Sensors and Materials, Vol. 38, No. 6 (2026) 3315–3336
MYU Tokyo

S & M 4505

A Robust Facial Image Restoration Support System for Human Environment Security Based on Generative Adversarial Networks with Hybrid and Multi-scale Spatial Attention Mechanisms

Chwei-Shyong Tsai,[1†] Hsien-Chu Wu,[2†] and Yen-Yu Chen[2*]

[1]Department of Information Management, National Chung Hsing University
145 Xingda Rd., South Dist., Taichung 40227, Taiwan
[2]Department of Artificial Intelligence and Computer Engineering, National Chin-Yi University of Technology

(Received January 20, 2026; accepted May 28, 2026)

Keywords: face restoration, image sensors, human environment support systems, generative adversarial networks, hybrid attention aggregation, multi-scale spatial attention

In intelligent human environments, image sensors are frequently hindered by occlusions, such as masks, which degrade the reliability of facial data for security and interactive support systems. In this paper, we propose a unified framework designed as a robust support system that integrates an occlusion segmentation network with a face image restoration network. To facilitate deployment in resource-constrained sensing nodes, the segmentation network employs depthwise separable convolutions to ensure computational efficiency while leveraging residual connections for multi-scale feature fusion. On the basis of precisely localized occluded areas, a generative adversarial network is introduced to reconstruct facial structures with high fidelity. The generator incorporates two novel feature enhancement components: a hybrid attention aggregation module, which strengthens global semantic consistency within skip connections, and a multi-scale spatial attention module, designed to capture fine-grained textures from sensor data across different spatial scales. Experimental results on the CelebFaces Attributes High-Quality (CelebA-HQ) dataset demonstrate that the proposed system effectively restores masked facial regions, achieving a *PSNR* of 35.01 dB and an *SSIM* of 0.931 under challenging 35–45% occlusion ratios. By significantly enhancing visual fidelity and recognition robustness, this framework provides a reliable solution for real-world vision-based support systems in human-centric environments.

1. Introduction

With the rapid advancement of smart city infrastructure and IoT, intelligent sensing systems have become indispensable in supporting human-centric environments. Among various sensing modalities, vision-based sensors play a critical role in applications such as public security

*Corresponding author: e-mail: vpcyy2233@gmail.com
†These authors contributed equally to this study.
https://doi.org/10.18494/SAM6179

ISSN 0914-4935 © MYU K.K.
https://myukk.org/

monitoring, healthcare assistance, and automated human–machine interaction. These sensors are designed to capture and analyze facial information to provide essential data for decision-making support systems. However, in real-world human environments, the data collected by vision sensors are frequently compromised by physical occlusions. Particularly during and after the global pandemic, the widespread use of face masks has introduced significant challenges to facial data integrity, leading to a substantial decline in the performance of subsequent recognition and authentication tasks within these support systems.

To maintain the reliability of support systems in such environments, it is imperative to develop robust image processing frameworks that can restore occluded facial information from corrupted sensor data. Existing restoration methods often struggle with a trade-off between visual fidelity and computational overhead. In the context of human environment support, these systems must not only enable the high-quality reconstruction of occluded regions but also ensure high efficiency for potential deployment on edge sensing devices. While generative adversarial networks (GANs) have shown promise in image inpainting, maintaining global semantic consistency and capturing fine-grained textures remain a bottleneck, particularly when dealing with large-scale mask occlusions. Therefore, a specialized vision-based support framework that balances precise segmentation, semantic restoration, and computational efficiency is urgently needed.

To address the limitations of conventional architectures, Liu *et al.* proposed partial convolution,[1] which is performed only on valid pixels while ignoring masked regions, resulting in more coherent and visually plausible restorations at the cost of increased computational complexity. Qin *et al.* introduced Multi-Scale Attention Network (MSA-Net),[2] an advanced encoder–decoder model that integrates multi-scale spatial and channel attention to better emphasize informative features and improve restoration quality. However, insufficient differentiation between spatial and channel attention during training may limit feature extraction effectiveness. In addition, Pathak *et al.* proposed the Context Encoder,[3] which combines an encoder–decoder structure with an adversarial discriminator to enhance realism, representing an early integration of GANs[4] and autoencoders[5] for image inpainting.

While traditional face recognition systems have matured, their reliability significantly degrades under large-scale occlusions. A recent study[6] has emphasized that restoring semantic integrity is a prerequisite for robust biometric authentication in secure environments. This necessitates a framework that can not only remove occlusions but also hallucinate identity-consistent features. Recently, generative models have seen a paradigm shift from GANs to diffusion models. For instance, Song *et al.*[7] utilized latent diffusion for facial de-occlusion, achieving impressive stochastic diversity. However, diffusion-based methods often require iterative sampling, which is computationally expensive for real-time applications. Concurrently, the LaMa architecture[8] demonstrated the effectiveness of fast Fourier convolutions for large mask inpainting. In contrast, our approach optimizes the balance between inference speed and structural accuracy by leveraging a lightweight GAN-based segmentation-restoration pipeline.

In this study, we aim to address face recognition challenges under occlusion by leveraging deep-learning-based face image restoration. By reconstructing occluded facial regions to closely resemble their original appearance, the proposed approach seeks to improve identity recognition

accuracy in partially occluded scenarios. The framework first processes masked face images and employs a segmentation network to localize regions requiring restoration. Paired occluded and unoccluded facial images are then used for training to generate natural and context-consistent restorations. This restoration process not only enables reliable recognition under occlusion but also mitigates security risks caused by identity concealment while improving system robustness and efficiency in real-world applications.

The main contributions of this work are summarized as follows:

(1) A mask-aware facial occlusion segmentation network based on depthwise separable convolution is proposed to reduce model parameters and computational cost while improving segmentation efficiency and accuracy.
(2) A GAN-based face restoration framework integrating two complementary attention mechanisms is developed. The multi-scale spatial attention module, deployed at the lower layers, expands the receptive field to capture features across multiple spatial scales, while the hybrid attention aggregation module combines channel attention and self-attention to enhance contextual feature fusion, leading to more realistic and high-quality facial reconstruction.

2. Methods

The proposed masked face restoration framework is designed as a modular support system tailored for intelligent vision-based sensing environments. As illustrated in Fig. 1, the system pipeline integrates an occlusion localization stage with a generative restoration stage to transform degraded sensor signals into high-fidelity facial images.

Specifically, the system first focuses on facial data acquisition within the human environment, where vision sensors frequently capture corrupted images due to mask occlusions. To automatically localize these regions and facilitate subsequent reconstruction, a dedicated occlusion segmentation network is employed. This network is enhanced with depthwise separable convolutions[9] to significantly reduce computational costs, ensuring the system's feasibility for deployment on resource-constrained sensing nodes. The resulting binary occlusion masks are then combined with the raw masked images to provide essential spatial guidance for the restoration engine.

On the basis of the localized occlusion data, a generative adversarial network serves as the backbone of the restoration stage. To address the challenge of jointly modeling complex spatial and channel information from sensor data, the framework introduces a feature extraction process based on split gated convolutions. The generator incorporates two key enhancement modules: the multi-scale spatial attention (MS2A) module, which aggregates features from receptive fields at different spatial scales to capture fine-grained textures, and the hybrid attention aggregation (H2A) module, which strengthens global semantic consistency by capturing inter-channel dependencies. Through this attention-guided fusion mechanism, the system effectively restores occluded facial structures, providing a robust solution for downstream support applications such as secure identity authentication.

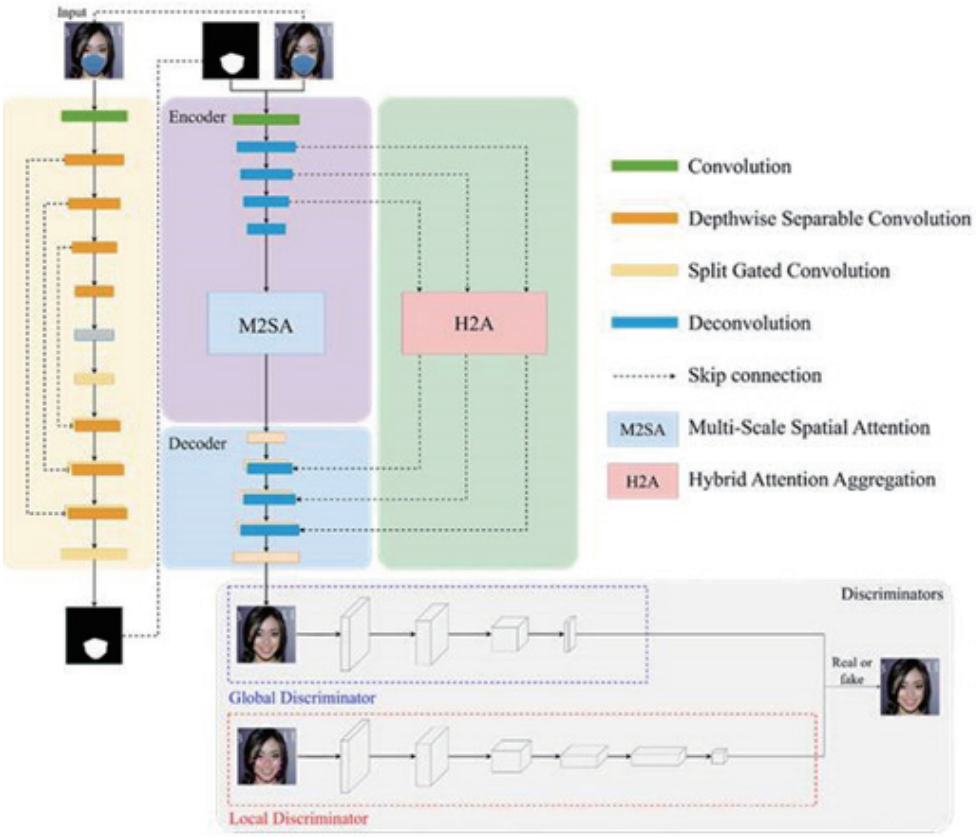

Fig. 1. (Color online) Proposed network flowchart.

2.1 Occlusion segmentation network

In the first stage, masked facial images are used as input to perform the binary segmentation of occluded regions. An autoencoder-based architecture is adopted for occlusion segmentation, where the encoder extracts representative features using depthwise separable convolutions and the decoder progressively restores spatial resolution. Skip connections are introduced between the corresponding encoder and decoder layers, allowing feature maps to be concatenated along the channel dimension. To ensure effective feature fusion, the concatenated feature maps share consistent spatial resolutions. Through this encoder–decoder structure with skip connections, the network outputs a predicted binary mask that accurately delineates the occluded facial regions.

2.2 Depthwise separable convolution

Depthwise separable convolution is an efficient convolutional operation that decomposes standard convolution into two successive stages. The first stage, referred to as depthwise

convolution, applies spatial convolution independently to each input channel using separate filters. This operation effectively preserves channel-specific spatial features but does not model inter-channel relationships, as no cross-channel aggregation is performed. To compensate for the lack of channel interaction, the second stage employs pointwise convolution, implemented as a 1×1 convolution, to linearly combine features across channels. This operation enables the network to learn channel-wise dependencies while maintaining a significantly reduced number of parameters. By combining depthwise and pointwise convolutions, the proposed segmentation network achieves substantial reductions in computational complexity and parameter count compared with standard convolution while preserving representational capability.

The overall architecture of the proposed occlusion segmentation network is illustrated in Fig. 2, and its processing pipeline consists of the following stages: First, the input image is fed into the encoder, where multiple convolutional layers progressively extract spatial and semantic features, capturing local patterns such as edges and textures.

During encoding, depthwise separable convolutions are employed to replace conventional convolution layers, effectively reducing model parameters and computational cost. By decoupling spatial convolution from channel aggregation, the encoder preserves spatial information while gradually embedding the input image into a semantically rich feature representation.

In the decoder, feature maps are gradually upsampled to restore the original spatial resolution. Upsampling is performed through transposed convolution or interpolation, followed by the application of convolutional layers to refine details and reconstruct spatial structures. To preserve high-resolution low-level features, skip connections are established between the corresponding encoder and decoder layers. Through feature concatenation or summation, multi-

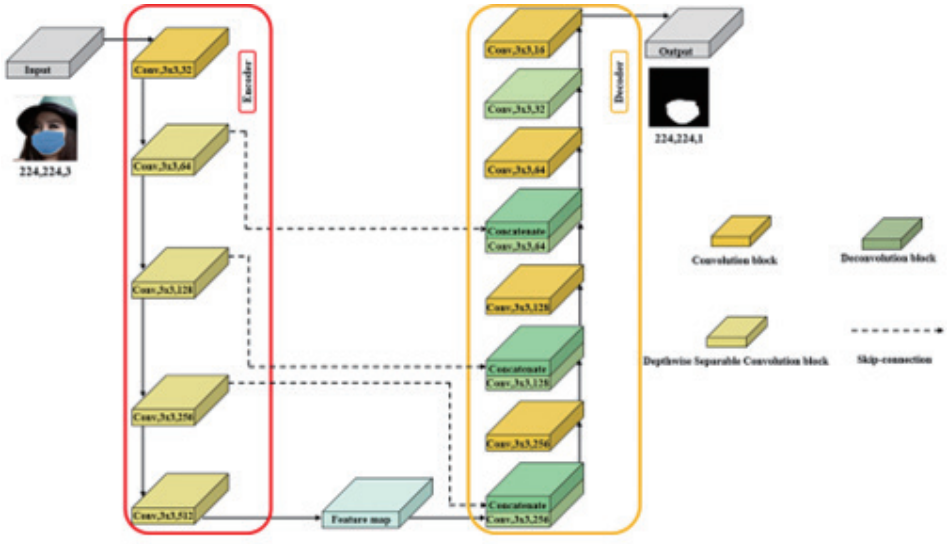

Fig. 2. (Color online) Proposed image segmentation network architecture diagram for occlusion detection.

level semantic and spatial information is fused, enhancing boundary localization and segmentation accuracy.

Finally, several convolutional layers are applied to integrate features and reduce channel dimensionality, producing a single-channel output map. Each pixel is classified in a binary manner, resulting in a predicted occlusion mask that completes the semantic segmentation task.

2.3 Face image restoration

In this stage, we focus on facial image restoration and adopt a GAN as the core framework, in which the generator is designed with an encoder–decoder architecture to reconstruct occluded facial regions. The encoder compresses the input image into a low-dimensional latent representation, enabling the extraction of semantically meaningful and highly discriminative features. To further enhance feature representation during encoding, a split gated convolution (SGC) mechanism is introduced, which dynamically filters and preserves task-relevant information through a gating strategy, thereby improving the expressive capability of the encoded features.

To strengthen feature learning for facial restoration, two feature enhancement modules, namely, the MS2A module and the H2A module, are incorporated. The MS2A module performs multi-scale feature aggregation using convolutional kernels of varying sizes, effectively enlarging the receptive field and improving the model's sensitivity to critical regions. In contrast, the H2A module emphasizes channel-wise attention modeling by learning inter-channel dependencies, allowing the network to highlight features that are particularly informative for restoration. The decoder subsequently upsamples the compressed features to reconstruct high-resolution feature maps and ultimately generates visually natural and structurally coherent restored facial images. Compared with existing attention mechanisms, the proposed MS2A and H2A modules are not designed as generic feature reweighting blocks, but as two task-specific and complementary modules for masked face restoration. Existing channel attention methods, such as SENet-style recalibration, mainly emphasize inter-channel importance but have limited ability to explicitly model long-range spatial dependencies. Conversely, conventional self-attention mechanisms are effective at capturing global correlations, yet they are often computationally expensive and do not explicitly address the multi-scale spatial variations caused by large contiguous facial occlusions. In addition, many existing attention-based inpainting methods apply a single attention form uniformly across the network, which may not be optimal for simultaneously recovering global facial semantics and fine-grained local textures.

To evaluate the realism of the generated results, a dual-discriminator architecture is optimized during training and a dual-discriminator strategy is employed. The global discriminator assesses the overall structural consistency of the restored face, ensuring global plausibility and coherence, whereas the local discriminator concentrates on fine-grained details, particularly facial components such as the eyes, nose, and mouth. The adversarial feedback from both discriminators is jointly propagated to the generator, guiding it to continuously refine restoration quality and structural integrity, thereby achieving high-fidelity facial image restoration.

2.3.1 SGC

The SGC module is designed to dynamically adjust the importance of feature responses during the encoding stage, enabling the network to capture more informative and semantically relevant representations. The overall architecture of the module is illustrated in Fig. 3.

During feature processing, a dual-branch structure is constructed, consisting of an input gate and a weight gate, to enhance feature selection and modulation. The input feature map is first divided into two parallel branches, each processed by independently parameterized convolutional filters followed by corresponding activation functions, allowing the extraction of diverse feature responses.

The input gate controls whether feature information is propagated to subsequent convolutional operations. By employing a ReLU activation function, negative responses are suppressed, ensuring that the network focuses on meaningful positive features and improves the efficiency of information flow. Conversely, the weight gate evaluates the relative importance of spatial locations within the feature map. A Sigmoid activation function is applied to constrain the output values to the range [0, 1], enabling the adaptive weighting of different regions and strengthening the representation of critical features.

The outputs of the two branches are then combined through element-wise multiplication, producing gated feature representations with enhanced discriminative capability. These gated features are subsequently integrated with the original feature map to achieve multi-level semantic fusion. Finally, to control computational complexity and channel dimensionality, a 3 × 3 convolution is applied at the end of the module to reduce the number of channels to half of the original size.

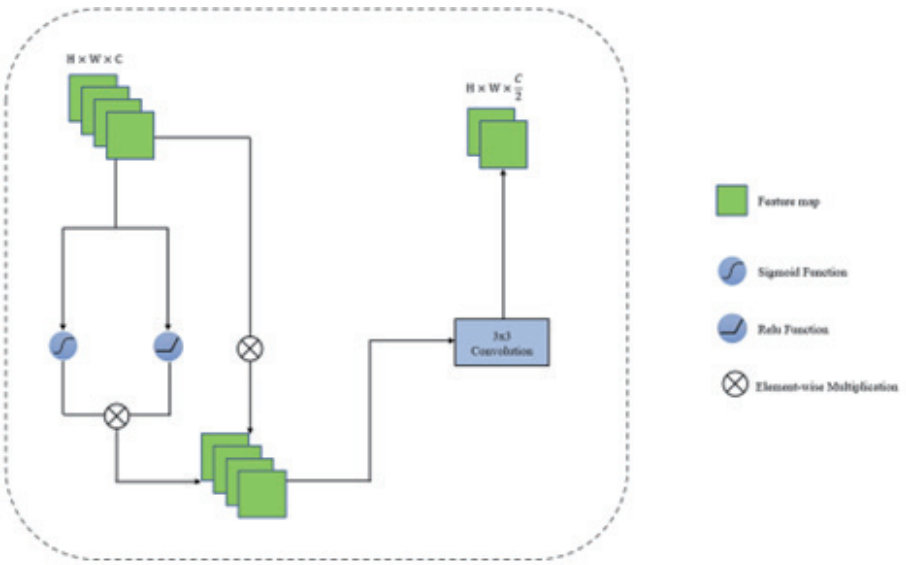

Fig. 3. (Color online) Split gate convolution architecture diagram.

2.3.2 MS2A

To enhance the network's ability to capture spatial features at multiple scales without increasing parameter count, we propose the MS2A module, which integrates multi-scale convolution with spatial attention. The module employs a multi-branch design to effectively expand the receptive field and strengthen feature representation as illustrated in Fig. 4.

The MS2A module consists of four branches. Three branches use convolution kernels of sizes 3 × 3, 5 × 5, and 7 × 7 to extract features at different spatial scales. These multi-scale feature maps are then element-wise summed to achieve feature fusion, enlarging the effective receptive field and improving the network's sensitivity to objects of varying scales.

The fourth branch computes spatial attention. The input feature map is first compressed via global max pooling to a 1 × 1 vector, retaining the maximum activation for each channel. This vector passes through a fully connected layer followed by a ReLU activation to learn inter-channel relationships. The resulting vector is then upsampled to the original feature map size and multiplied element-wise with the input feature map, providing preliminary spatial enhancement.

Finally, the fused multi-scale feature map is normalized using a Sigmoid function to generate a weight map. This weight map is applied element-wise to the spatially enhanced feature map

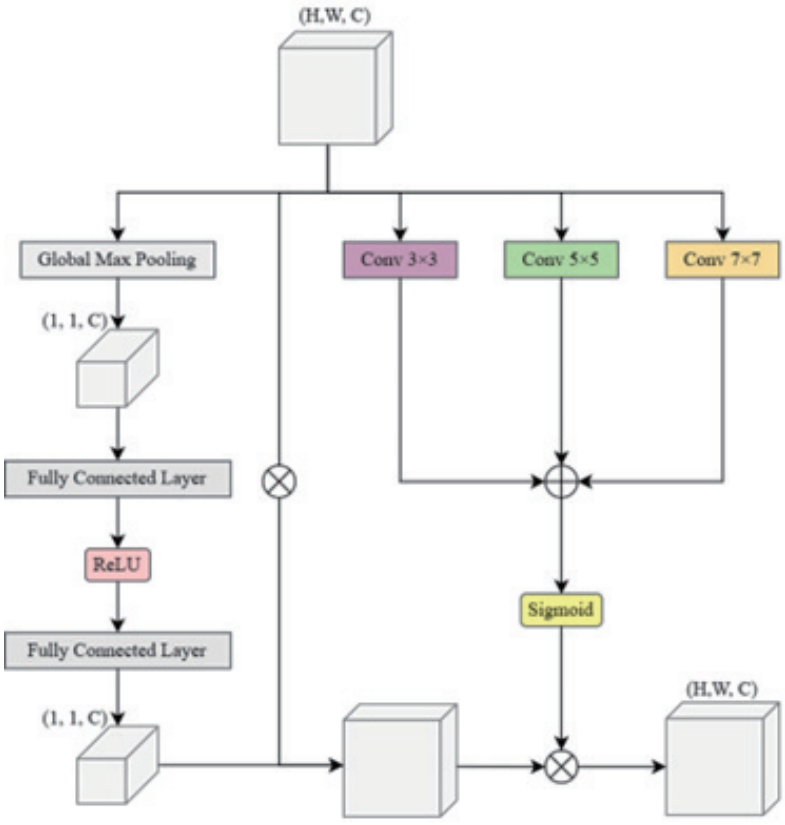

Fig. 4. (Color online) MS2A architecture diagram.

from the attention branch, adjusting the importance of each channel and spatial position, producing the final output feature map of the MS2A module.

Compared with conventional multi-scale convolution approaches, the MS2A module integrates spatial attention within multi-branch convolutional paths, enabling the adaptive weighting of features across different receptive fields without significantly increasing computational cost.

2.3.3 H2A module

The H2A module is designed to integrate contextual feature consistency by combining self-attention and channel attention mechanisms as illustrated in Fig. 5. It takes feature maps from the encoder path as input and, through attention-guided convolutions, captures critical features while effectively aggregating global contextual information.

The channel attention branch is inspired by SENet,[10] which models inter-channel dependencies through squeeze-and-excitation operations. Concurrently, the self-attention branch[11] calculates attention weights on the basis of feature similarities, allowing the network to selectively focus on important spatial regions and capture long-range dependencies.

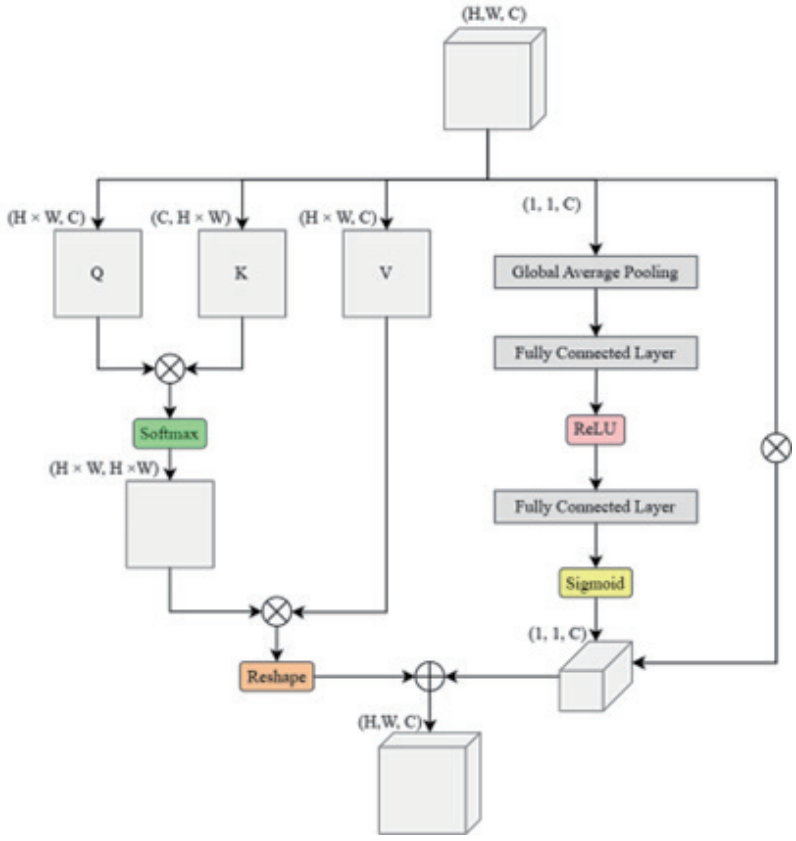

Fig. 5. (Color online) H2A architecture diagram.

Specifically, the self-attention branch transforms the input feature map into a Key matrix, which is then multiplied with the corresponding Query matrix derived from the original feature map to compute spatial correlations. The resulting similarity map is normalized via a Softmax function to generate spatial attention weights, which are applied element-wise to the input feature map, producing spatially enhanced features.

In parallel, the channel attention branch performs global average pooling to compress the input feature map to a 1 × 1 vector per channel. These channel descriptors are processed through fully connected layers with ReLU activations to learn nonlinear inter-channel relationships, followed by a Sigmoid activation to generate channel-wise weights. The element-wise multiplication of these weights with the input feature map produces features enhanced along the channel dimension.

Unlike the existing channel-only attention mechanisms such as SENet and spatial attention modules that focus on local or global regions independently, the proposed H2A module integrates both self-attention and channel attention to jointly model global dependencies and inter-channel relationships.

Finally, the outputs from the self-attention and channel attention branches are fused via element-wise addition, resulting in feature maps with both spatial and channel selectivities. This dual-attention mechanism enables the network to more precisely capture salient features while suppressing redundant information, thereby improving overall recognition performance.

The proposed face image restoration network is illustrated in Fig. 6. The training workflow is as follows:

The model takes as input a masked face image along with its corresponding binary mask. Features are first extracted and selectively filtered through the SGC, effectively preserving

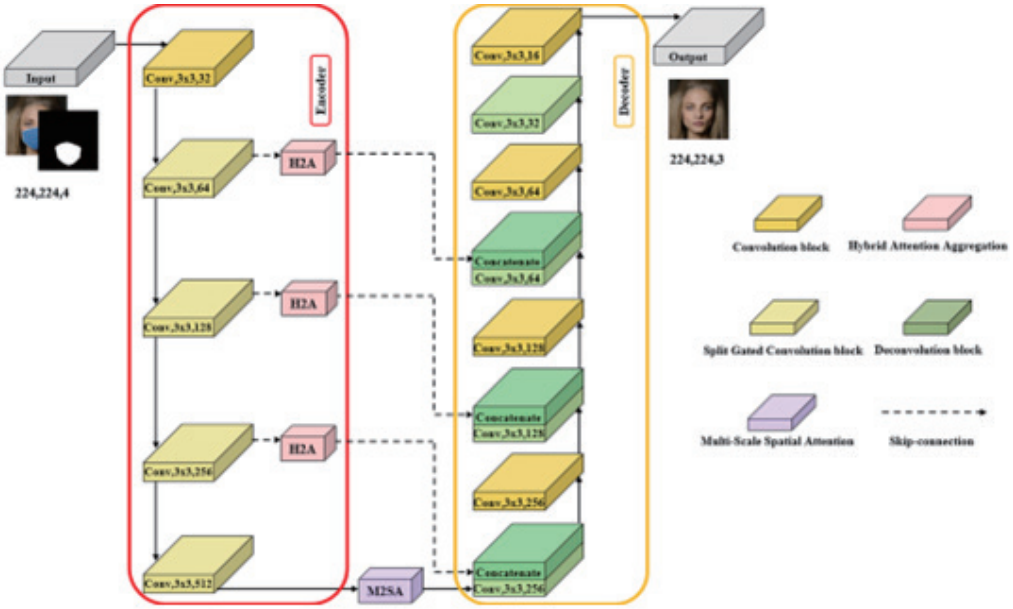

Fig. 6. (Color online) Proposed facial image restoration network architecture diagram.

semantically informative representations and enhancing overall feature expressiveness. At the network bottleneck, the MS2A module is embedded, which applies parallel convolutions with kernels of varying sizes to enlarge the receptive field and capture multi-scale semantic information. A spatial attention mechanism is incorporated within the MS2A module to further emphasize critical regions and strengthen feature representation.

The deep features generated by the MS2A module are passed to the decoder, where successive deconvolution operations progressively restore spatial resolution, reconstructing features to match the original input size. To improve restoration quality, skip connections link the decoder with the corresponding encoder layers. Prior to fusion, these skip-connected features are enhanced using the H2A module, which refines discriminative channel-wise representations within the masked regions. The fused features integrate semantic and contextual consistency, thereby enhancing the naturalness and realism of the generated content.

Finally, a 3 × 3 convolution is applied at the output layer to adjust the channel dimension to three, producing the final high-quality restored face image.

3. Discriminator

The architecture of the proposed discriminator network is illustrated in Fig. 7, and the training procedure is described as follows.

3.1 Global discriminator

The global discriminator receives a pair of input images, including the ground-truth image and the restored image produced by the generator. Feature extraction is performed using four consecutive 3 × 3 convolutional layers, where fixed-size kernels slide over the input images to capture local spatial features through convolution operations. Each convolutional block is followed by a 2 × 2 pooling layer to reduce the spatial resolution while retaining the most representative features.

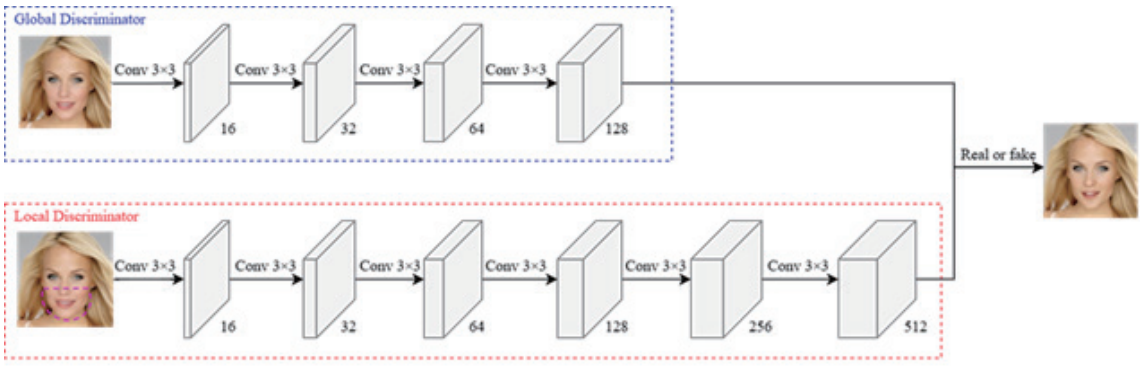

Fig. 7. (Color online) Discriminator network architecture diagram.

For nonlinear activation, LeakyReLU is adopted to alleviate the vanishing gradient problem and improve training stability. At the classification stage, a Sigmoid activation function is applied to output the final probability, indicating whether the input image is real or generated.

To enhance the realism of the generated results, the output of the global discriminator is incorporated into the adversarial loss and backpropagated to update the generator parameters. This adversarial feedback encourages the generator to produce restored images that increasingly resemble real facial images. The adversarial loss is defined as

$$L_{adv}^{global} = \underset{D}{\text{Max}} \underset{G}{\text{Min}} E\left[\log D(I_{inpaint}, I_{gt})\right] + E\left[\log\left(1 - D(G(I_{in}, I_{mask}))\right)\right], \tag{1}$$

where G denotes the generator network and D represents the discriminator network. The generator G takes the masked input image I_{in} and the corresponding binary mask I_{mask} as inputs, and produces the restored image $I_{inpaint}$. The final inpainted image is obtained by the following formulation:

$$I_{inpaint} = G(I_{in}, I_{mask}). \tag{2}$$

3.2 Local discriminator

The local discriminator is designed to focus on fine-grained details within the masked regions. During training, it receives a pair of input images consisting of the ground-truth image and the restored image produced by the generator. Feature extraction is performed using six stacked 3 × 3 convolutional layers to capture detailed local spatial patterns. Each convolutional block is followed by a 2 × 2 pooling layer to reduce the feature map resolution while preserving the most discriminative local information.

LeakyReLU is employed as the activation function to mitigate the vanishing gradient problem and ensure stable training. At the classification stage, a Sigmoid function outputs the probability indicating whether the input image corresponds to a real or generated sample.

To further enhance the realism of the restored results, the output of the local discriminator is incorporated into the adversarial loss and propagated back to the generator for parameter updates.

$$L_{adv}^{local} = \underset{D}{Max}\ \underset{G}{Min}\, E\left[\log D(I_{mask\_inpaint}, I_{gt})\right] + E\left[\log\left(1 - D(G(I_{in}, I_{mask}))\right)\right] \tag{3}$$

Unlike the global adversarial loss, the local discriminator operates exclusively on the masked regions of the inpainted image, denoted as $I_{mask\_inpaint}$, thereby enforcing realistic reconstruction within occluded facial areas. The corresponding local adversarial loss is formulated as

$$I_{mask\_inpaint} = I_{in} \otimes \left(1 - I_{mask}\right) + \left(I_{inpaint} \otimes I_{mask}\right). \tag{4}$$

4. Experimental Results

In this study, experiments were conducted on a Windows 10 Pro operating system using Python 3.8.13. The hardware environment (detailed in Table 1) includes an NVIDIA GeForce RTX 3070 8G graphics card. The proposed method was implemented using the Pytorch 1.9.0+cu111 framework.

Currently, no publicly available dataset provides paired images of the same subject with and without a mask, posing a significant challenge for supervised training. To address this, we used the CelebA-HQ dataset,[12] which contains diverse facial features and expressions.

In this study, masked facial images were synthesized using the MaskTheFace[13] tool, which accurately overlays various mask templates onto facial regions to generate realistic appearances. We selected a subset of the CelebA-HQ dataset for our experiments, consisting of 25000 images for training and 4565 images for testing. To maintain consistency across the network inputs, all images were resized to 224 × 224 pixels.

Under this configuration, the mask occlusions typically cover approximately 35 to 45% of the total facial area, primarily concentrated on critical semantic features such as the nose, mouth, and lower jaw. This high occlusion ratio poses a significant challenge for image restoration, requiring the model to demonstrate robust generative capabilities to maintain global coherence and local structural fidelity. Sample images of masked and unmasked faces are shown in Figs. 8 and 9, respectively.

For quantitative evaluation, two widely used image restoration metrics were employed: peak signal-to-noise ratio (*PSNR*) and structural similarity index (*SSIM*). *PSNR* measures signal fidelity between the reconstructed and original images, with higher values indicating lower distortion. *SSIM* assesses perceptual similarity by considering luminance, contrast, and structural information, simulating human visual perception. Together, these metrics provide a comprehensive assessment of restoration quality. *PSNR* is mathematically defined as

$$PSNR = 10 \times \log_{10} \frac{MAX^2}{MSE}. \tag{5}$$

PSNR is a widely used objective metric for assessing the quality of reconstructed images by quantifying the difference between the original and restored images. The core idea is to compare the maximum pixel value with the mean squared error (*MSE*) between the original image and the reconstructed image $I(i,j)$ at pixel location (i,j). *MSE* is calculated as

Table 1
Hardware specifications.

Operating system	Windows 10 Pro
CPU	Intel(R) Core(TM) i7-11700K CPU @ 3.60 GHz
GPU	NVIDIA GeForce RTX 3070 8G
System Memory	32 GB

Fig. 8. (Color online) Face images with masks.

Fig. 9. (Color online) Face images.

$$L_{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[O(i,j) - I(i,j)]^2, \tag{6}$$

where MAX_I is the maximum possible pixel value of the image. Higher *PSNR* values indicate smaller reconstruction errors, reflecting higher visual quality and similarity to the original image.

SSIM evaluates image similarity by simultaneously considering three components: luminance $l(i,j)$, contrast $c(i,j)$, and structural consistency $s(i,j)$. These components are mathematically defined as

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1},\ c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2},\ s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}, \tag{7}$$

where μ_x and μ_y are the mean intensities of the original and reconstructed images, respectively, σ_x and σ_y are their standard deviations, and σ_{xy} is the covariance. The constants c_1, c_2, and c_3 are included to stabilize the division and prevent numerical instability. *SSIM* values range from 0 to 1, where a value closer to 1 indicates higher perceptual similarity and structural preservation, while values closer to 0 indicate larger structural differences. *SSIM* is particularly effective for evaluating visual quality in applications that emphasize structural fidelity.

The results of the proposed method are shown in Fig. 10. By leveraging the mask as guiding information, the model can accurately identify regions requiring restoration and generate semantically consistent segmentation outputs. During training, the masked regions were

Fig. 10. (Color online) Obstruction image segmentation results.

weighted to emphasize semantically important areas, allowing the model to focus on meaningful regions and thus improving the reconstruction quality and overall visual fidelity.

Figure 11 illustrates the restoration results on the test dataset. These examples demonstrate that the reconstructed images closely resemble the original faces, with well-preserved details and no visible artifacts. The model exhibits stable restoration even under challenging conditions, such as oblique angles or severe occlusion, highlighting its robustness.

Quantitative evaluation further confirms the effectiveness of our approach. The proposed model achieved high scores in both *PSNR* and *SSIM*, indicating superior reconstruction quality and structural fidelity. As shown in Fig. 12, the model can reconstruct natural and visually

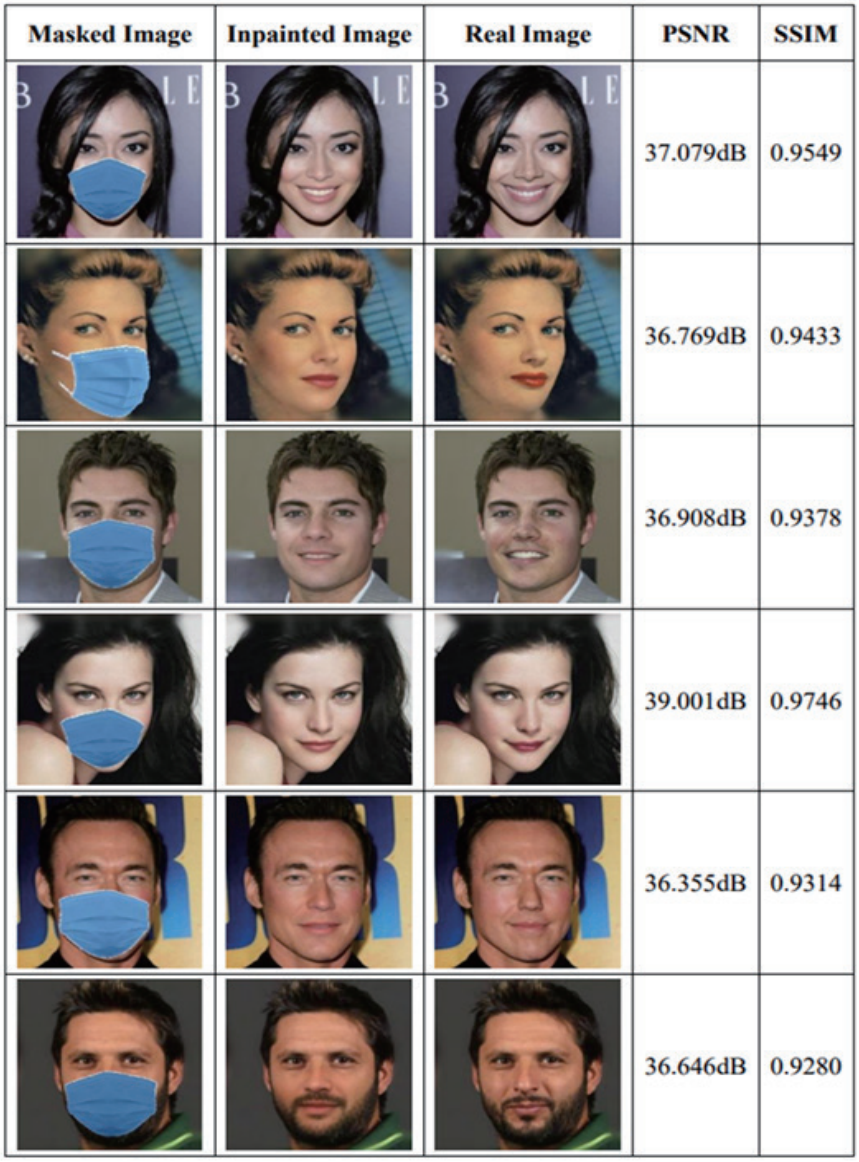

Fig. 11. (Color online) Repair results on the test set.

accurate facial images even under varying lighting conditions and camera angles, demonstrating strong generalization capabilities.

To verify the individual contributions of the key components within the proposed architecture, we conducted ablation studies focusing on the H2A and MS2A modules. As illustrated in Table 2 and Fig. 13, the integration of both modules yielded reconstructions that are significantly more consistent with the ground truth. This is particularly evident in regions requiring high semantic integrity, such as the mouth and jawline, where the combined model produced sharper and more natural textures than the degraded outputs of the subconfigurations.

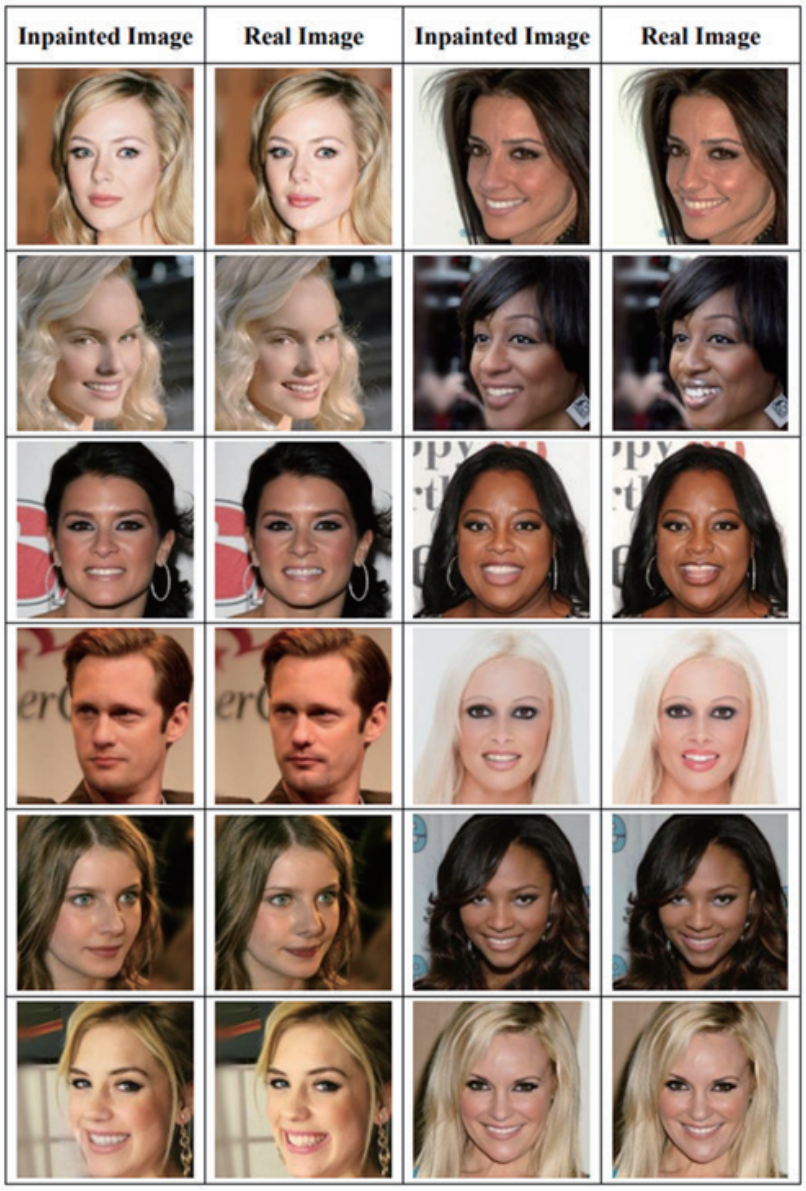

Fig. 12. (Color online) Restoration results across diverse skin tones and angles.

Table 2
Quantitative results of the ablation study on CelebA-HQ.

Module	w/o H2A	w/o MS2A	With all
PSNR	34.56	34.45	35.01
SSIM	0.924	0.923	0.931

Furthermore, we analyzed the effect of inserting the H2A and MS2A modules at different network stages. The results indicated that placing the MS2A module at the bottleneck and the H2A module within skip connections achieves the highest overall performance, demonstrating the importance of appropriate module positioning in the proposed architecture.

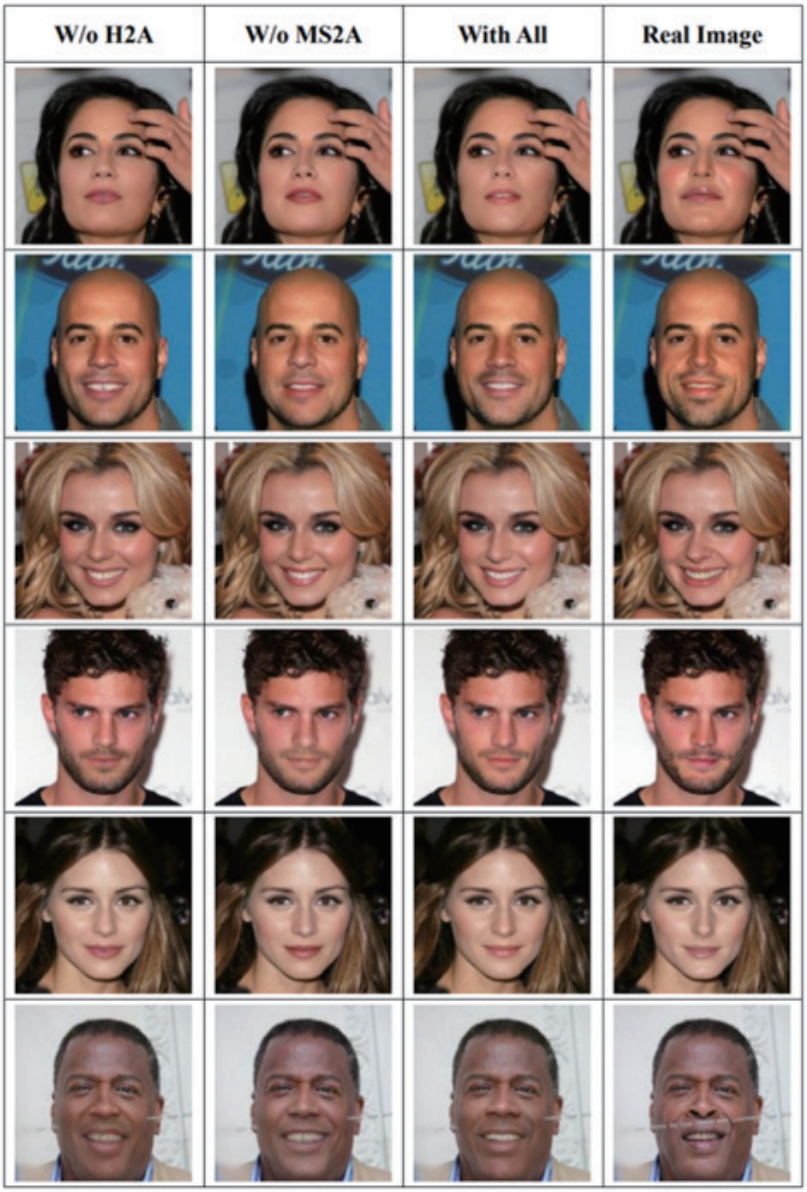

Fig. 13. (Color online) Ablation experiments using the H2A and M2SA modules.

Quantitatively, removing either the H2A or MS2A module led to a notable decline in performance. Specifically, the model utilizing only a single attention component achieved *PSNR* values of approximately 34.45–34.56 dB. In contrast, the full configuration (With All) achieved a *PSNR* of 35.01 dB, representing a performance gain of 0.45–0.56 dB. Similarly, *SSIM* improved from 0.923 to 0.931, indicating a refined structural restoration. These results demonstrate a synergistic effect between the two modules: while the H2A module optimizes discriminative channel features for global consistency, the MS2A module effectively captures fine-grained spatial details. Together, they enhance the model's perception of semantically critical regions, significantly elevating the overall fidelity of the facial restoration.

To evaluate the effectiveness of the proposed unified framework, we conducted a comprehensive quantitative comparison against four state-of-the-art methods: Ud Din *et al*.,[14] Wang *et al*.,[15] RFA-Net,[16] and Yingnan *et al*.[17] For a fair evaluation, all baseline models were tested under the same experimental conditions using the CelebA-HQ dataset with simulated face mask occlusions (occupying 35–45% of the facial area).

Table 3 shows the quantitative comparison in terms of *PSNR* and *SSIM*. The selected baseline methods are primarily GAN-based and attention-based approaches, which are closely related to the proposed framework and allow for fair comparison under similar experimental settings.

As shown in Table 3, the proposed method achieved a superior *PSNR* of 35.01 dB, outperforming the most recent RFA-Net[16] by 0.41 dB. While earlier GAN-based methods[14–17] struggle to maintain structural integrity under large-scale contiguous occlusions, the proposed framework leverages the H2A and MS2A modules to ensure both global semantic consistency and fine-grained texture restoration. Although the *SSIM* of the proposed method (0.931) is slightly lower than that of RFA-Net, note that RFA-Net is primarily designed to handle small and irregularly distributed missing pixels, whereas the proposed method is optimized for large and continuous masked areas. This difference in target scenarios explains the variation in structural similarity performance. Overall, the results demonstrate that the proposed approach achieves a higher balance between pixel-wise fidelity and structural realism in real-world masked face restoration scenarios.

In addition, the proposed modules introduce only marginal computational overhead while consistently improving *PSNR* and *SSIM*, demonstrating a favorable trade-off between performance and efficiency.

In addition to reconstruction quality, the computational efficiency of the proposed model is also an important consideration. All experiments were conducted on the hardware environment shown in Table 1.

The proposed framework incorporates depthwise separable convolutions to reduce computational complexity. Compared with standard convolution, which requires $K \times K \times M \times N$ parameters (where K is the kernel size and M and N denote the numbers of input and output channels, respectively), depthwise separable convolution decomposes the operation into a depthwise convolution and a pointwise convolution, requiring only $K \times K \times M + M \times N$ parameters. This significantly reduces both parameter count and computational cost.

As a result, the proposed model achieves a favorable trade-off between restoration performance and efficiency, making it suitable for practical and resource-constrained

Table 3
Quantitative comparison on the CelebA-HQ dataset under 35–45% mask occlusion.

Method	*PSNR* (dB)	*SSIM*
Ud Din *et al*.[14]	30.96	0.921
Wang *et al*.[15]	33.12	0.935
Chen *et al*.[16]	34.60	0.948
Yingnan *et al*.[17]	31.24	0.928
Ours (Proposed)	35.01	0.931

applications. Although detailed benchmarking against other methods is not provided owing to differences in implementation settings, the lightweight design highlights the efficiency advantage of the proposed approach.

5. Discussion

The proposed framework demonstrates strong and stable performance in restoring facial images under practical occlusion conditions, particularly when the occluded region accounts for approximately 35–45% of the facial area. This range aligns well with common real-world scenarios such as standard mask usage covering the lower half of the face. By focusing on this representative setting, the experimental design remains closely aligned with realistic application requirements. The results confirm that the model can reconstruct semantically consistent and visually plausible facial structures within this scope.

However, several limitations should be noted. The model may struggle under extreme occlusion scenarios where a large portion of facial information is missing, as well as under significant pose variations and challenging illumination conditions. These factors introduce additional complexity beyond the training distribution and may degrade restoration quality.

From the perspective of downstream applications, the proposed restoration framework has the potential to improve face recognition robustness under occlusion by recovering semantically meaningful facial features. Nevertheless, restoration does not guarantee perfect identity preservation, and reconstruction artifacts may still affect recognition performance in certain cases.

In terms of data formulation, the model is trained on synthetically generated masked faces to ensure a controlled supervision and stable learning of occlusion-to-complete face mappings. While this approach is widely adopted owing to the lack of large-scale paired real masked datasets, differences between synthetic and real-world conditions—such as mask texture, lighting interaction, shadow effects, and environmental noise—may affect generalization performance.

In recent years, diffusion-based models have demonstrated strong performance in image inpainting, particularly for large missing regions, by generating high-quality results through iterative denoising processes. However, their high computational cost and long inference time limit their suitability for real-time or resource-constrained applications such as face recognition systems. In contrast, the proposed framework achieves a balance between restoration quality and efficiency by leveraging attention mechanisms and lightweight convolutional design, making it suitable for practical deployment, including edge scenarios.

Future work will focus on addressing these limitations by exploring more challenging occlusion settings, incorporating real-world masked face datasets, and investigating domain adaptation strategies to further enhance robustness and generalization capability.

6. Conclusion

In this paper, we proposed a unified GAN-based framework for facial image restoration under mask occlusion in intelligent human-centered sensing environments. By integrating an occlusion segmentation network with a restoration network, the proposed system effectively processes facial data acquired from vision-based sensor systems, identifies occluded regions, and focuses reconstruction on semantically meaningful facial structures. To enhance both representation capability and computational efficiency, we introduce split gated and depthwise separable convolutions, enabling a favorable balance between restoration quality and model complexity for resource-constrained sensing devices and edge deployment scenarios.

Quantitative evaluations on the CelebA-HQ dataset demonstrate the effectiveness of the proposed method. Under mask occlusion ratios of approximately 35–45%, the model achieves an average *PSNR* of 35.01 dB and an *SSIM* of 0.931, outperforming several state-of-the-art approaches. Qualitative results further confirm that the reconstructed images preserve natural appearance and structural consistency, supporting reliable downstream face recognition tasks.

Ablation studies validate that the proposed H2A and MS2A modules play a critical role in capturing both global contextual information and fine-grained spatial details, leading to improved restoration quality. Note also that, while the proposed framework performs robustly on synthetic masked data, generalization to real-world scenarios remains an important direction for future research.

In conclusion, the proposed method provides an efficient and high-fidelity solution for masked face restoration, enhancing the robustness and reliability of vision-based intelligent sensing systems in practical applications.

References

1 G. Liu, F. A. Reda, K. J. Shih, T. C. Wang, A. Tao, and B. Catanzaro: Proc. European Conf. Computer Vision (Springer, 2018) 85–100. https://doi.org/10.1007/978-3-030-01252-6_6
2 J. Qin, H. Bai, and Y. Zhao: Comput. Vis. Image Underst. **204** (2021) 103155. https://doi.org/10.1016/j.cviu.2020.103155
3 D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros: Proc. IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2016) 2536–2544. https://doi.org/10.1109/CVPR.2016.278
4 I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio: Proc. Advances in Neural Information Processing Systems (NIPS, 2014) 2672–2680. https://papers.nips.cc/paper/5423-generative-adversarial-nets
5 A. Ng: Sparse Autoencoder, CS294A Lecture Notes (Stanford University, 2011). http://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf
6 L. Zhu, X. Wang, J. Chen, Y. Li, and H. Zhang: IEEE Trans. Inf. Forensics Secur. **20** (2025) 1125.
7 Y. Song, L. Shen, Z. Yang, H. Zhang, and D. Zhou: Proc. Advances in Neural Information Processing Systems (NeurIPS, 2024) 2845–2856.
8 R. Suvorov, E. Logacheva, and A. Mashikhin et al: Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (IEEE, 2022) 2102–2112. https://doi.org/10.1109/WACV51458.2022.00323
9 F. Chollet: Proc. IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2017) 1251–1258. https://doi.org/10.1109/CVPR.2017.195
10 J. Hu, L. Shen, and G. Sun: Proc. IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2018) 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

11 H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena: Proc. Int. Conf. Machine Learning (PMLR, 2019) 7354–7363. https://proceedings.mlr.press/v97/zhang19d.html
12 T. Karras, T. Aila, S. Laine, and J. Lehtinen: Proc. Int. Conf. Learning Representations (ICLR, 2018). https://openreview.net/forum?id=Hk99zCeAb
13 A. Anwar and A. Raychowdhury: arXiv (2020). https://arxiv.org/abs/2008.11104
14 N. Ud Din, K. Javed, S. Bae, and J. Yi: IEEE Access **8** (2020) 44276. https://doi.org/10.1109/ACCESS.2020.2977386
15 M. Wang, W. Lu, J. Lyu, and X. Zhang: Displays **75** (2022) 102321. https://doi.org/10.1016/j.displa.2022.102321
16 M. Chen, S. Zang, Z. Ai, and Y. Li: Eng. Appl. Artif. Intell. **119** (2023) 105814. https://doi.org/10.1016/j.engappai.2022.105814
17 S. Yingnan, F. Yao, and Z. Ningjun: Optik **242** (2021) 167101. https://doi.org/10.1016/j.ijleo.2021.167101