Lightweight Multi-Feature Fusion GAN with Deformable Attention for HMD-Occluded Face Reconstruction

Yingying Li[1], Ajune Wanis Ismail[2], Muhammad Anwar Ahmad[3], Norhaida Mohd Suaib[4], Fazliaty Edora Fadzli[5]
Mixed and Virtual Reality Research Lab-ViCubeLab-Faculty of computing, Universiti Teknologi Malaysia (UTM), Skudai, Johor 81310, Malaysia[1, 2, 3, 4, 5]
Shanxi Agricultural University, Jinzhong 030801, China[1]

***Abstract*—Head-mounted displays (HMDs) enhance virtual reality (VR) experiences, but occlude the upper face, hindering realistic user representation. To address this, some studies employ sensors to capture facial expressions under occlusion, while deep learning methods typically rely on image inpainting to restore missing regions. However, these approaches often suffer from limitations such as insufficient shallow feature representation, high computational complexity, and redundant model structures. This study proposes a lightweight generative adversarial network (GAN) that utilizes multi-feature fusion and deformable attention for face reconstruction under HMD occlusion. Specifically, a Lie group feature learning module is used to enhance shallow geometric representations, while reference-guided deformable attention dynamically focuses on occluded regions, improving both structural fidelity and efficiency. Experiments across multiple face datasets show that the proposed method outperforms existing mainstream approaches regarding structural fidelity, detail restoration capability, and model efficiency. The proposed framework offers a promising solution for integration with HMDs equipped with facial tracking, enabling more realistic and expressive avatars in VR applications.**

Keywords*—Generative adversarial network; Lie group feature learning; deformable attention; face reconstruction; virtual reality; head-mounted displays

I. Introduction

Virtual reality (VR) is a technology that creates virtual environments by simulating real-world sensory experiences (such as vision, hearing, and touch) using computers [1]. The rapid growth of VR has enabled its widespread use in domains such as education [2],[3],[4], healthcare [5],[6],[7], entertainment [8],[9],[10], remote collaboration [11],[12],[13], and more. As a VR interaction device, the head-mounted display (HMD) provides users with a highly immersive communication experience and enables more natural and vivid forms of interaction [14]. However, since the HMD covers the upper face, it obstructs the effective extraction of facial features. It makes it challenging to integrate facial information fully into the virtual scene [15]. The occlusion reduces the realism of virtual communication and affects users' social experiences [16]. Therefore, removing HMD occlusion from facial images and restoring missing facial information has become a key challenge in enhancing the realism of VR interaction.

There are two main approaches for reconstructing the face under HMD occlusion [17]: model-based methods and image-based methods. Model-based methods use sensors to capture facial information in the occluded areas and achieve full face reconstruction by fitting a 3D face model [18],[19],[20],[21]. For example, Chen et al. [18] reconstructed the users' facial expression and restored eye gaze direction by using three infrared cameras to directly capture the occluded face beneath the HMD. Kin et al. [19] proposed a facial expression recognition system that combines facial electromyography (fEMG) and electrooculography (EOG). The system collects signals via tiny electrodes placed around the eyes and uses machine learning to estimate virtual blendshape weights, enabling natural mapping of expressions to 3D avatars. However, such methods require the integration of additional sensors into the HMD, leading to problems such as reduced wearing comfort and increased cost [22]. In recent years, advanced HMDs such as the Meta Quest Pro have been gradually adopted in industrial applications. By integrating similar sensors into the HMDs, they have improved wearing comfort and enhanced facial expression tracking [23]. Nevertheless, in practical applications, these devices still cannot capture complete facial texture information, making it challenging to render high-fidelity facial appearances.

Image-based methods primarily use deep learning models to generate the occluded facial regions, aiming to restore the missing information in facial images realistically [17], [24],[25],[26]. Numan et al. [17] introduced a GAN-based HMD removal framework that simultaneously reconstructs the incomplete color and depth information in RGB-D facial images. Gupta et al. [24] introduced spatial supervision and landmark prediction modules to improve facial image quality in the de-occlusion task. They optimized the reconstruction of the peri-ocular region by leveraging the inherent structure of the eye and enhanced feature extraction by incorporating an attention module. Ghorbani Lohesara et al. [25] incorporated a self-attention into a GAN, enabling HMD occlusion removal by exploiting multiple reference video frames. Bai et al. [26] proposed a universal facial encoding system for consumer-grade HMDs. By employing self-supervised learning and training on large-scale unlabeled HMD camera data, their method achieved cross-view face reconstruction without relying on 3D models or high-quality labels, significantly improving overall model performance.

In recent years, diffusion models [27] have made significant breakthroughs in image generation and have demonstrated superior performance to traditional GANs in tasks such as image restoration and image synthesis [28]. The central idea is to produce high-quality images via a progressive denoising process, which leads to more stable training and greater fidelity in fine image details. However, such methods typically require multi-step sampling, which leads to long generation times, low inference efficiency, and high demands on computational resources [29], [30]. Therefore, given the resource constraints in virtual reality applications, this study adopts a GAN-based generation approach.

Although GAN-based methods have shown promising progress in HMD-occluded face reconstruction, the current literature still presents several notable research gaps. First, most existing studies rely heavily on deep convolutional semantic features and overlook shallow geometric cues, which are crucial for maintaining structural consistency in reconstructed facial regions. Second, the attention mechanisms commonly adopted in prior work are dominated by global self-attention, whose quadratic computational complexity results in high memory consumption and low efficiency—making these models unsuitable for real-time or resource-constrained VR applications. Third, many existing approaches employ complex and redundant GAN architectures with large parameter counts, limiting the deployability and hindering practical integration into lightweight VR systems. These gaps highlight the need for a lightweight, geometry-aware, and computationally efficient framework for high-fidelity HMD-occluded face reconstruction.

In response to the identified research gaps, this study presents a lightweight, geometry-aware GAN architecture that incorporates multi-feature fusion and an efficient deformable attention module. The core design of the approach is described as follows:

1) To introduce the Lie group feature learning method to perform a structured representation of shallow features. By combining deep semantic features with shallow visual information, it effectively improved the integrity and geometric consistency of feature expression.

2) To propose a deformable attention mechanism guided by facial reference images, effectively reducing memory usage and computational complexity.

3) To propose a lightweight GAN model that combines parallel dilated convolution, Lie group feature modeling, and an attention module. This significantly reduces the model complexity and inference time while maintaining good reconstruction quality.

The primary contributions of this research can be outlined as follows:

1) Proposed a multi-feature fusion strategy combining shallow and deep features, and the Lie group feature learning method is introduced to significantly enhance the model's ability to model facial structure.

2) Designed a deformable attention mechanism guided by facial reference images (face deformable attention), which effectively reduces the computational complexity, improves the focus on key areas, and reduces the impact of irrelevant regions.

3) Constructed a lightweight GAN architecture that ensures high-fidelity reconstruction while having good adaptability.

The remainder of this study is organized as follows: Section II reviews the related work on HMD-occluded face reconstruction. Section III presents the proposed lightweight multi-feature fusion GAN with deformable attention. Section IV describes the experimental setup, implementation details, and performance evaluation. Section V concludes the study and outlines potential directions for future research.

II. Related Work

A. Occluded Face Reconstruction

Occluded face reconstruction is a challenging problem that has recently garnered significant attention. Current approaches can be categorized into two groups broadly [25]: model-based methods and image-based methods. Model-based methods rely on statistical models, such as 3D morphable models (3DMMs) [31], to estimate facial geometry and texture. Purps et al. [32] identified unoccluded facial regions, extracted facial landmarks, followed the Facial Action Coding System (FACS) to achieve facial muscle activation, and created realistic virtual images through 3D modeling software to achieve facial expression presentation. However, the above methods are mainly aimed at small-scale facial occlusions and have particular difficulties for large-scale occlusions, such as the difficulty in capturing detailed facial features. To overcome the limitations of small-scale facial occlusions, some studies have adopted more complex techniques to deal with large-scale facial occlusions. For example, He et al. [33] fitted a 3D facial model, used 3D facial information to reconstruct the occluded parts, and combined Gabor-based occlusion dictionary learning to increase feature diversity and better represent occluded faces. Li et al. [34] separated the occluded areas through an abnormal region segmentation network, avoiding the model fitting error caused by occlusion, achieving more accurate model fitting positioning, and improving the quality of occluded facial reconstruction.

Image-based methods usually use deep learning models to complete face images. Ju et al. [35] randomly added masks to the face dataset to synthesize an occluded face dataset, achieved self-supervised training by generating damaged images with simulated occlusion and rotation, and combined CFR-GAN to repair texture in occluded areas. Yu et al. [36] introduced a reference-guided face inpainting approach that restores missing pixels using a reference image of the same identity as the occluded face. Similarity constraints were employed to synthesize finer detailed texture information. Lu et al. [37] proposed a face inpainting method that combines a multi-stage generative adversarial network with a global attention mechanism. Their framework leverages a generator with skip connections, an encoder–decoder structure, and a local refinement network to enhance inpainting quality. Luo et al. [38] proposed a two-stage control framework that disentangles the reference image into identity features and texture details. They achieved identity-preserving face image

completion in the case of large-area missing photos. Hassanpour et al. [39] developed a GAN-based 'eye-to-face network' (E2F-Net) that targets the restoration of the periocular region to improve the realism and completeness of the reconstructed face images. While image-based approaches perform better on large occlusions, they primarily rely on deep convolutional semantic features and often lack explicit modeling of shallow geometric structures, which are essential for preserving facial consistency under severe occlusion. These limitations suggest that a more geometry-aware and lightweight GAN architecture is necessary—motivating the method proposed in this work.

B. Lie Group Feature Learning

Lie group machine learning [40] is a paradigm that integrates Lie group theory with machine learning algorithms, leveraging the mathematical structure of Lie groups to improve feature representation, generalization, and robustness, particularly for data exhibiting symmetry and geometric constraints [41]. Lie group feature learning is widely used to extract shallow features of target objects. Its theoretical basis stems from the advantages of Lie groups in describing geometric transformations and symmetries [42],[43],[44],[45]. Shallow features typically encompass the geometric structure of the target object, local texture details, and basic statistical properties, capturing essential patterns in the data's primary representation. These features are vital in deep learning systems because they provide a reliable basic representation for subsequent high-level feature learning. Xu et al. [42] first applied Lie group feature learning to shallow feature extraction of remote sensing images, effectively learning and representing shallow features such as target size and shape in remote sensing images, and combined with CNN models to achieve more discriminative feature extraction. The results show that Lie group feature learning can effectively enhance the feature representation ability of deep learning models [43]. Xu et al. [44],[45] further applied Lie group feature learning to the deep learning model for scene classification, allowing the model to capture and represent a more diverse set of features, enhancing the interpretability of the model, and thus effectively improving the model's scene classification performance. Cai et al. [46] used Lie groups to learn and implement natural motion data representation. Then they used CNN to discover and classify Lie group features, improving the accuracy of human motion recognition and saving computing time. Yang et al. [47] designed a Lie algebraic residual network (LARNet), which effectively improved face recognition accuracy by combining Lie groups and residual networks, and was robust to face posture. Although Lie group feature learning has shown promise in capturing geometric structures across different fields, it remains largely unexplored in occluded face reconstruction.

C. Attention Mechanism

The attention mechanism, inspired by human visual perception, allows the model to focus on the most relevant regions of the input data [48]. Initially, this mechanism was mainly used in natural language processing and significantly improved model performance. With the development of technology, researchers have gradually extended the attention mechanism to the field of computer vision and achieved many breakthrough results [49],[50],[51]. However, while improving the accuracy of the model, this mechanism also introduces new problems [52], such as high computational complexity and extensive memory usage. To solve these problems, scholars have proposed some solutions, such as Deformable Attention (DAT) [53], Slide-Transformer [54], Single-Head Vision Transformer (SHViT) [55], etc., which reduce the redundancy in attention calculation through different strategies and significantly improve its computational efficiency. Lu et al. [37] used the global attention mechanism to enhance global feature interaction and reduce information dispersion, thereby better realizing face information restoration. Wan et al. [56] introduced an unsupervised face restoration approach that integrates contrastive learning with an attention mechanism. The feature attention module focuses on key feature information and establishes long-range dependencies to improve the face restoration effect. Xu et al. [27] proposed a face inpainting approach that combines parallel visual attention (PVA) with a diffusion model. By inserting a parallel attention matrix into the denoising network and focusing on the reference image features extracted by the identity encoder, the identity preservation ability of the restored face is effectively improved. Chen et al. [57] proposed a channel attention layer with spatial activation, combined with a sandwich-style feed-forward network structure, to achieve efficient spatial modeling and context understanding of damaged images at multiple scales. While these methods significantly enhance restoration fidelity, most existing works either assume global attention, which is computationally heavy, or do not fully leverage facial cues that could better guide reconstruction under occlusion, such as HMDs.

III. Proposed Method

A. Overview

0 illustrates the complete framework of the model. Initially, the Lie group feature learning is adopted to extract shallow features from the reference face image, while a convolutional neural network captures high-level features. These features are then fused and used as input for the subsequent stage; then, a generator is constructed, in which we propose a deformable attention mechanism guided by facial reference images to reconstruct the face under HMD occlusion; finally, through adversarial training between the generator and discriminator, images containing various facial detail information are gradually generated to achieve high-fidelity face reconstruction.

B. Feature Learning

1) Shallow feature learning: The image samples are first projected onto the Lie group manifold space to derive their corresponding Lie group representations. Then, feature extraction is performed within the manifold space. Inspired by the method proposed in [42], key features such as color and gradient information are extracted at each pixel, and a Lie group-based regional covariance matrix is constructed based on these features. This matrix characterizes the shallow structural properties of the sample and provides a richer feature representation for subsequent analysis and processing. The formulation is as follows:

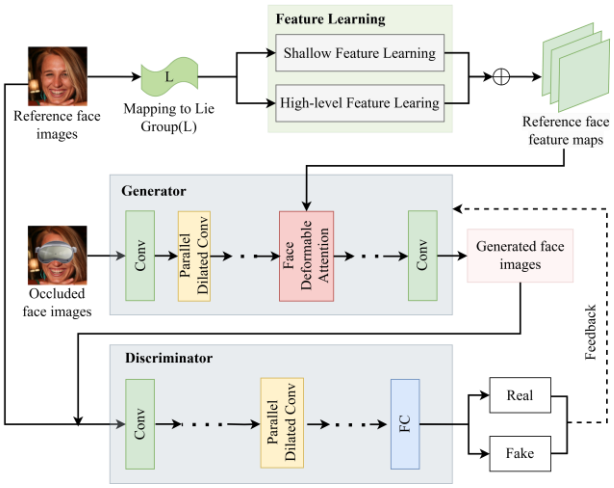

Fig. 1. Framework overview.

$$F(x,y)=\begin{bmatrix} x, y, HSV, \left|\frac{\partial I(x,y)}{\partial x}\right|, \left|\frac{\partial I(x,y)}{\partial y}\right|, \left|\frac{\partial^2 I(x,y)}{\partial x^2}\right|, \left|\frac{\partial^2 I(x,y)}{\partial y^2}\right|, \\ \left|\frac{\partial^2 I(x,y)}{\partial x \partial y}\right|, HOG(x,y), Gabor(x,y) \end{bmatrix}^T \quad (1)$$

where, the pixel position (x, y) and the gradient ($\left|\frac{\partial I(x,y)}{\partial x}\right|, \left|\frac{\partial I(x,y)}{\partial y}\right|$) are the basic characteristics of the target object, the gradient can provide key features such as texture, edge, and direction. Compared with [42], we increased the image pixel in the x and y directions of the second-order mixed partial derivative $\left|\frac{\partial^2 I(x,y)}{\partial x \partial y}\right|$ to capture local curvature change, capture complex and essential features such as eyebrows, corners of the eyes, and corners of the mouth more accurately. And $\left|\frac{\partial^2 I(x,y)}{\partial x \partial y}\right|$ of intersection is very sensitive and can detect the image of the cross and the nonlinear change, enhancing the model's perception of delicate features.

In the research [42], RGB and YCbCr are used as basic color features to enhance the representation of target objects in remote sensing scenes. This research focuses on the color distribution of different areas, such as skin tone, occlusions, etc. However, in a different light, the value of skin color in the RGB space will change significantly, especially when the non-uniform light performance is worse. In [58], the authors show that HSV is more robust to changes in facial illumination. Therefore, we capture color features in the HSV color space. Hue H and saturation S can describe skin color well and have certain invariances to brightness changes, making skin color feature extraction more reliable.

HOG counts the local gradient change around each pixel. In [59], the authors show that HOG features are susceptible to object deformation, so introducing HOG features can effectively extract facial edge and shape information. The Gabor feature is widely used for characterizing image texture information [60]. It effectively captures and discriminates textures because their frequency and orientation selectivity align with the response characteristics of the human visual system.

2) High-level feature learning: In this section, a multi-layer convolutional neural network (CNN) is used to extract high-level semantic features from the reference face image. Fig. 2 shows the high-level feature learning network structure used in this research. The semantic features of the reference face image are gradually extracted through multiple convolutions (including standard convolutions and parallel dilated convolutions) combined with batch normalization (BN) and SeLU activation functions. To extract the high-level features of the reference face image at a deeper level, three residual modules are used to deepen the network structure and reduce feature loss.

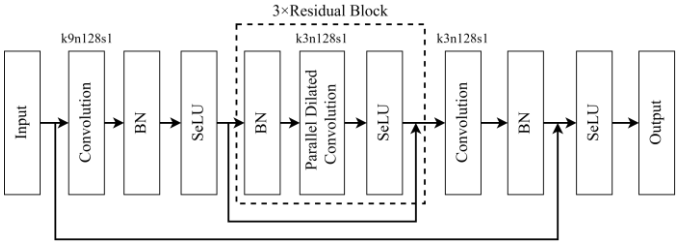

Fig. 2. High-level feature learning network. k, n, and s indicate kernel size, number of channels, and stride, respectively.

This study uses dilated convolutions with r=1, 2, and 3 dilation rates to perform parallel operations. As shown in Fig. 3, given a feature map Ft, it is segmented into four parts, Ftc1, Ftc2, Ftc3, and Ftc4, along the channel. A 3×3 convolution kernel is used to perform dilated convolution operations on two adjacent parts, so that each dilated convolution can share some parameters. Then the output of the dilated convolution is fused with the original feature. Finally, the number of channels is readjusted through a 1×1 convolutional network.

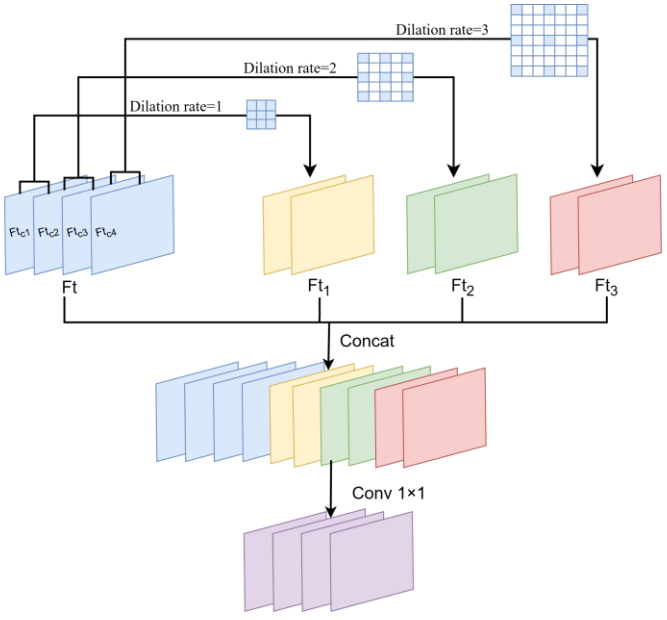

Fig. 3. The principle of parallel dilated convolution.

TABLE I. THE NUMBER OF PARAMETERS OF PARALLEL DILATED CONVOLUTION AND ORDINARY CONVOLUTION

Methods	Kernel Size	Input Channel	Output Channel	Layer	Parameters Size	Total(M)
Ordinary	3×3	512	512	Conv1 Conv2 Conv3	512×512×3×3=2,359,296 512×512×3×3=2,359,296 512×512×3×3=2,359,296	7,077,888 ≈ 7.08
	5×5	512	512	Conv1 Conv2 Conv3	512×512×5×5=6,553,600 512×512×5×5=6,553,600 512×512×5×5=6,553,600	19,660,800 ≈ 19.66
Parallel	5×5	512	512	Conv1 Conv2 Conv3	512×512×5×5=6,553,600	6,553,600 ≈ 6.55

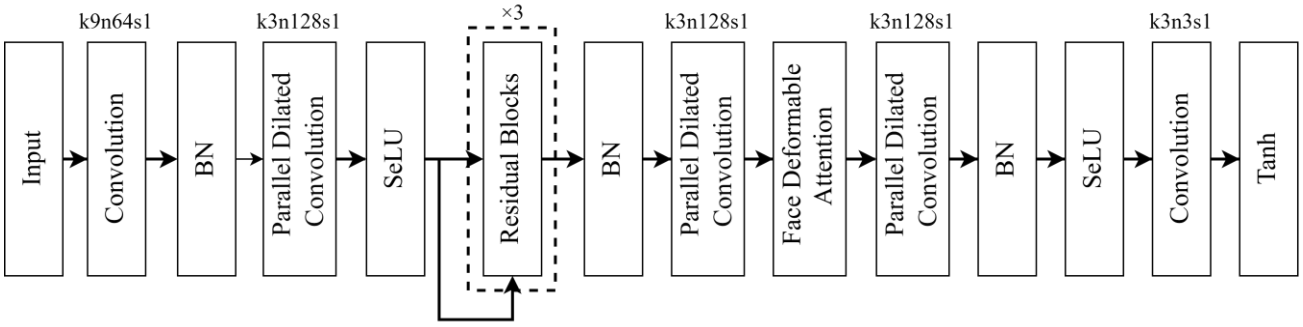

Fig. 4. The generator structure. k, n, and s indicate kernel size, number of channels, and stride, respectively.

Compared to traditional convolution, parallel dilated convolution effectively enlarges the receptive field and decreases the number of parameters by sharing parameters across branches with different dilation factors, as summarized in TABLE I. When the input and output channels are 512 and the convolution kernel size is 5×5, the parameter count of standard convolution is approximately three times higher than that of parallel dilated convolution. Notably, the parameter counts of parallel dilated convolution using a 5×5 convolution kernel are only 6.55M, even less than the parameters of three traditional convolution operations using a 3×3 convolution kernel. Therefore, the parallel dilated convolution effectively reduces the model size while ensuring the feature extraction effect.

C. Generator

1) Generator structure: In this section, a generator consisting of convolutional layers, residual blocks, and an attention module is constructed for face reconstruction under HMD occlusion, as presented in Fig. 4. This model proposes a face deformable attention module guided by a reference face image to reconstruct the face under HMD occlusion and generate finer textures in the occluded area. In the model, the BN layer is placed before the convolutional layer. This operation has been proven in research [45] to effectively speed up model convergence, improve gradient propagation, and enhance model generalization ability.

Three residual blocks are used in the generator to effectively avoid the gradient vanishing or gradient exploding problems that occur as the network depth increases. The residual block structure is presented in Fig. 5. Firstly, the input features undergo batch normalization, followed by parallel dilated convolution, which enlarges the receptive field while retaining fine details. To improve non-linear representation and ensure numerical stability, the SeLU activation function is applied. Finally, a skip connection directly adds the original input features to the convolutional outputs, mitigating feature loss as network depth increases. The formulation is as follows:

$$x_{i+1} = x_i + SeLU(PDConv(BN(x_i)) \quad (2)$$

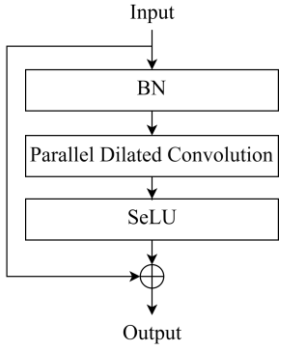

Fig. 5. The residual block network.

where, x_i denotes the input to a residual block, while x_{i+1} represents its output, which also serves as the input to the subsequent residual block, and $PDConv(\cdot)$ represents the parallel dilated convolution operation.

2) Face deformable attention: This section proposes a deformable attention mechanism guided by face reference images, as illustrated in Fig. 6. Given the HMD occluded face feature map $F_{masked} \in \mathbb{R}^{H\times W\times C}$, it is projected to query tokens Q by linear transformation. Given the reference face feature map $F_{ref} \in \mathbb{R}^{H\times W\times C}$, the reference points $p(x, y) \in \mathbb{R}^{H_r\times W_r}$

are generated by uniform grid sampling in F_{ref}. The reference points p are represented as 2D coordinates normalized to the range [-1,1] according to the mesh shape. By feeding Q into the offset network, a set of offset values $\Delta p = offset(Q)$ can be obtained. Adding the positional offset Δp to the reference points p, the reference points can be dynamically adjusted towards critical regions (e.g., the eye area), yielding new coordinate points $\hat{p}$. Sampling at points $\hat{p}$ on the reference feature map F_{ref} produces a new feature map $F_{ref}{}'$, which contains authentic reference features of the eye region under HMD occlusion. These features further support reconstructing a complete face under HMD occlusion. Key K and Value V can be obtained by linear transformation of $F_{ref}{}'$ using projection matrices W_k and W_v. The attention score is then obtained by calculating the similarity between Q and K, which represents the relevance of the query to each key. Finally, the attention score is normalized, and the values are weighted to generate the final weighted feature. Thus, the face information of the HMD occluded area can be recovered to help realize high-fidelity face reconstruction under HMD occlusion.

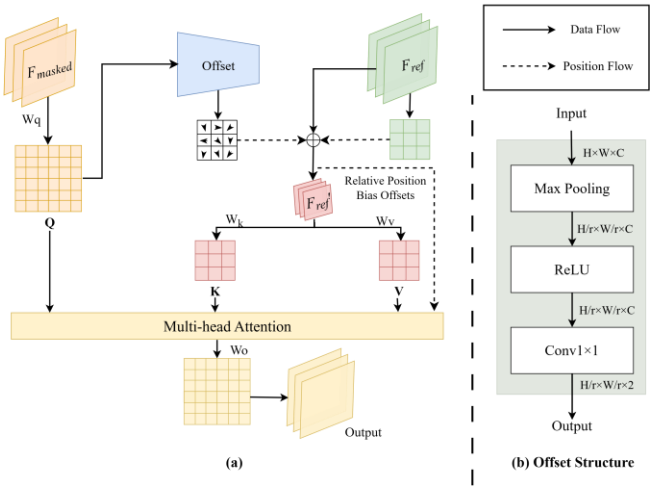

Fig. 6. Face deformable attention structure.

Fig. 6(b) shows the offset network structure. Q is downsampled through a Max pooling layer, and then nonlinearity is introduced through ReLU. The channels are compressed and mapped using 1×1 convolution. Finally, the position offset 2D coordinates are output. In the offset network, a predefined factor s is used for scaling to prevent the training process from being unstable due to excessive offset. Face deformable attention is calculated as follows:

$$Q = F_{masked}W_q,\ K = F_{ref}{}'W_k,\ V = F_{ref}{}'W_v \tag{3}$$

$$F_{out} = softmax\left(\frac{QK^T}{\sqrt{d}} + b\right)V \tag{4}$$

where, W_q, W_k, W_v are learnable projection matrices optimized during training, enabling the attention can effectively capture the relationships between input features. F_{masked} is the input HMD occluded face feature map, $F_{ref}{}'$ is the grid sampling result of the reference face feature map, and b is the relative position offset value.

D. Discriminator

The discriminator is constructed to differentiate reconstructed face images from real ones, as illustrated in Fig. 7. The reference face image and the reconstructed face image are randomly selected as the input of the discriminator. The input image first undergoes convolution to obtain initial feature representations, followed by the SeLU activation function to introduce nonlinearity. Subsequently, the network uses three parallel dilated convolutions to extract image features further and capture contextual information. The architecture combines average pooling to reduce spatial dimensions and aggregate features, and then uses a fully connected layer for advanced feature processing. Finally, through the Sigmoid activation, the input image is assigned a discrimination result.

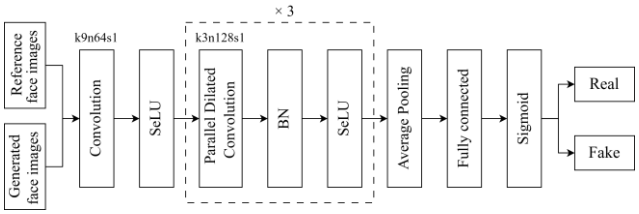

Fig. 7. The structure of the discriminator. k, n, and s indicate kernel size, number of channels, and stride, respectively.

The discriminator evaluates the input image and feeds the result back to the generator, guiding it to iteratively optimize its parameters and generate more realistic images. During adversarial training, the discriminator continuously updates its parameters to enhance its capability to differentiate between authentic and reconstructed images. Through this adversarial process, the generator gradually enhances the quality of its outputs, making the reconstructed images increasingly similar to real ones. Meanwhile, the discriminator delivers effective feedback to the generator by extracting deep feature representations, helping it to capture finer details and more complex features, thus facilitating better parameter optimization. Ultimately, the discriminator ensures that the facial images generated by the generator achieve high visual quality and detailed expression, effectively addressing the challenge of high-fidelity face reconstruction under HMD occlusion.

IV. Experimental Results and Analysis

A. Experimental Dataset

Due to the difficulty of obtaining real HMD-occluded faces with corresponding ground truth, synthetic datasets were used to ensure quantitative evaluation and reproducibility [32],[61]. The baseline datasets used in this study are FFHQ and CelebA. This study adopts FFHQ and CelebA, two widely used benchmark face datasets. FFHQ provides high-resolution (1024×1024) facial images with rich diversity in age, ethnicity, pose, and expression, making it suitable for high-fidelity reconstruction tasks and supervised learning with clean, unobstructed ground truth. CelebA offers a large-scale collection of face images with substantial identity and attribute variation, which supports evaluating the model's generalization ability across different facial characteristics. Based on these datasets, 68,274 synthetic HMD-occluded images were generated from FFHQ and 143,669 from CelebA. The model

was trained on the occluded FFHQ dataset and evaluated on the FFHQ and CelebA test sets to verify its generalization capability.

B. Experiment Setup

Experiments were conducted on a computing platform equipped with an NVIDIA RTX 4090 GPU (24 GB VRAM) and a 12th Gen Intel(R) Core(TM) i7-12700K CPU. The implementation was based on PyTorch 2.0.0 with CUDA 11.8 for GPU acceleration, using Python 3.8. Model training employed the Adam optimizer with first-order and second-order momentum estimates ($\beta1 = 0.5,\ \beta2 = 0.999$) and a batch size of 8. The learning rate for both the generator and discriminator was set to 0.0002, and training was performed for 100 epochs. The detailed experimental configuration is summarized in TABLE II.

TABLE II. EXPERIMENTAL SETUP

Category	Configuration
Hardware	NVIDIA RTX 4090 GPU (24GB), Intel Core i7-12700K CPU
Software	Python 3.8, PyTorch 2.0.0, CUDA 11.8
Optimizer	Adam($\beta1 = 0.5,\ \beta2 = 0.999$)
Batch Size	8
Learning Rate	0.0002
Epoch	100

C. Result and Analysis

1) Qualitative analysis: To verify the effectiveness of the proposed method in face reconstruction under HMD occlusion, this study conducts a qualitative analysis on the test set and compares the differences in visualization between the proposed approach and several widely used contemporary methods. Fig. 8 shows the face reconstruction results of different approaches in the FFHQ dataset.

As presented in the comparison results in Fig. 8, the existing methods can restore the basic contours of the occluded face to a certain extent, but there are generally problems such as blurred reconstruction areas, discontinuous boundaries, or structural distortion, especially in key areas such as the eyes and brow bones, which make it difficult to restore the true face accurately. As shown in Fig. 8(c), this method exposes the problems of global structure loss and regional incoherence in the generated results. Obviously, it lacks overall coordination, making it challenging to model facial semantics consistently. As shown in Fig. 8(d), MAT has enhanced local texture modeling by introducing the Transformer structure, but its structural alignment and detail continuity performance are still unsatisfactory. For example, in the 4th row and 4th column, there is a misalignment phenomenon in the reconstruction of the face contour under the side view, indicating that its modeling ability under complex postures still has room for improvement. AOT-GAN also exhibits similar problems, and from the visualization results in Fig. 8(e), it can be observed that its reconstruction results are obviously insufficient in diversity, and the eye structures generated in multiple samples are highly similar, lacking personalized performance. In addition, noticeable color differences and texture breakage often occur in the occluded area, affecting the overall visual consistency. In contrast, our approach is more natural, realistic, and diverse in visual effects. Meanwhile, our results also generated facial features and structures that were closely aligned with the ground truth, such as the distance between the eyebrows and eyes, the proportion of the facial features, the color of the eyes, etc. It can not only accurately restore the structural contours of the occluded area, but also show better coherence and consistency in fusion with the non-occluded area.

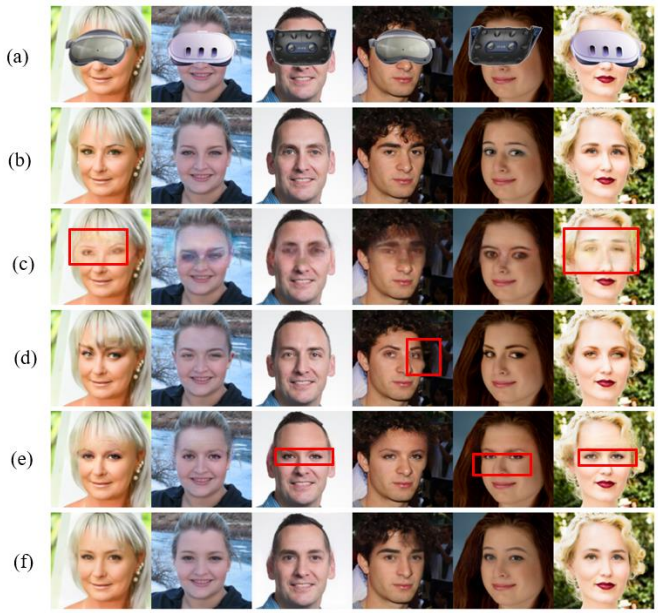

Fig. 8. Vision comparison. Each row is: (a) Input image, (b) Ground truth, (c) RFI[38], (d) MAT[62], (e) AOT-GAN[63], and (f) the proposed method.

2) Quantitative analysis: The proposed model is also verified on the CelebA dataset. To objectively measure the performance of the proposed method, this study uses three common image quality evaluation indicators: SSIM, PSNR, and LPIPS, and evaluates the results against existing leading image inpainting approaches. SSIM quantifies the structural similarity between the reconstructed and the original images. The higher the value, the closer the structure; PSNR primarily reflects the overall quality of image restoration. Higher values indicate lower image distortion; LPIPS evaluates image similarity based on the perceptual characteristics of the deep network, which can better capture the differences in human visual perception. The lower the value, the closer the reconstructed face is to the real face. We further report the number of parameters and FLOPs to verify the model's lightweight design and computational efficiency. FLOPs indicate the number of floating-point operations required for a single forward pass and serve as a measure of computational complexity. Since runtime can vary across different devices, this research did not use it as an evaluation metric.

A comparison of our method with multiple leading approaches is presented in TABLE III. Compared with mainstream GAN-based models, our method achieves higher SSIM and PSNR, as well as lower LPIPS, indicating superior reconstruction quality. Although RFI achieves lower FLOPs (5.67G) than ours (10.55G), its parameter count is much higher (128.87M). When compared with attention-based models such as AOR and MAT, our method achieves better image-level metrics while significantly reducing both parameter counts and FLOPs, demonstrating the efficiency of the face deformable attention module with dynamic sampling. Compared with lightweight GAN models like FastGAN, although our FLOPs (10.55G) are slightly higher than theirs (9.6G), our model has fewer than one-third of their parameters while achieving higher image quality, showing that we maintain a lightweight design without compromising reconstruction performance. Although diffusion-based methods have recently shown promising results, they were not included in the quantitative comparison due to the lack of publicly available code. Overall, these results indicate that our model effectively balances high-quality face reconstruction with a compact and computationally efficient design.

TABLE III. COMPARISON WITH STATE-OF-THE-ART RESEARCH

Research	SSIM	PNSR	LPIPS	Parameters (M)	FLOPs (G)
Lafin[64]	0.902	26.25	0.92	25.32	75.03
EC[65]	0.846	25.28	2.82	60.95	88.64
AOR[66]	0.918	29.02	0.07	43.75	100.27
EVI-HRnet[25]	0.899	25.83	0.03	26.01	62.31
MAT[62]	0.894	21.26	0.11	56.94	68.25
AOT-GAN[63]	0.901	22.01	0.11	15.20	18.22
RFI[38]	0.852	20.61	0.13	128.87	**5.67**
FastGAN [67]	0.865	22.34	0.51	9.44	**9.6**
PI-MFO-GAN [68]	0.878	27.50	0.18	50.63	42.65
MGAN-CRCM[69]	0.892	26.30	0.27	25.85	24.3
Ours	**0.912**	**30.64**	**0.03**	**2.23**	10.55

D. Ablation Study

To assess the effect of each component within the proposed model for HMD face reconstruction, an ablation study was carried out. Key components were selectively removed, and the effectiveness of the resulting model variants was systematically compared. TABLE IV. summarizes the experimental configurations and the corresponding outcomes, comprising the following three model variants:

w/o Lie: Remove the Lie group feature learning module and keep other components.

w/o FDA: Remove the Face Deformable Attention (FDA) module and keep other components.

Ours: The complete model, including two key components: Lie group feature learning and face deformable attention.

TABLE IV. PERFORMANCE COMPARISON: FULL MODEL VS. ABLATED VARIANTS

	SSIM↑	PNSR↑	LPIPS↓
w/o Lie	0.892	28.36	0.04
w/o FDA	0.843	23.48	0.08
Ours	**0.912**	**30.64**	**0.03**

The quantitative results demonstrate that removing the Lie group feature learning module leads to a notable decline in performance, with SSIM dropping to 0.892 and PSNR decreasing to 28.36. This indicates a weakened ability to restore structural information and preserve fine details. Upon further removal of the attention mechanism, the performance deteriorates, with LPIPS increasing to 0.080, reflecting a significant decline in perceptual quality. In contrast, the whole model consistently achieves superior results across all evaluation measures, underscoring the contribution and necessity of the proposed components.

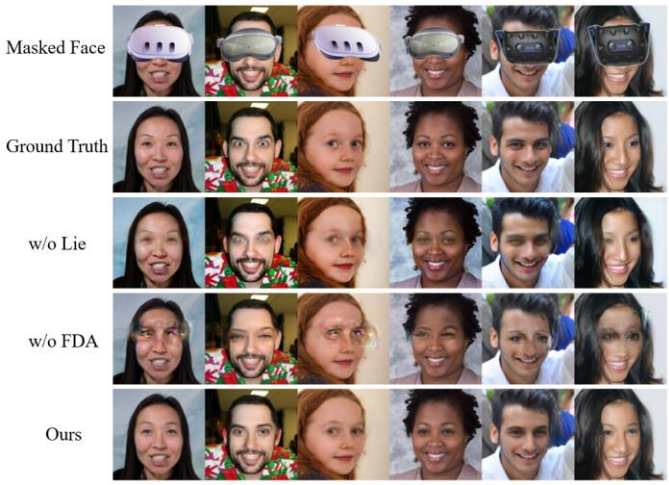

Fig. 9. Visual example of the ablation experiment.

To further illustrate the contribution of each module to the reconstruction performance, Fig. 9 presents a visual comparison of reconstructed results generated by different model variants on occluded facial images. In the w/o Lie configuration, the reconstructed outputs exhibit noticeable blurring and incomplete structural restoration in the occluded regions. The absence of fine-grained details highlights the importance of the Lie group feature learning module in providing low-level geometric and structural information. Without this module, the model struggles to capture local facial features, impairing its ability to reconstruct detailed and coherent textures. In the w/o FDA configuration, the results suffer from prominent artifacts and structural misalignments. Some examples display unnatural facial deformations, indicating that the proposed FDA module effectively guides the model to focus on critical facial regions and enhances its

ability to maintain global structural consistency by leveraging reference image information. The removal of this component hinders the integration of semantic cues, resulting in a noticeable decline in reconstruction quality. In contrast, the complete model achieves the most natural and structurally coherent visual results. It successfully restores precise facial details and maintains high texture and structural consistency across occluded and non-occluded regions. These observations further validate the synergistic effectiveness of combining Lie group feature learning with the face deformable attention mechanism in improving overall reconstruction performance.

In summary, the ablation study comprehensively demonstrates the substantial contributions of the proposed key modules to overall model performance, as evidenced by both quantitative metrics and qualitative visual comparisons. Although the Lie group feature learning module and the face deformable attention mechanism function through distinct mechanisms, they complement each other in enhancing structural restoration and maintaining visual coherence. Their synergistic integration constitutes a core strength of the model architecture. This design improves reconstruction quality under occlusion conditions and offers a scalable and generalizable framework for more complex facial restoration tasks in future applications.

V. Conclusion

This study proposes an innovative model based on GAN to solve the problem of face reconstruction under HMD occlusion. By introducing the Lie group feature learning module, the model effectively captures the geometric structure and local changes, enriches the feature representation, and enhances the robustness. In addition, we design a deformable attention mechanism guided by the reference face image to dynamically adjust the model's attention area, so that the model can not only accurately repair the occluded area, but also maintain the natural and realistic reconstruction effect, effectively improving the consistency and credibility of the generated image. The proposed model is verified on the synthetic FFHQ and CelebA datasets. Experimental results indicate that it outperforms existing methods across multiple evaluation metrics, including structural similarity, perceptual consistency, and model efficiency.

Despite these encouraging results, several limitations remain. First, the current model operates only at the single-image level and does not incorporate temporal information, which constrains its ability to generate temporally consistent reconstructions for dynamic HMD-occluded facial sequences. Second, although the architecture is designed to be lightweight, real-time deployment on resource-constrained standalone VR headsets may still require further optimization, particularly given that the present reconstruction results are limited to 2D image outputs. Future work will focus on further lightweight optimization and the incorporation of temporal cues to achieve consistent dynamic reconstruction, as well as exploring 3D-level facial texture completion.

Beyond the technical findings, this work also provides broader insights into the future development of HMD-occluded face reconstruction and VR realism. By demonstrating the value of combining Lie group–based geometric modeling with deformable attention, this study highlights a promising direction for designing generative models that balance structural fidelity with computational efficiency. Such geometry-aware lightweight architectures may reshape current approaches to face inpainting, offering an alternative to purely convolutional or diffusion-based pipelines. Furthermore, the ability to restore plausible and identity-consistent facial textures has potential implications for social VR, avatar expressiveness, and immersive communication, where natural facial cues directly influence user experience and presence. Although this research does not involve direct system deployment, the findings underscore the importance of sustainable and responsible facial synthesis, especially in an era where generative diffusion models and identity-sensitive applications continue to evolve. These reflections position the proposed framework not only as a technical contribution but also as a foundation for future interdisciplinary advances in VR-oriented facial reconstruction.

Acknowledgment

We would like to extend our heartfelt gratitude to the Ministry of Higher Education for the financial support received under the Fundamental Research Grant Scheme (FRGS), UTM, FRGS/1/2023/ICT10/UTM/02/2. Thanks to ViCubeLab at Universiti Teknologi Malaysia for their invaluable support and facilities.

This work is also partially supported by the Hengshui University Education and Teaching Reform and Research Project (jg2024051) and the Hebei Provincial Statistical Science Research Project (2025HL51).

References

[1] D. Wu, Z. Yang, P. Zhang, R. Wang, B. Yang, and X. Ma, "Virtual-reality interpromotion technology for metaverse: A survey," IEEE Internet of Things Journal, vol. 10, no. 18, pp. 15788–15809, 2023, doi: 10.1109/JIOT.2023.3265848.

[2] H. Wu and H. Tu, "USING DEEP LEARNING AND VIRTUAL REALITY TO BUILD AN ANIMATION GAME FOR THE HEALTHCARE EDUCATION," J. Mech. Med. Biol., vol. 23, no. 04, p. 2340052, May 2023, doi: 10.1142/S0219519423400523.

[3] A. A. Cantone et al., "Contextualized Experiential Language Learning in the Metaverse," in Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter, Torino Italy: ACM, Sep. 2023, pp. 1–7. doi: 10.1145/3605390.3605395.

[4] A. Dubiel, D. Kamińska, G. Zwoliński, B. Ramić-Brkić, D. Agostini, and M. Zancanaro, "Virtual reality for the training of soft skills for professional education: trends and opportunities," Interactive Learning Environments, vol. 33, no. 5, pp. 3261–3281, May 2025, doi: 10.1080/10494820.2025.2450634.

[5] J. Gao et al., "Pilot Study of a Virtual Reality Educational Intervention for Radiotherapy Patients Prior to Initiating Treatment," J Canc Educ, vol. 37, no. 3, pp. 578–585, Jun. 2022, doi: 10.1007/s13187-020-01848-5.

[6] E. Dhar et al., "A scoping review to assess the effects of virtual reality in medical education and clinical care," DIGITAL HEALTH, vol. 9, p. 20552076231158022, Jan. 2023, doi: 10.1177/20552076231158022.

[7] F. D. Chiesa, F. Bourhaleb, C. Pardi, and A. M. Soccini, "Virtual Reality Training for Advanced Radiotherapy," in 2024 IEEE Gaming, Entertainment, and Media Conference (GEM), IEEE, 2024, pp. 1–6. doi: 10.1109/GEM61861.2024.10585508.

[8] D.-I. D. Han, Y. Bergs, and N. Moorhouse, "Virtual reality consumer experience escapes: preparing for the metaverse," Virtual Reality, vol. 26, no. 4, pp. 1443–1458, Dec. 2022, doi: 10.1007/s10055-022-00641-7.

[9] S. Li, F. Yuan, and J. Liu, "Smart city VR landscape planning and user virtual entertainment experience based on artificial intelligence," Entertainment Computing, p. 100743, 2024, doi: 10.1016/j.entcom.2024.100743.

[10] L. G. Gilberto, F. Bermejo, F. C. Tommasini, and C. García Bauza, "Virtual Reality Audio Game for Entertainment & Sound Localization Training," ACM Trans. Appl. Percept., p. 3676557, Jul. 2024, doi: 10.1145/3676557.

[11] M. Barreda-Ángeles and T. Hartmann, "Psychological benefits of using social virtual reality platforms during the covid-19 pandemic: The role of social and spatial presence," Computers in Human Behavior, vol. 127, p. 107047, 2022, doi: 10.1016/j.chb.2021.107047.

[12] A. Gallace and M. Girondini, "Social touch in virtual reality," Current Opinion in Behavioral Sciences, vol. 43, pp. 249–254, 2022, doi: 10.1016/j.cobeha.2021.11.006.

[13] H. Tian, G. A. Lee, H. Bai, and M. Billinghurst, "Using virtual replicas to improve mixed reality remote collaboration," IEEE Transactions on Visualization and Computer Graphics, vol. 29, no. 5, pp. 2785–2795, 2023, doi: 10.1109/TVCG.2023.3247113.

[14] S. Zhu, W. Hu, W. Li, and Y. Dong, "Virtual Agents in Immersive Virtual Reality Environments: Impact of Humanoid Avatars and Output Modalities on Shopping Experience," International Journal of Human–Computer Interaction, vol. 40, no. 19, pp. 5771–5793, Oct. 2024, doi: 10.1080/10447318.2023.2241293.

[15] Z. Chen, Z. Zhang, J. Yuan, Y. Xu, and L. Liu, "Show Your Face: Restoring Complete Facial Images from Partial Observations for VR Meeting," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 8688–8697. doi: 10.1109/WACV57701.2024.00849.

[16] T. Combe, R. Fribourg, L. Detto, and J.-M. Normand, "Exploring the influence of virtual avatar heads in mixed reality on social presence, performance and user experience in collaborative tasks," IEEE Transactions on Visualization and Computer Graphics, 2024, doi: 10.1109/TVCG.2024.3372051.

[17] N. Numan, F. Ter Haar, and P. Cesar, "Generative rgb-d face completion for head-mounted display removal," in 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), IEEE, 2021, pp. 109–116. doi: 10.1109/VRW52623.2021.00028.

[18] S.-Y. Chen, Y.-K. Lai, S. Xia, P. L. Rosin, and L. Gao, "3D face reconstruction and gaze tracking in the HMD for virtual interaction," IEEE Transactions on Multimedia, vol. 25, pp. 3166–3179, 2022, doi: 10.1109/TMM.2022.3156820.

[19] C. Kim, H.-S. Cha, J. Kim, H. Kwak, W. Lee, and C.-H. Im, "Facial Motion Capture System Based on Facial Electromyogram and Electrooculogram for Immersive Social Virtual Reality Applications," Sensors, vol. 23, no. 7, p. 3580, 2023, doi: 10.3390/s23073580.

[20] A. Karande, C. Purps, J. Deuchler, D. Hepperle, and M. Wölfel, "Sensor-Based Occluded Face-Part Reconstruction: Eye Tracking and Facial Expressions," in 2023 IEEE 2nd International Conference on Cognitive Aspects of Virtual Reality (CVR), IEEE, 2023, pp. 000045–000050. doi: 10.1109/CVR58941.2023.10395459.

[21] Y. Xuan, V. Viswanath, S. Chu, O. Bartolf, J. Echterhoff, and E. Wang, "SpecTracle: Wearable Facial Motion Tracking from Unobtrusive Peripheral Cameras," Aug. 14, 2023, arXiv: arXiv:2308.07502. doi: 10.48550/arXiv.2308.07502.

[22] H. Guillen-Sanz, D. Checa, I. Miguel-Alonso, and A. Bustillo, "A systematic review of wearable biosensor usage in immersive virtual reality experiences," Virtual Reality, vol. 28, no. 2, p. 74, 2024, doi: 10.1007/s10055-024-00970-9.

[23] S. Aziz, D. J. Lohr, L. Friedman, and O. Komogortsev, "Evaluation of Eye Tracking Signal Quality for Virtual Reality Applications: A Case Study in the Meta Quest Pro," Mar. 11, 2024, arXiv: arXiv:2403.07210. doi: 10.1145/3649902.3653347.

[24] S. Gupta, S. S. Jinka, A. Sharma, and A. Namboodiri, "Supervision by Landmarks: An Enhanced Facial De-occlusion Network for VR-Based Applications," in Computer Vision – ECCV 2022 Workshops, vol. 13805, L. Karlinsky, T. Michaeli, and K. Nishino, Eds., in Lecture Notes in Computer Science, vol. 13805. , Cham: Springer Nature Switzerland, 2023, pp. 323–337. doi: 10.1007/978-3-031-25072-9_21.

[25] F. Ghorbani Lohesara, K. Egiazarian, and S. Knorr, "Expression-aware video inpainting for HMD removal in XR applications," in Proceedings of the 20th ACM SIGGRAPH European Conference on Visual Media Production, London United Kingdom: ACM, Nov. 2023, pp. 1–9. doi: 10.1145/3626495.3626497.

[26] S. Bai et al., "Universal Facial Encoding of Codec Avatars from VR Headsets," Jul. 17, 2024, arXiv: arXiv:2407.13038. doi: 10.48550/arXiv.2407.13038.

[27] J. Xu et al., "Personalized face inpainting with diffusion models by parallel visual attention," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 5432–5442. doi: arXiv:2312.03556v1.

[28] S. Kim, S. Suh, and M. Lee, "Rad: Region-aware diffusion models for image inpainting," in Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 2439–2448. Accessed: Sep. 22, 2025.

[29] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 9, pp. 10850–10869, 2023, doi: 10.1109/TPAMI.2023.3261988.

[30] H. Chen et al., "Comprehensive exploration of diffusion models in image generation: a survey," Artif Intell Rev, vol. 58, no. 4, p. 99, Jan. 2025, doi: 10.1007/s10462-025-11110-3.

[31] T. Yenamandra et al., "i3dmm: Deep implicit 3d morphable model of human heads," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12803–12813. doi: 10.1109/CVPR46437.2021.01261.

[32] C. F. Purps, S. Janzer, and M. Wölfel, "Reconstructing Facial Expressions of HMD Users for Avatars in VR," in ArtsIT, Interactivity and Game Creation, vol. 422, M. Wölfel, J. Bernhardt, and S. Thiel, Eds., in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 422. , Cham: Springer International Publishing, 2021, pp. 61–76. doi: 10.1007/978-3-030-95531-1_5.

[33] H. He, J. Liang, Z. Hou, H. Liu, and X. Zhou, "Occlusion recovery face recognition based on information reconstruction," MACHINE VISION AND APPLICATIONS, vol. 34, no. 5, Sep. 2023, doi: 10.1007/s00138-023-01423-0.

[34] C. Li, A. Morel-Forster, T. Vetter, B. Egger, and A. Kortylewski, "Robust model-based face reconstruction through weakly-supervised outlier segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 372–381. doi: 10.1109/CVPR52729.2023.00044.

[35] Y.-J. Ju, G.-H. Lee, J.-H. Hong, and S.-W. Lee, "Complete face recovery gan: Unsupervised joint face rotation and de-occlusion from a single-view image," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2022, pp. 3711–3721. doi: 10.1109/WACV51458.2022.00124.

[36] J. Yu, K. Li, and J. Peng, "Reference-guided face inpainting with reference attention network," Neural Comput & Applic, vol. 34, no. 12, pp. 9717–9731, Jun. 2022, doi: 10.1007/s00521-022-06961-8.

[37] X. Lu, R. Lu, W. Zhao, and E. Ma, "Facial image inpainting for big data using an effective attention mechanism and a convolutional neural network," Frontiers in Neurorobotics, vol. 16, p. 1111621, 2023, doi: 10.3389/fnbot.2022.1111621.

[38] W. Luo, S. Yang, and W. Zhang, "Reference-guided large-scale face inpainting with identity and texture control," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 10, pp. 5498–5509, 2023, doi: 10.1109/TCSVT.2023.3257271.

[39] A. Hassanpour, F. Jamalbafrani, B. Yang, K. Raja, R. Veldhuis, and J. Fierrez, "E2F-Net: Eyes-to-face inpainting via StyleGAN latent space," Pattern Recognition, vol. 152, p. 110442, 2024, doi: 10.1016/j.patcog.2024.110442.

[40] M. Lu and F. Li, "Survey on lie group machine learning," Big Data Mining and Analytics, vol. 3, no. 4, pp. 235–258, 2020, doi: 10.26599/BDMA.2020.9020011.

[41] H. Yang, H. He, W. Zhang, Y. Bai, and T. Li, "Lie group manifold analysis: an unsupervised domain adaptation approach for image classification," Appl Intell, vol. 52, no. 4, pp. 4074–4088, Mar. 2022, doi: 10.1007/s10489-021-02564-3.

[42] C. Xu, G. Zhu, and J. Shu, "A lightweight and robust lie group-convolutional neural networks joint representation for remote sensing scene classification," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–15, 2021, doi: 10.1109/TGRS.2020.3048024.

[43] H. Kumar, A. Parada-Mayorga, and A. Ribeiro, "Lie group algebra convolutional filters," IEEE Transactions on Signal Processing, 2024, doi: 10.1109/TSP.2024.3365950.

[44] C. Xu, G. Zhu, and J. Shu, "A combination of lie group machine learning and deep learning for remote sensing scene classification using multi-layer heterogeneous feature extraction and fusion," Remote Sensing, vol. 14, no. 6, p. 1445, 2022, doi: 10.3390/rs14061445.

[45] C. Xu, J. Shu, and G. Zhu, "Adversarial Remote Sensing Scene Classification Based on Lie Group Feature Learning," Remote Sensing, vol. 15, no. 4, p. 914, 2023, doi: 10.3390/rs15040914.

[46] L. Cai, C. Liu, R. Yuan, and H. Ding, "Human action recognition using Lie Group features and convolutional neural networks," Nonlinear Dyn, vol. 99, no. 4, pp. 3253–3263, Mar. 2020, doi: 10.1007/s11071-020-05468-y.

[47] X. Yang, X. Jia, D. Gong, D.-M. Yan, Z. Li, and W. Liu, "Larnet: Lie algebra residual network for face recognition," in International Conference on Machine Learning, in Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 11738–11750. Accessed: Nov. 21, 2024.

[48] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," Information Fusion, vol. 108, p. 102417, 2024, doi: 10.1016/j.inffus.2024.102417.

[49] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," ACM Comput. Surv., vol. 54, no. 10s, pp. 1–41, Jan. 2022, doi: 10.1145/3505244.

[50] M.-H. Guo et al., "Attention mechanisms in computer vision: A survey," Comp. Visual. Med., vol. 8, no. 3, pp. 331–368, Sep. 2022, doi: 10.1007/s41095-022-0271-y.

[51] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," Artif Intell Rev, vol. 57, no. 4, p. 99, Mar. 2024, doi: 10.1007/s10462-024-10721-6.

[52] K. Han et al., "A survey on vision transformer," IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 1, pp. 87–110, 2022, doi: 10.1109/TPAMI.2022.3152247.

[53] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 4794–4803. doi: arXiv:2201.00520v3.

[54] X. Pan, T. Ye, Z. Xia, S. Song, and G. Huang, "Slide-transformer: Hierarchical vision transformer with local self-attention," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 2082–2091. doi: 10.1109/CVPR52729.2023.00207.

[55] S. Yun and Y. Ro, "Shvit: Single-head vision transformer with memory efficient macro design," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5756–5767. doi: arXiv:2401.16456v2.

[56] W. Wan, S. Chen, L. Yao, and Y. Zhang, "Unsupervised masked face inpainting based on contrastive learning and attention mechanism," Multimedia Systems, vol. 30, no. 4, p. 209, Aug. 2024, doi: 10.1007/s00530-024-01411-y.

[57] S. Chen, A. Atapour-Abarghouei, and H. P. Shum, "HINT: High-quality inpainting transformer with mask-aware encoding and enhanced attention," IEEE Transactions on Multimedia, 2024, doi: 10.1109/TMM.2024.3369897.

[58] S. R. Prakash and P. N. Singh, "Background region based Face orientation prediction through HSV skin color model and K-Means clustering," Int. j. inf. tecnol., vol. 15, no. 3, pp. 1275–1288, Mar. 2023, doi: 10.1007/s41870-023-01174-1.

[59] A. B. Ahadit and R. K. Jatoth, "A novel multi-feature fusion deep neural network using HOG and VGG-Face for facial expression classification," Machine Vision and Applications, vol. 33, no. 4, p. 55, Jul. 2022, doi: 10.1007/s00138-022-01304-y.

[60] A. W. Muzaffar et al., "Gabor contrast patterns: A novel framework to extract features from texture images," IEEE Access, vol. 11, pp. 60324–60334, 2023, doi: 10.1109/ACCESS.2023.3280053.

[61] T. Lu, Z. Peng, X. Xing, X. Xu, and J. Pang, "A general method of realistic avatar modeling and driving for head-mounted display users," IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 3, pp. 916–925, 2021, doi: 10.1109/TCDS.2021.3080588.

[62] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10758–10768. doi: 10.1109/CVPR52688.2022.01049.

[63] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Aggregated contextual transformations for high-resolution image inpainting," IEEE transactions on visualization and computer graphics, vol. 29, no. 7, pp. 3266–3280, 2022, doi: 10.1109/TVCG.2022.3156949.

[64] Y. Yang, X. Guo, J. Ma, L. Ma, and H. Ling, "LaFIn: Generative Landmark Guided Face Inpainting," Nov. 26, 2019, arXiv: arXiv:1911.11394. doi: 10.48550/arXiv.1911.11394.

[65] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in Proceedings of the IEEE/CVF international conference on computer vision workshops, 2019, pp. 0–0. doi: 10.1109/ICCVW.2019.00408.

[66] S. Gupta, A. Shetty, and A. Sharma, "Attention based occlusion removal for hybrid telepresence systems," in 2022 19th Conference on Robots and Vision (CRV), IEEE, 2022, pp. 167–174. Accessed: Jan. 03, 2025.

[67] B. Liu, Y. Zhu, K. Song, and A. Elgammal, "Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis.," in iclr, 2021. doi: 10.48550/arXiv.2101.04775.

[68] Y. Li, X. Zhan, H. Li, and W. Zhang, "Selection and guidance: high-dimensional identity consistency preservation for face inpainting," Vis Comput, Nov. 2024, doi: 10.1007/s00371-024-03702-x.

[69] N. A. Asad et al., "MGAN-CRCM: a novel multiple generative adversarial network and coarse refinement-based cognizant method for image inpainting," Neural Comput & Applic, vol. 37, no. 7, pp. 5459–5480, Mar. 2025, doi: 10.1007/s00521-024-10886-9.