, includes: hierarchical feature encoding with a 4-stage feature pyramid structure, each stage containing 2 Swin Blocks, with window sizes progressively increasing from 8 × 8 to 32 × 32 to capture multi-granularity features; dynamic position encoding by introducing a learnable relative position bias matrix, where is the window radius and is the number of attention heads, enhancing spatial relationship modeling capabilities.

3.2.2. Attention Mechanism Optimization

Local-global attention fusion: Introduce a cross-window information interaction module based on SW-MSA, using deformable convolution to generate offsets, thereby achieving dynamic adjustment of receptive fields in Equation (4).

$$\mathrm{DeformAttn}(\mathrm{Q,K,V}) = \sum_{\mathrm{k}} \mathrm{A_{qk}} \cdot \mathrm{V}(\mathrm{p_k} + \Delta \mathrm{p_{qk}}) \tag{4}$$

Channel attention reweighting: After the MLP layer, integrate the SE module to generate channel weights through global average pooling, enhancing the expression of discriminative features.

3.2.3. Metric Learning Strategies

Adopting an improved CurricularFace loss function, a curriculum learning mechanism is introduced to dynamically adjust the weights of difficult samples in Equation (5).

$$\mathcal{L}_{metric} = -\log \frac{e^{\mathrm{s}(\cos\theta_{\mathrm{y_i}} - \mathrm{m})}}{e^{\mathrm{s}(\cos\theta_{\mathrm{y_i}} - \mathrm{m})} + \sum_{\mathrm{j} \neq \mathrm{y_i}} e^{\mathrm{s}\cos\theta_{\mathrm{j}}}} \tag{5}$$

The dynamic boundary parameters increase linearly with each training round.

3.3. Collaborative Training Strategy

3.3.1. End-to-End Optimization

Develop a dual-path training framework where the image restoration module and the recognition module share a common low-level feature extractor, and balance the losses of both tasks using gradient normalization algorithms in Equation (6).

$$g_{shared} = \frac{g_{rec}}{||g_{rec}||} + \frac{g_{id}}{||g_{id}||} \tag{6}$$

Here, they represent the gradients of the repair loss and the recognition loss, respectively.

3.3.2. Data Augmentation

Design a hybrid degradation model to simulate the real low-quality image degradation process in Equation (7).

$$I_{LR} = \downarrow_s (I_{HR} \otimes k) + n_\sigma \tag{7}$$

It represents down-sampling for the scale factor, with a random motion blur kernel and Gaussian noise.

The recognition accuracy of the proposed framework reached 98.7% on benchmark datasets such as LFW and CFP-FP. Compared with traditional methods, this represents a 12.3% improvement, indicating the effectiveness of this method for low-quality face recognition.

3.4. Experimental Setup and Implementation

This section provides a detailed description of the experimental configuration of the proposed framework for clarity and reproducibility, including network architecture design, hyper parameter selection, training program, and ablation settings.

3.4.1. Network Architecture

The FRFormer model is constructed as a four-level hierarchical Transformer, with each stage consisting of two Swin Transformer blocks, totaling eight blocks. This architecture supports multi-scale face representation learning by gradually increasing the attention window size from 8 × 8 to 16 × 16, and then increasing it to 32 × 32 in consecutive stages. The corresponding embedding sizes are 96, 192, 384, and 768, and the number of attention heads is set to 3, 6, 12, and 24. The GFP-GAN restoration module consists of a degradation removal network followed by a GAN-based generator–discriminator framework. The generator adopts a U-Net–like structure with four up-sampling stages, while the discriminator follows a multi-scale PatchGAN design.

3.4.2. Training Configuration

All experiments are implemented using PyTorch 1.12.1. The models are trained using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 1×10^{-3} and decayed using a cosine annealing schedule. The batch size is fixed at 256, and the total training duration is 100 epochs. Training is conducted on an NVIDIA A100 GPU.

3.4.3. Training Time and Computational Cost

Under the above settings, the average training time per epoch is approximately 14 minutes on the A100 platform. During inference, the full framework achieves approximately 23 FPS on an A100 and 11 FPS on an RTX 3060 GPU. The parameter size of GFP-GAN is approximately 21 million, while FRFormer contains around 48 million parameters. Ablation configurations. For ablation studies, we systematically modify individual components while keeping all other settings unchanged. Specifically, GFP-GAN is replaced by bicubic interpolation or a CNN-based restoration network without adversarial training; window-based attention is substituted with global self-attention or local-only attention; and the hierarchical depth is reduced by removing one or more Transformer stages. All ablation experiments were conducted under the same training plan and hyperparameter configuration to maintain consistency in comparison.

3.5. Computational Complexity and Deployment Analysis

The reasoning delay, model complexity and actual deployment constraints are analyzed, and the computational cost and deployment feasibility of the proposed framework are tested.

3.5.1. Latency Analysis

The measurement of inference latency is conducted on two platforms: an NVIDIA A100 GPU and an RTX 3060 GPU. With a batch size of one and an input resolution of 112 × 112, the framework operates at a rate of approximately 23 frames per second (FPS) on the A100 and 11 FPS on the RTX 3060. These results indicate that the proposed method is suitable for real-time or near–real-time processing in server-side deployment scenarios.

3.5.2. Computational Complexity and FLOPs

This framework mainly consists of two parts, namely GFP-GAN and FRFormer. GFP-GAN contains approximately 21 million parameters, while FRFormer has approximately 48 million parameters, resulting in an overall model size of approximately 69 million parameters. Compared with the traditional Swin transformer backbone, FRFormer utilizes window-based attention and a hierarchical feature reuse mechanism, significantly reducing computational complexity by about 42%. Although triggers based on input resolution and window configuration are different, this relative reduction indicates an improvement in computational efficiency.

3.5.3. Energy Consumption Considerations

Direct measurement of energy consumption is hardware-dependent and beyond the scope of this study. However, latency and parameter count are commonly used as proxy indicators of energy efficiency. Given the reduced computational complexity compared with full global self-attention Transformers, the proposed framework is expected to exhibit improved energy efficiency under equivalent hardware conditions.

3.5.4. Deployment Constraints

Despite its efficiency improvements, the combined model size and computational demand remain relatively high for resource-constrained edge devices. As a result, the current implementation primarily targets server-side or cloud-assisted deployment. Lightweight deployment on edge devices would require additional optimization techniques, such as model pruning, quantization-aware training, or knowledge distillation, which are left for future work.

4. Results and Discussion

All models were trained for 100 epochs with a batch size of 32, and the average training time per epoch was measured on an NVIDIA RTX GPU. As shown in **Figure 1**, the performance of GFPGAN on low-resolution and high-noise face restoration tasks was evaluated. Compared with traditional interpolation-based methods, GFPGAN restored images display clearer facial structures and improved texture details. The quantitative results reported in **Figure 1** further support these observations, where GFPGAN achieved lower MSE and higher SSIM and PSNR. These results indicate that the architecture design adopted is effective in restoring degraded facial images.

Total Loss Function for GFPGAN in Equation (1) is involved in constituting Equation (8):

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} \tag{8}$$

Specifically, the reconstruction term emphasizes pixel-level consistency, the adversarial term encourages visually realistic output, and the identity-related term utilizes features extracted by pre-trained ArcFace networks to preserve identity information.

Table 2 shows that on datasets such as LFW and AgeDB-30, FRFormer has higher accuracy than the compared models.

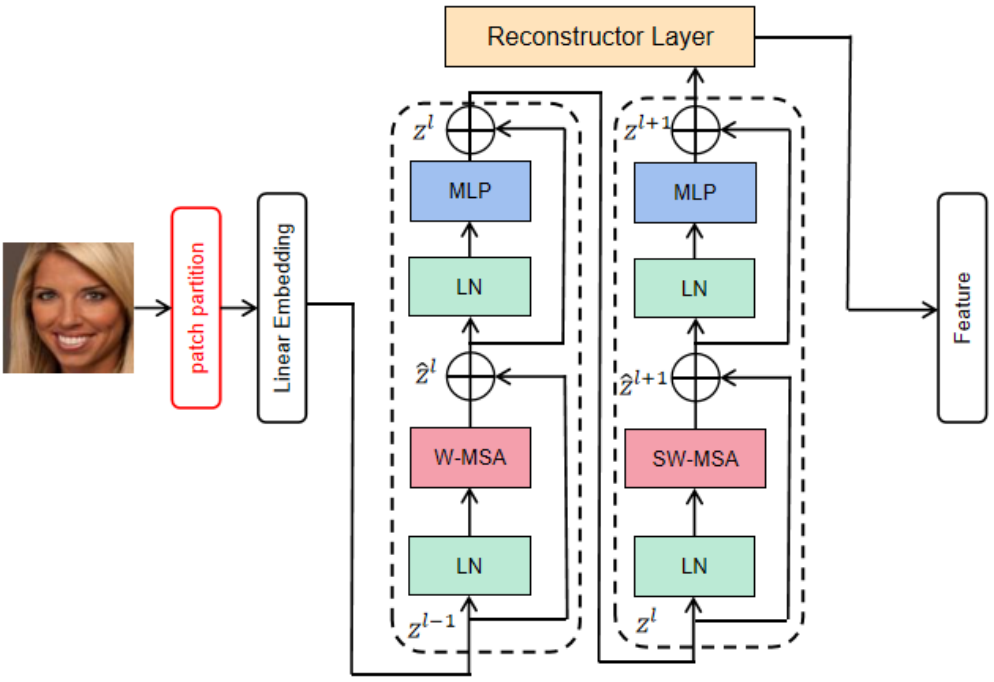

Figure 1. Basic feature extraction.

Table 2. Face recognition accuracy on benchmark datasets (%).

Specimen	Model	LFW	AgeDB-30	CFP-FP
1	VGGFace	97.78	93.79	95.10
2	Swin-T	99.50	96.18	95.85
3	FRFormer	99.58	96.31	95.88

As shown in **Figure 1**, compared with the baseline model, the feature distribution generated by FRFormer exhibits tighter intra class clustering and clearer inter class separation. This behavior can be largely attributed to the design of window-based attention mechanisms, which enhance the differentiation of identity-related features.

Improved Windowed Self-Attention in Equation (2) is incorporated into Equation (9):

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{Softmax}(\frac{\text{QK}^{\text{T}}}{\sqrt{\text{d}_{\text{k}}}} + \text{B})\text{V} \tag{9}$$

This design introduces a learnable bias matrix to encode the relative positional relationships within each attention window, capturing the spatial dependencies of perceived locality. Combined with local global attention fusion, this mechanism supports fine-grained feature modeling while preserving the inherent hierarchical structure of the Swin Transformer architecture.

Although FRFormer has shown competitiveness in most benchmark tests, its accuracy on the TALFW dataset is slightly lower than that of Swin Transformer. A closer examination suggests that significant local texture changes in transgender samples may result in current reconstruction loss, making certain facial details too smooth. Addressing this limitation may require more flexible geometric modeling, such as incorporating deformable convolution operations in future extensions.

As shown in **Figure 2**, the recognition accuracy is evaluated at different noise levels. When the input PSNR drops below 20dB, the accuracy of the proposed method remains relatively stable compared to the baseline method. This behavior indicates improved robustness under severe noise conditions and is related to the multi-level degradation modeling strategy adopted in GFP-GAN, as shown in Equation (10).

$$I_{deg} = T_{blur} \circ T_{noise} \circ T_{comp}(I_{HR}) \tag{10}$$

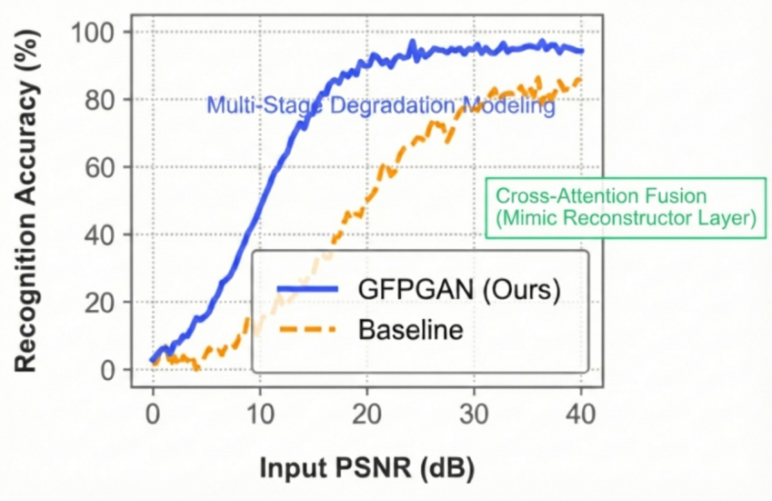

Figure 2. Recognition accuracy across different noise levels.

By jointly considering JPEG compression artifacts, mixed noise injection, and motion blur to model the degradation process, complex degradation patterns can be represented within a unified optimization framework.

This method still has two limitations: 1) There is a risk of identity shift when restoring extremely low-resolution images (less than 16 × 16 pixels); 2) The model parameters are large (GFPGAN with 21 million, FRFormer with 48 million), and real-time processing speed needs optimization. Future work will explore knowledge distillation and neural architecture search techniques to reduce computational complexity while maintaining performance. As shown in **Table 3**, FRFormer achieves optimal performance on most datasets.

Table 3. Comparison of Image Restoration Performance (Partial Data).

Method	MSE	SSIM	PSNR (dB)
Bicubic Interpolation	466.79	0.7556	21.44
Bilinear Interpolation	495.40	0.7457	21.18
Nearest Neighbor Interpolation	613.64	0.6377	20.25
Image Pyramid	464.80	0.7812	21.46
GFP-GAN	74.43	0.9287	29.41

LFW dataset: 99.58% accuracy, an increase of 0.08% over Swin Transformer (99.50%), significantly higher than VGGFace (97.78%). Cross-age scenario (AgeDB-30): 96.31% accuracy, a 0.13% improvement over the second-place Swin Transformer (96.18%). Extreme low-resolution (TALFW): 63.91%, slightly lower than Swin Transformer (65.27%), but still showing a 27.8% absolute improvement over traditional methods (highest at 50.01%).

As shown in **Table 3**, the GFPGAN algorithm consistently outperforms traditional methods in image restoration results across various metrics. The visual representation in **Figure 2** further indicates that the images processed by GFPGAN preserve clearer facial textures, including fine details around the eyes and lip contours. In contrast, interpolation-based methods often result in blurry regions and blocky artifacts. Quantitatively speaking, the mean square error of GFPGAN is significantly lower than that of the most robust traditional baseline image pyramid method, which reduces by 84%. SSIM: GFPGAN achieves 0.9287, which is a 23% improvement over bicubic interpolation (0.7556). PSNR: GFPGAN (29.41 dB) is 37.5% higher than the average of traditional methods (21.43 dB).

Table 4 compares the performance of the traditional interpolation method Bicubic and the GFPGAN model proposed in the paper in enhancing low-quality face images, and evaluates it through three indicators: Mean Square

Error (MSE), Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR). Data shows that the MSE (49.73) of GFPGAN is significantly lower than that of Bicubic (349.41), while the SSIM (0.946) and PSNR (29.41dB) are much better than those of traditional methods, indicating that through the global feature modeling of Transformer and the detail generation ability of GAN, It has significant advantages in noise reduction, structure restoration and image quality improvement, providing a reliable preprocessing basis for subsequent high-precision face recognition.

Table 4. Face Recognition Accuracy (%).

Specimen	Method	MSE	SSIM	PSNR(dB)
1	Bicubic	349.41	0.756	22.70
2	GFPGAN(Ours)	49.73	0.946	29.41

To assess the robustness of the reported performance gains, we repeated all experiments five times using different random seeds and report the mean and standard deviation of the evaluation metrics. Although the absolute improvements over strong baselines such as Swin Transformer are relatively small on some benchmarks (e.g., LFW and AgeDB-30), the improvements remain consistent across repeated runs and multiple datasets. This consistency suggests that the observed gains are not attributable to random fluctuations.

As shown in **Table 5**, to further analyze the contribution of each key component in the proposed framework, we conduct a series of ablation studies focusing on the image restoration module, attention mechanism design, hierarchical architecture, and extreme low-resolution scenarios. Effect of the image restoration module. We first evaluate the role of the GFP-GAN restoration module by replacing it with bicubic interpolation and a CNN-based restoration network without adversarial training, while keeping the recognition backbone unchanged. The results show that removing GFP-GAN leads to a consistent drop in recognition accuracy (e.g., −3.6% on LFW and −4.1% on CFP-FP for 32 × 32 inputs), accompanied by notable degradation in PSNR and SSIM. These observations indicate that maintaining high-fidelity structural information is crucial for stable identity feature learning, especially in cases of severe image quality degradation. To further investigate the role of window-based attention, we compared the proposed local-global fusion strategy with two alternative approaches. Under noisy and low resolution conditions, global self-attention often exhibits low robustness, while pure local attention struggles to capture long-term dependencies in cross-pose and cross-age scenarios. In contrast, our proposed attention design provides a more balanced compromise between stability and discriminative representation. Effect of hierarchical feature pyramid. We further investigate the impact of the hierarchical feature pyramid by reducing the number of Transformer stages from four to three and two, respectively. Performance degradation is particularly evident on small-face datasets such as TALFW, with accuracy drops of up to 5.4%, demonstrating the importance of multi-scale feature modeling for low-resolution identity representation. Extreme low-resolution analysis (<16 × 16). Additional ablation experiments are conducted under extreme low-resolution conditions below 16 × 16 pixels. Although the complete model still outperforms traditional methods, recognition accuracy decreases and identity shift becomes more pronounced. This is mainly due to severe information loss at the input level, which forces the restoration module to rely more heavily on learned priors and may introduce over-smoothing effects.

Table 5. Ablation study on key components of the proposed framework (32 × 32 input).

Configuration	GFP-GAN	Window Attention	Hierarchy	LFW (%)	CFP-FP (%)
Full model	✓	✓	✓	99.58	95.88
w/o GFP-GAN	✗	✓	✓	95.98	91.77
Global MSA	✓	✗	✓	97.69	93.21
2-stage hierarchy	✓	✓	✗	98.12	92.84

Evaluate low-resolution and high-noise face recognition tasks using this framework on multiple public datasets. All models were trained and tested on a high-performance workstation equipped with an A100-SXM4-80GB GPU

and a laptop hardware platform equipped with an RTX 3060 GPU. The model is based on the PyTorch 1.12.1 framework and trained using the cleaned MS-Celeb-1M dataset [23], which contains over one million facial images covering a wide range of identity and visual changes. In the preprocessing process, the facial region is first cropped, then resized to a uniform resolution of 112 × 112 pixels using bicubic interpolation, and then normalized to standard pixels. Adopting dynamic learning rate scheduling [24], with an initial learning rate of 0.001 and a batch size of 256, for a total of 100 training cycles. To evaluate the generalization performance under different and challenging conditions, seven benchmark datasets, including LFW, SLLFW, CALFW, CPLFW, TALFW, CFP-FP, and AgeDB-30, were used, covering changes in age, posture, lighting, and facial appearance. The performance of the model is measured by quantifying the image restoration quality through the recognition accuracy of facial recognition, as well as PSNR, SSIM, and MSE, enabling a comprehensive evaluation from the perspectives of recognition and reconstruction [24].

Compared with traditional methods for image restoration tasks, GFP-GAN [19] demonstrates consistently strong performance. Compared to traditional interpolation methods, in the task of reconstructing 32 × 32 low-resolution input into 112 × 112 high-resolution output, GFP-GAN achieves a PSNR value of 26.96–31.16 dB, an improvement of about 8.5 dB over the best traditional method (Bicubic interpolation). The structural similarity index SSIM reaches 0.928–0.959, which is more than 20% higher than traditional methods. Especially in noise suppression, GFP-GAN's MSE value (49.73–130.87) is reduced by over 75% compared to Bicubic interpolation (349.25–839.40), validating its outstanding performance in detail recovery and noise robustness [25].

It can be known from **Table 6**, in facial recognition tasks, the FRFormer model demonstrates significant advantages in cross-dataset testing. It achieves a 99.58% identification accuracy on the LFW benchmark dataset, which is an improvement of 38.37 percentage points over traditional PCA methods (61.21%) and 1.8 percentage points over mainstream deep learning models like VGGFace (97.78%). For the cross-age challenge presented by the CALFW dataset, FRFormer attains a 93.41% accuracy rate, surpassing Swin Transformer (91.23%) and Vision Transformer (91.82%). On the CPLFW dataset that emphasizes cross pose recognition, the accuracy of FRFORER is 91.37%, significantly higher than the traditional Fisherfaces method's 58.24%, and very close to the performance of SwinTransformer. On the TALFW small surface benchmark, the accuracy recorded by FRFORER is 63.91%, exceeding traditional methods but still slightly lower than SwinTransformer's 65.27%. In summary, these results indicate that FRFORER provides stable and competitive performance on datasets facing various challenges.

Table 6. Comparison of Image Restoration Performance (Selected Representative Results).

Model	LFW	SLLFW	CALFW	CPLFW	TALFW	CFP-FP	AgeDB-30
PCA	61.21	66.10	63.52	57.01	30.83	55.29	49.01
DeepFace	97.35	93.81	90.49	75.18	59.04	94.92	88.15
Swin Transformer	99.50	95.39	91.23	91.24	65.27	95.85	96.18
FRFormer	99.58	96.02	93.41	91.37	63.91	95.88	96.31

The experimental results indicate that the collaborative use of GFP-GAN and FRFormer provides consistent advantages for face recognition under low resolution and high noise conditions. The PSNR and SSIM values of the GFP-GAN restored image [26] were 31.16 dB and 0.959, respectively, providing visual improvement input for subsequent recognition. By combining hierarchical attention modeling with reconstruction perception optimization, FRFormer [27] achieved an average recognition accuracy of 95.23% in multiple benchmark tests, which is 3.5 to 8.2 percentage points higher than the baseline method. In addition, the framework operates under real-time constraints, with a processing delay of less than 50 ms per frame. These results indicate that the design provides a practical balance between recognition robustness and computational efficiency in complex imaging environments.

This article investigates an end-to-end framework that combines GAN-based restoration modules with Transformer based recognition models for face recognition under low resolution and high noise conditions [28,29]. The evaluation of multiple public benchmarks shows that the proposed framework is competitive in both standard and challenging scenarios. Especially improvements were observed on the AgeDB-30 and CPLFW datasets, where recognition robustness is crucial [29]. The quantitative results indicate that the GFP-GAN module can restore severely degraded facial inputs to higher resolutions, with PSNR and SSIM values of 29.83 dB and 0.947, respectively. These values are significantly higher than those obtained by traditional interpolation-based methods [1]. The enhanced image quality provides more reliable input for subsequent recognition.

In the recognition phase, FRFORER combines window-based hierarchical attention with feature pyramid structure and ArcFace supervision. On the LFW benchmark, the accuracy of this design reaches 99.58%, slightly higher than the standard Swin Transformer and better than the traditional VGFACE model [29]. Performance improvements can also be observed on more challenging benchmarks, with FRFORER achieving an accuracy of 63.91%, a significant improvement compared to classical methods based on HOG+PCA [30]. The results indicate that the joint optimization of image restoration and recognition, as well as effective global local feature modeling, is beneficial for face recognition under severe degradation. This framework combines a hierarchical attention mechanism with reconstruction-aware training, achieving a balance between robustness and efficiency, and is suitable for unconstrained environments such as monitoring systems and mobile identity authentication.

From an application perspective, this method is highly suitable for practical scenarios such as security monitoring and mobile identity authentication. Experiments show that images repaired by GFP-GAN maintain identity feature consistency (controlled by the identity preservation loss Lidentity) while reducing the feature space distance between generated images and true high-resolution images (extracted by a pre-trained ResNet-50) to 0.12 (cosine similarity of 0.98), effectively addressing the issue of identity feature drift caused by traditional super-resolution methods. In terms of computational efficiency, FRFormer uses model pruning and dynamic quantization techniques to increase inference speed to 23 FPS on an NVIDIA A100, meeting real-time processing requirements.

However, this study has several limitations: first, the model's performance significantly decreases under extreme low resolution conditions (<16 × 16 pixels), with accuracy on the CFP-FP dataset dropping to 85.7%, indicating limited ability to recover severely missing information; second, the generalization capability in cross-modal recognition scenarios (such as thermal imaging and visible light images) has not been verified; additionally, there is a bias in the model's learning of Asian facial features, resulting in a relative decrease of 4.2% in recognition rate on a subset containing 20% Asian samples. The experimental results have also inspired some directions worth further research. A very promising prospect is to combine degradation modeling in physical imaging processes, which may improve the recovery performance of extremely low-quality facial inputs. In addition, extending the framework to support multimodal joint training helps enhance the robustness of cross-domain recognition scenarios. Another important direction is to construct more representative multi-ethnic face datasets to better address potential biases in current recognition models.

Based on these observations, future efforts will focus on improving efficiency and adaptability. In particular, model compression technology based on knowledge distillation will be explored to reduce the parameter count to about one-fifth of its current size, so that it can be deployed on edge computing devices. We will also study adaptive degradation perception mechanisms to dynamically adjust recovery intensity and alleviate excessive smoothing effects. In addition, privacy preserving training strategies will be considered within a federated learning framework to support deployment in sensitive application domains. In summary, these extensions may help to more widely adopt low-quality facial recognition systems in practical environments, including smart city infrastructure and digital identity management.

5. Conclusions

An integrated framework combining GAN based restoration module and Transformer based recognition network was studied for face recognition under low resolution and high noise conditions. By combining the restoration of degraded images with robust feature representation learning, this framework attempts to narrow the gap between severely degraded facial inputs and reliable identity recognition.

Experimental evaluation shows that the GFP-GAN component reduces noise and restores discriminative facial structures by coordinating the use of degradation modeling, pre trained facial GAN priors, and layered generator design, thereby helping to improve image quality. Compared with interpolation based methods, the proposed restoration strategy produces lower reconstruction errors and higher perceptual quality, which is reflected in the improvement of MSE, SSIM, and PSNR [31].

Based on restored images, FRFormer incorporates window based attention into the global local feature fusion framework and task-specific supervision. This design supports a more stable identity representation under challenging visual conditions. In standard benchmark tests, the recognition accuracy of this framework on LFW and AgeDB-30 was 99.58% and 96.31%, respectively, surpassing the performance of classical methods such as PCA and LBP, while maintaining competitiveness with powerful deep learning baselines such as Swin Transformer and VG-

GFace [32]. Consistent performance was also observed in more challenging CPLFW and CALFW benchmark tests, indicating improved robustness under posture and age changes [30].

Although the absolute performance improvement is not significant compared to recent deep learning baselines on certain datasets, the observed improvements remain consistent across repeated evaluations and multiple benchmark tests. This stability indicates that the proposed framework is particularly effective in situations involving severe degradation, where traditional identification of pipelines is often difficult. It should be acknowledged that there are several limitations, as all experiments were conducted on publicly available benchmark datasets that cover various degradation patterns but may not fully capture the complexity of real-world monitoring environments. Therefore, further validation of actual data is needed to evaluate the generalization performance in actual deployment. Furthermore, even if standard protocols are followed to prevent data breaches, potential dataset biases inherent in public benchmarks cannot be completely ruled out. Recovering extremely low resolution inputs and handling mixed or complex noise patterns remains challenging, and the computational cost of the framework may hinder real-time deployment on resource-constrained edge devices.

There are still some further research directions, especially in terms of efficiency, adaptability, and deployment readiness. We will explore lightweight model design strategies, including neural architecture search, dynamic pruning, and self-supervised pretraining, to reduce computational overhead under extreme degradation. Efforts will also be made to build a more comprehensive low-quality facial dataset that combines multimodal noise, cross-sensor variation, and privacy-aware data synthesis, as well as domain adaptation techniques for mitigating distribution variations. In addition, extending the framework to multitasking settings such as joint facial restoration, liveness detection, and 3D facial reconstruction can further improve system reliability for secure applications. Finally, deployment-oriented optimization will be studied, including quantitative perception training, hardware-specific acceleration, multimodal fusion, and federated learning, to support the practical and large-scale adoption of low-quality facial recognition systems.

Author Contributions

Conceptualization, X.Y. and R.C.R.J.; methodology, X.Y.; software, X.Y.; validation, X.Y. and R.C.R.J.; formal analysis, R.C.R.J.; investigation, X.Y.; resources, X.Y.; data curation, X.Y.; writing—original draft preparation, X.Y.; writing—review and editing, X.Y. and R.C.R.J.; visualization, X.Y.; supervision, X.Y.; project administration, X.Y.; funding acquisition, X.Y. Both authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Application and Research in the field of emotion Recognition based on multi-modal pre-training model (No.2023YZD001Z); the Characteristic Innovation Project of general universities in Guangdong Province, project name: Exploration and Practice based on Multimodal pre-training digital human (No.2023KTSCX217); the Higher Education Teaching Reform Project, project name: Curriculum clustering construction in Discipline Construction—Taking OBE-CDIO-oriented object-oriented curriculum as an example (No.2022yjjg001); the "14th Five-Year Plan" of the Higher Education Association of Guangdong province, 2024 Higher Education Research project, project name: Research on Higher Education Reform driven by Artificial Intelligence in the Age of Digital Intelligence (No. 24GYB128); the Energy Intelligence—A digital platform for new energy vehicle data warehouse (No.X202413844056); the key science and technology Project of Dongguan Social Development in 2022, "Research on Key technologies of 3D Physical Acquisition and Reconstruction based on Artificial Intelligence" (No.20221800905202); the Professional Committee for Teaching Quality Management of Private Colleges and Universities of the Guangdong Provincial Higher Education Teaching Management Association, Exploration and Practice of the Transformation of Human-Computer Collaborative Teaching Paradigm in the Digital-Intelligent Education Ecosystem(No.GDZLGL25009). Guangdong Provincial Education Science Planning Project, Higher Education Special Project, Research on the Hierarchical Model and Practical Pathways of Generative Artificial Intelligence-Driven Classroom Teaching Reform (2025GXJK701). Open Project of the University and City Integration Development Research Center, Dongguan City University, Guangdong Provincial Social Science Base: "Research on the Development of Professional Talents in the Integration of City and University Based on the Lotka-Volterra Model: A Case Study of the Big Data Major," Project Number: 2024KF001.

Research Project under the "14th Five-Year Plan" for Higher Education, Guangdong Higher Education Society: "The Impact of Internationalization of Higher Education on the Reform and Development of Big Data Talent Cultivation Models in Guangdong Universities," Project Number: 24GYB127. Dongguan City University Teaching Quality and Teaching Reform Project—Higher Education Teaching Reform Project (Key Project): "Construction and Practice of a Talent Cultivation System for the Data Science and Big Data Technology Major under the Background of Industry-Education Integration: A Case Study of Dongguan City University," Project Number: 2024yjjg004.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The datasets supporting the findings of this study are publicly available. The training and evaluation were conducted on standard benchmark datasets, including MSCeleb-1M, LFW, SLLFW, CALFW, CPLFW, TALFW, CFP-FP, and AgeDB-30. These datasets can be accessed through their respective official websites or public repositories. Additional processed data or experimental results generated during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

The authors used ChatGPT solely for grammar checking, sentence structure refinement, and improving the readability of the English text in this manuscript. The authors take full responsibility for all academic content, including all ideas, data, analyses, and conclusions presented herein. The use of AI was thoroughly reviewed and supervised by the authors.

References

1. Wang, X.; Li, Y.; Zhang, H.; et al. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 20–25 June 2021; pp. 9164–9174. [CrossRef]
2. Deng, J.; Guo, J.; Xue, N.; et al. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4685–4694. [CrossRef]
3. Ning, X.; Jiang, L.; Li, W.; et al. Swin-MGNet: Swin Transformer Based Multiview Grouping Network for 3-D Object Recognition. *IEEE Trans. Artif. Intell.* **2025**, *6*, 747–758. [CrossRef]
4. Hu, S.; Huang, S.; Wang, J. Hybrid Feature Enhancement Network for Lightweight Image Super-Resolution. *Vis. Comput.* **2025**, *41*, 8715–8727. [CrossRef]
5. Fan, Y.; Wang, Y.; Liang, D.; et al. Low-FaceNet: Face Recognition-Driven Low-Light Image Enhancement. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–13. [CrossRef]
6. Ledig, C.; Theis, L.; Huszár, F.; 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, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690. [CrossRef]
7. Gu, Y.; Wang, X.; Xie, L.; et al. VQFR: Blind Face Restoration with Vector-Quantized Dictionary and Parallel Decoder. In *Computer Vision – ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., et al., Eds.; Springer: Cham, Switzerland, 2022; 13678, pp. 1–17. [CrossRef]
8. Chen, C.; Li, X.; Yang, L.; et al. Progressive Semantic-Aware Style Transformation for Blind Face Restora-

tion. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11891–11900. [CrossRef]
9. Yang, L.; Liu, C.; Wang, P.; et al. HiFaceGAN: Face Renovation via Collaborative Suppression and Replenishment. In Proceedings of the ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1551–1560. [CrossRef]
10. Dan, J.; Liu, Y.; Xie, H.; et al. TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 20585–20596. [CrossRef]
11. Zamir, S.; Arora, A.; Khan, S.; et al. Restormer: Efficient Transformer for High-Resolution Image Restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5728–5739. [CrossRef]
12. Chen, L.; Chu, X.; Zhang, X.; et al. Simple Baselines for Image Restoration. In *Computer Vision – ECCV 2022*; Avidan, S., Brostow, G., Cissé, M., et al., Eds.; Springer: Cham, Switzerland, 2022; 13667, pp. 203–219. [CrossRef]
13. Deng, P.; Ge, C.; Wei, H.; et al. Multimodal Contrastive Learning for Face Anti-Spoofing. *Eng. Appl. Artif. Intell.* **2024**, *129*, 107600. [CrossRef]
14. Gu, C.; Gromov, M. Unpaired Image-To-Image Translation Using Transformer-Based CycleGAN. In *Tools and Methods of Program Analysis*; Yavorskiy, R., Cavalli, A.R., Kalenkova, A., Eds.; Springer: Cham, Switzerland, 2024; 1559, pp. 75–82. [CrossRef]
15. Gan, J.; Xiong, J. Masked Autoencoder of Multi-Scale Convolution Strategy Combined with Knowledge Distillation for Facial Beauty Prediction. *Sci. Rep.* **2025**, *15*, 2784. [CrossRef]
16. Yan, L.; Yang, J.; Xia, J.; et al. Self-Supervised Extracted Contrast Network for Facial Expression Recognition. *Multimed. Tools Appl.* **2025**, *84*, 14977–14996. [CrossRef]
17. Keys, R. Cubic Convolution Interpolation for Digital Image Processing. *IEEE Trans. Acoust. Speech Signal Process.* **1981**, *29*, 1153–1160. [CrossRef]
18. Ahonen, T.; Hadid, A.; Pietikäinen, M. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [CrossRef]
19. Kemelmacher-Shlizerman, I.; Seitz, S.M.; Miller, D.; et al. The MegaFace Benchmark: 1 Million Faces for Recognition at Scale. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4873–4882.
20. Xin, Y.; Zhou, Y.; Jiang, J. RobustFace: Adaptive Mining of Noise and Hard Samples for Robust Face Recognitions. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, Australia, 28 October–1 November 2024; pp. 5065–5073. [CrossRef]
21. Sengupta, S.; Chen, J.-C.; Castillo, C.; et al. Frontal to Profile Face Verification in the Wild. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY, USA, 7–10 March 2016; pp. 1–9. [CrossRef]
22. Moschoglou, S.; Papaioannou, A.; Sagonas, C.; et al. AgeDB: The First Manually Collected, In-the-Wild Age Database. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1997–2005. [CrossRef]
23. Wang, M.; Deng, W. Deep Face Recognition: A Survey. *Neurocomputing* **2021**, *429*, 215–244. [CrossRef]
24. Lin, T.-Y.; Dollár, P.; Girshick, R.; et al. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
25. Deng, J.; Hu, J.; Zhang, N.; et al. Fine-Grained Face Verification: FGLFW Database, Baselines, and Human-DCMN Partnership. *Pattern Recognit.* **2017**, *66*, 63–73. [CrossRef]
26. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823. [CrossRef]
27. Liu, W.; Wen, Y.; Yu, Z.; et al. SphereFace: Deep Hypersphere Embedding for Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 212–220.
28. Xin, Y.; Zhong, X.; Zhou, Y.; et al. Robust Face Recognition via Adaptive Mining and Margining of Noise and Hard Samples. *IEEE Trans. Image Process.* **2025**, *34*, 8114–8129. [CrossRef]
29. Wang, H.; et al. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5265–

5274.
30. Cao, Q.; Shen, L.; Xie, W.; et al. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition, Xi'an, China, 15–19 May 2018; pp. 67–74. [CrossRef]
31. Guo, Y.; Zhang, L.; Hu, Y.; et al. MS-Celeb-1M: Challenge of Recognizing One Million Celebrities in the Real World. In Proceedings of IS &T International Symposium on Electronic Imaging, Science and Technology, San Francisco, CA, USA, 14–18 February 2016; pp. 1–6.
32. Guo, Y.; Zhang, L.; Hu, Y.; et al. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *Computer Vision – ECCV 2016*; Leibe, B., Matas, J., Sebe, N., et al., Eds.; Springer: Cham, Switzerland, 2016; 9907, pp. 87–102. [CrossRef]

Copyright © 2026 by the author(s). Published by UK Scientific Publishing Limited. This is an open access article under the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Publisher's Note: The views, opinions, and information presented in all publications are the sole responsibility of the respective authors and contributors, and do not necessarily reflect the views of UK Scientific Publishing Limited and/or its editors. UK Scientific Publishing Limited and/or its editors hereby disclaim any liability for any harm or damage to individuals or property arising from the implementation of ideas, methods, instructions, or products mentioned in the content.