electronics

MDPI

Article

Visibility-Prior Guided Dual-Stream Mixture-of-Experts for Robust Facial Expression Recognition Under Complex Occlusions

Siyuan Ma, Long Liu *, Mingzhi Cheng, Peijun Qin, Zixuan Han, Cui Chen, Shizhao Yang and Hongjuan Wang *

School of New Media, Beijing Institute of Graphic Communication, Beijing 102600, China; 20240752012@stu.bigc.edu.cn (S.M.)

* Correspondence: liulong@bigc.edu.cn (L.L.); wanghongjuan@bigc.edu.cn (H.W.)

Abstract

Facial occlusion induces sample-wise reliability shifts in facial expression recognition (FER), where the usefulness of global context and local discriminative cues varies dramatically with the amount of visible facial information. Existing occlusion-robust FER studies often evaluate under limited or homogeneous occlusion settings and commonly adopt static fusion strategies, which are insufficient for complex and heterogeneous real-world occlusions. In this work, we establish a rigorous occlusion robustness evaluation protocol by constructing a fixed offline test benchmark with diverse synthetic occlusion patterns (e.g., masks, sunglasses, texture blocks, and mixed occlusions) on top of public FER test splits. We further propose a Dual-Stream Adaptive Weighting Mixture-of-Experts framework (DS-AW-MoE) that fuses a global contextual expert and a local discriminative expert via an occlusion-aware weighting network. Crucially, we introduce a facial visibility assessment as a task-agnostic prior to explicitly regulate expert contributions, enabling dynamic reallocation of model capacity according to input-dependent feature reliability. Extensive experiments on public datasets and the constructed occlusion benchmark demonstrate that DS-AW-MoE achieves more stable recognition under complex occlusions, characterized by a smaller and more consistent performance drop. To support reproducibility under dataset license constraints, we will release an anonymous, fully runnable repository containing the complete occlusion synthesis pipeline, evaluation protocol, and configuration files, allowing researchers to reproduce the benchmark after obtaining the original datasets.

Keywords: facial expression recognition; occlusion robustness; dual-stream features; adaptive weighting; feature fusion

Check for updates

Academic Editors: Silvia Liberata Ullo and Li Zhang

Received: 19 February 2026
Revised: 12 March 2026
Accepted: 12 March 2026
Published: 16 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Facial expression recognition (FER) is a critical research area within affective computing, aimed at enabling computer systems to interpret human emotional states [1]. Although deep learning methods have significantly enhanced FER performance in controlled experimental settings [2–4], the robustness of FER systems in real-world uncontrolled scenarios remains limited, particularly in the presence of facial occlusions [5,6]. In recent years, increasing attention has been devoted to occlusion-aware FER [7–9]; however, most existing studies focus on specific occlusion types or relatively constrained experimental conditions, while real-world occlusions are often heterogeneous and complex [8]. Mainstream FER datasets such as RAF-DB and FER+ [10,11] contain limited systematic coverage of complex

mixed occlusion scenarios, and cross-corpus studies further reveal significant performance degradation under domain shifts [6]. To provide a more rigorous robustness evaluation protocol, this study constructs a Self-Constructed Occlusion Dataset with diverse synthetic occlusion patterns inspired by occlusion-aware benchmark construction paradigms [12], including masks, sunglasses, texture blocks, and mixed occlusions.

Existing approaches for handling occluded facial features can generally be categorized into image restoration and feature enhancement strategies. Image restoration methods typically reconstruct occluded regions using generative adversarial networks (GANs), such as partial convolution-based inpainting [13] and edge-guided adversarial completion models [14]. While such approaches aim to recover missing regions, reconstruction artifacts in expression-relevant areas may negatively affect subsequent classification. In contrast, feature enhancement methods leverage attention mechanisms to emphasize visible and discriminative facial regions [4,7]. Nevertheless, conventional convolutional neural networks (CNNs) are inherently constrained by localized receptive fields, limiting their ability to model long-range semantic dependencies under severe occlusion [3]. Transformer-based architectures, such as Vision Transformer (ViT) [15], enable global dependency modeling via self-attention mechanisms originally introduced in [16]. However, the quadratic computational complexity of self-attention with respect to sequence length [16,17] imposes substantial computational overhead, which can restrict real-time deployment. Recent state-space model architectures, including Mamba and VMamba variants [18,19], attempt to address this limitation through linear-complexity sequence modeling. Furthermore, existing feature fusion strategies in emotion recognition often rely on static late-fusion or predefined attention mechanisms [20,21], lacking dynamic mechanisms to adaptively re-balance global and local features under varying occlusion severities.

To address these challenges, this paper proposes an end-to-end Dual-Stream Adaptive Weighting Mixture-of-Experts (DS-AW-MoE) framework to improve the robustness of facial expression recognition under complex occlusion conditions. Unlike existing facial expression recognition methods based on dual-stream architectures or mixture of experts models, the proposed DS-AW-MoE does not rely on static feature fusion or heuristic weighting strategies; instead, it explicitly models the variation in feature reliability under complex occlusion conditions via an occlusion-aware adaptive weighting mechanism. This framework dynamically adjusts the contribution ratio between global and local feature experts according to the occlusion status of the input image, thereby more effectively addressing the feature uncertainty caused by occlusions in real-world scenarios.

To explicitly clarify the methodological boundaries and position the novelty of this research relative to prior dual-stream and mixture-of-experts fusion methodologies, a definitive distinction is established between the adopted foundational architectures and the proposed original contributions. This study reuses established state-of-the-art visual backbones as internal feature extractors, specifically employing the VMamba architecture to model global topological dependencies and the ResNet-18-based Facial Occlusion Perception Module to capture local fine-grained regional details. Building upon these reused foundations, the core novelty of the proposed DS-AW-MoE framework fundamentally lies in three original dimensions. First, this work introduces the extraction of a continuous facial visibility score to serve as a distinct, task-agnostic prior. Instead of relying on implicit feature-level gating, this explicit prior directly orchestrates the dynamic allocation of weights between the global and local experts. Second, a comprehensive dynamic expert handover analysis is systematically conducted to interpretably demonstrate how the adaptive weighting network intelligently shifts its reliance from local convolutional features to global state-space contexts as the spatial occlusion severity intensifies. Third, to overcome the evaluation inconsistencies prevalent in existing literature, this study pro-

poses a progressive stochastic occlusion synthesis pipeline and publicly establishes the Self-Constructed Occlusion Dataset benchmark. This fixed offline occlusion benchmark strictly guarantees the fairness, reproducibility, and rigorous comparative assessment of facial expression recognition models under complex real-world occlusion scenarios.

2. Related Work

This section first reviews representative deep learning-based methods in FER, focusing on the challenges encountered under occlusion conditions in real-world scenarios. Subsequently, research related to mixture of experts models and multi-stream network architectures is presented, along with a review of commonly used facial expression recognition datasets and occlusion synthesis strategies.

2.1. Deep Learning and Visual State-Space Models in Facial Expression Recognition

The research paradigm of FER has progressively transitioned from traditional handcrafted feature extraction to deep learning-based approaches. Early studies primarily relied on CNNs for facial feature modeling, as demonstrated by Cheah et al. [2] and Patel and Ajoodha [22]. They attempted to enhance the model's capacity for capturing fine-grained texture features by integrating local binary patterns with CNNs. However, CNNs, limited by their local receptive fields, encounter challenges in occlusion scenarios: when key facial regions (such as the eyes or mouth) are occluded, the model finds it difficult to effectively capture long-range semantic dependencies across the occluded areas, leading to a notable decline in recognition accuracy. To mitigate the above issues, attention mechanisms and ViT have been progressively introduced into the FER field. Wang et al. [3] employ attention mechanisms to direct the model's focus towards unoccluded regions, while Lu et al. [23] and Hazmoune and Bougamouza [24] leverage the Transformer's self-attention mechanism to capture global contextual information. Although Transformers have demonstrated strong performance on unobstructed or mildly occluded datasets, their computational complexity increases quadratically with sequence length $(O(N^2))$, thereby imposing limitations on inference efficiency and real-time applicability. Furthermore, existing Transformer architectures generally employ static feature fusion strategies, which lack the ability to dynamically adjust the weighting between local detail and global semantic information in response to varying degrees of occlusion.

Furthermore, existing studies predominantly evaluate models using synthetic and simplistic occlusion datasets. These occlusion settings partially facilitate the analysis of model behavior but inadequately capture the challenges posed by complex, multimodal occlusions in real-world uncontrolled scenarios, thereby readily causing domain shift issues—that is, the generalization capability of models between training domains and practical application scenarios is constrained.

To address the issues of computational redundancy and insufficient dynamic adaptability, the recently proposed Vision Mamba visual state-space model offers a novel research paradigm for facial expression recognition under occlusion conditions. The Mamba architecture achieves linear computational complexity $(O(N))$ through a selective scan mechanism, ensuring efficient inference while maintaining strong global modeling capability [6,7]. Existing studies (such as Ma et al. [25] and Li et al. [26]) have preliminarily explored the application of Mamba in facial expression feature modeling and validated its potential for feature enhancement. However, these methods typically lack dedicated dynamic feature scheduling mechanisms specifically designed to handle semantic uncertainty under complex occlusion conditions.

In this context, the proposed DS-AW-MoE architecture introduces a collaborative modeling framework combining a Mamba-based global expert with a local texture expert,

and employs a learnable arbitration network to achieve adaptive weighting between local perception and global inference under varying occlusion conditions. Meanwhile, by integrating the constructed complex occlusion testing benchmark with the Progressive Stochastic Occlusion Augmentation Training Strategy, the model's robustness in real-world scenarios is further improved.

2.2. Research on Facial Expression Recognition Under Occlusion

Although existing high-performance affective computing models are predominantly developed under the assumption of relatively complete data [1,27], Cheheb [5]'s empirical study demonstrates that in uncontrolled real-world scenarios, facial occlusion is a widespread phenomenon that compromises the integrity of facial topology and negatively impacts model performance. To address this issue, related research has progressed from handling single, regular occlusions to investigating complex multimodal occlusions. Early studies predominantly concentrated on singular, regular occlusion types and sought to compensate for the absence of visual information via multimodal fusion strategies. For instance, Dixit and Satapathy [20] and Wang et al. [28] incorporated audio or physiological signals to enhance visual modeling, whereas Filho et al. [29] further examined the joint modeling of physiological signals and speech features. Within the purely visual domain, Jekauc et al. [30], Zhang and Chai [31], and Pei et al. [32] exploited deep convolutional networks and micro-expression analysis to extract higher-level semantic information. Liu et al. [21] proposed a multi-attention fusion method based on ensemble learning and a Transformer to alleviate the effects of local occlusion.

However, when confronted with complex and multimodal occlusion scenarios in real-world contexts, these methods still reveal certain limitations, primarily manifested in the inadequate dynamic adaptability of the feature fusion mechanism and the discrepancy between evaluation benchmarks and real-world applications. Firstly, most existing fusion strategies (such as late fusion [20] or simple feature concatenation) adopt a static design, maintaining constant dependency weights on local features and global context across varying degrees of occlusion. Existing studies have indicated that an ideal recognition system should possess decision interpretability and environmental adaptability; however, current methods struggle to effectively adjust inference strategies under severe occlusion. Additionally, the domain shift problem remains prevalent. Hu et al. [33] and Ryumina et al. [6] highlight the current lack of high-quality testing benchmarks encompassing diverse complex occlusion types, noting that model training primarily relies on simplified synthetic occlusion data, thereby limiting generalization capability in real-world scenarios.

Beyond affective computing, the challenges of visual occlusion are also prominent in real-world safety monitoring applications. For example, in construction-site environments, surveillance and body-worn camera imagery frequently suffer from severe partial occlusions of workers' faces and heads, complicating the reliable identification of non-PPE compliance. To address this, advanced domain-adaptive frameworks [34] have been explicitly developed to mitigate localized information losses in chaotic real-world conditions. These cross-domain challenges are highly analogous to the occlusion issues encountered in unconstrained facial expression recognition. Just as safety monitoring systems must dynamically adapt to visual degradation, our proposed DS-AW-MoE framework utilizes a visibility-prior guided dynamic weighting mechanism to reliably recover and balance occluded facial cues under similar real-world constraints.

To systematically synthesize the aforementioned research paradigms and to explicitly highlight the structural novelty and comparative advantages of the proposed methodology, a comprehensive comparison is presented in Table 1. Existing approaches for facial expression recognition under occlusion predominantly fall into three categories encompassing

local region-based attention, global context modeling, and generative restoration. While local attention mechanisms excel at extracting fine-grained features from unoccluded areas, they frequently suffer catastrophic failure when critical facial landmarks are entirely obscured. Conversely, global context models utilizing standard self-attention mechanisms capture long-range dependencies but inherently introduce quadratic computational overhead and remain highly susceptible to global token corruption caused by unstructured occlusion noise. Furthermore, generative restoration techniques involve high inference latency and often introduce synthetic artifacts that mislead the classification boundary. Overcoming these entrenched limitations, the proposed Dual-Stream Adaptive Weighting Mixture-of-Experts framework dynamically integrates a local convolutional expert with a global state-space expert. Guided by an explicitly learned visibility prior, this architecture adaptively balances fine-grained regional cues and global topological structures while maintaining linear computational complexity, thereby fundamentally resolving the robustness bottlenecks prevalent in existing single-paradigm methods.

Table 1. Summary and comparison of existing methodology paradigms for facial expression recognition under occlusion. The table explicitly outlines the core mechanisms and primary limitations of local, global, and generative approaches, contrasting them with the structural advantages of the proposed DS-AW-MoE framework.

Methodology Paradigm	Core Mechanism	Primary Limitations
Local Region Focus	Isolates visible facial components via landmark guidance or attention cropping.	Fails under severe occlusion when critical keypoints are entirely obscured.
Global Context Modeling	Captures long-range dependencies using (ViT) and self-attention.	Incurs quadratic computational overhead and vulnerability to unstructured occlusion noise.
Generative Restoration	Reconstructs missing facial pixels using GANs prior to classification.	Introduces synthetic artifacts, increases inference latency, and complicates end-to-end training.
DS-AW-MoE (Ours)	Integrates local and global experts via visibility-prior guided dynamic weighting.	Balances local fine-grained cues and global context with linear complexity, ensuring robust adaptation.

In summary, although existing methods have partially mitigated the impact of occlusion on facial expression recognition, there remain prevalent issues of limited dynamic adjustment capabilities within feature fusion mechanisms and insufficient model generalization performance under complex and variable real-world occlusion scenarios. Current research has not sufficiently addressed the critical challenge of adaptively balancing global structural information with local detailed features. To address these issues, this paper proposes a dual-stream adaptive weighting Mixture of Experts (MoE) framework. This approach employs a defensive training strategy combined with a dynamic weighting mechanism to adaptively allocate weights to global and local experts according to the occlusion status of the input image, aiming to enhance model robustness under complex occlusion conditions.

3. Method

This section provides a detailed exposition of the proposed DS-AW-MoE framework. The architecture is designed to address the inherent unreliability of fixed feature extraction under varying facial occlusions by dynamically balancing global structural context and local fine-grained details. We first present the overall architectural blueprint, followed by

a granular description of the individual experts and the adaptive gating mechanism that dynamically arbitrates between them based on facial visibility priors.

3.1. Overall Algorithm Framework

The proposed DS-AW-MoE framework is designed to address the issue of "feature reliability fluctuating with the availability of facial information" in facial expression recognition under complex occlusion conditions. Unlike traditional dual-stream or multi-branch networks that utilize fixed weights or static fusion strategies, this framework treats the global and local feature branches as two complementary experts whose reliability varies with the input, and employs a visibility-prior-based adaptive weighting mechanism to perform sample-level adjustments of each expert's contribution. This design enables the model to dynamically reallocate expert weights according to variations in effective information within input samples, thereby improving discrimination stability under diverse occlusion conditions.

As illustrated in Figure 1, the processing workflow of the DS-AW-MoE system primarily comprises three stages: the defensive augmentation phase, the dual-stream feature extraction stage, and the adaptive weighting fusion phase.

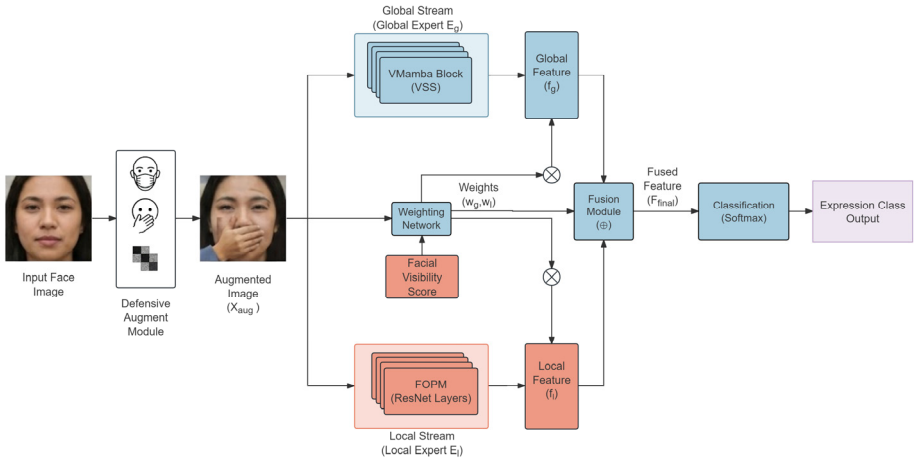


Figure 1. Pipeline of the proposed DS-AW-MoE for occlusion-robust facial expression recognition. The facial visibility score guides the weighting network to adaptively weight the global (VMamba) and local (FOPM) streams, whose features are fused for final classification. ⊗ denotes feature re-weighting by the visibility-aware gating weights.

Defensive augmentation phase: To reduce the model's overreliance on pristine textures, input images are first processed through an online augmentation module. This module generates training samples containing structured occlusions such as masks, biometric occlusions such as hands, and unstructured noise basGlobal streamed on a randomized mixing strategy, aiming to enhance the model's adaptability under conditions of information loss.

Dual-Stream Feature Extraction Phase: The augmented images are fed into two parallel expert networks. The global stream employs the VMamba architecture based on a state-space model to capture long-range dependencies and model the global topology of facial components. The local stream adopts the FOPM (Facial Occlusion Parsing Module) network to extract fine-grained discriminative texture features from visible facial regions.

Adaptive Weighting Fusion Phase: To integrate dual-stream representations, an adaptive weighting network is introduced. By incorporating the estimated facial visibility score, the network dynamically computes normalized weights for the global and local experts and performs weighted fusion to generate the final representation, enabling stable prediction under varying occlusion conditions.

Formally, let the input image be X, and the data augmentation function be $\Phi(\cdot)$. The augmented image representation is as follows:

$$X_{aug} = \Phi(X). \tag{1}$$

The global and local feature extractors are denoted as $E_g(\cdot)$ and $E_l(\cdot)$, respectively, with output feature representations:

$$f_g = E_g(X_{aug}), \tag{2}$$

$$f_l = E_l(X_{aug}) \tag{3}$$

where $f_g, f_l \in \mathbb{R}^D$.

To dynamically adjust the contribution of experts based on occlusion levels by introducing a weighting function $G(\cdot)$, whose inputs are the augmented image and the corresponding occlusion intensity score S_{vis}, and outputs a normalized weight vector $[w_g, w_l]$:

$$[w_g, w_l] = G(X_{aug}, S_{\text{vis}}), w_g + w_l = 1 \tag{4}$$

Here, $S_{\text{vis}} \in [0, 1]$ represents the facial visibility score, where a higher value indicates greater visibility of discriminative facial regions.

The final fused feature representation is:

$$F_{final} = w_g \cdot f_g + w_l \cdot f_l \tag{5}$$

and serves as the input to the subsequent classifier for expression prediction.

3.2. Global-Local Dual-Stream Feature Extraction Network

To balance the overall semantic consistency of facial expressions with the discriminative power of local fine-grained cues, this study designs a parallel dual-stream feature extraction architecture, as illustrated in Figure 2. This architecture comprises two structurally and functionally complementary branches: the global stream based on the visual state space model (VMamba) and the local stream based on the residual convolutional network (ResNet).

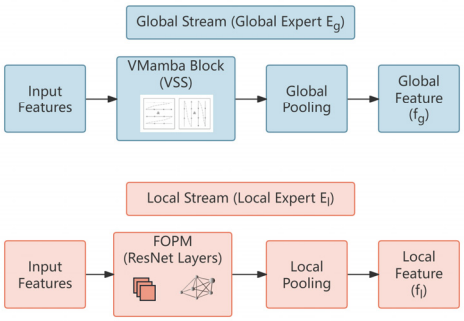


Figure 2. Dual-stream feature extraction module. The global stream (global expert E_g) employs VMamba (VSS blocks) followed by global pooling to produce the global feature f_g, while the local stream (local expert E_l) uses the FOPM (ResNet-based) branch with local pooling to generate the local feature f_l.

3.2.1. Global Stream: Long-Range Context Modeling Based on VMamba

In occlusion scenarios, the holistic facial topology and the correlation information between key regions are often incomplete. Traditional CNNs, limited by local receptive fields, struggle to adequately model cross-regional semantic dependencies when partial

information is absent. Although the ViT incorporates global modeling capability via the self-attention mechanism, its computational complexity increases quadratically with the input resolution ($O(N^2)$), certain limitations still remain in high-resolution or efficiency-critical application scenarios. To balance global modeling capability and computational efficiency, this paper introduces VMamba, based on a visual state space model, as the backbone network of the global stream. This model is constructed upon the State Space Model (SSM), preserving a global receptive field while its computational complexity scales linearly with the sequence length ($O(N)$), thereby efficiently capturing long-range dependencies between pixels or features. The Visual State Space Block (VSS Block) constitutes the core component of the global stream, with its theoretical foundation derived from the discretized form of continuous state space equations. For the input sequence $x(t)$ and output $y(t)$, whose evolution process can be described by the hidden state $h(t)$ described:

$$\begin{aligned} h'(t) &= \mathrm{A}h(t) + \mathrm{B}x(t) \\ y(t) &= \mathrm{C}h(t) \end{aligned} \tag{6}$$

In deep learning implementations, the aforementioned continuous equations are discretized using the Zero-Order Hold (ZOH) method, enabling the model to exhibit state update characteristics analogous to recurrent neural networks (RNN) during inference, while maintaining parallel computation similar to convolutional networks during training.

Since the standard state-space model primarily targets one-dimensional sequence modeling, making it difficult to apply directly to two-dimensional image features, this paper further introduces a two-dimensional Selective Scan mechanism (2D-Selective Scan, SS2D). This mechanism unfolds the two-dimensional feature map along multiple directions into one-dimensional sequences for scanning, thereby modeling long-range dependencies across spatial positions without incurring significant computational overhead. This omnidirectional scanning strategy effectively mitigates feature sensitivity to orientation, allowing the model to establish semantic relationships between non-adjacent facial regions such as forehead and chin. When local regions are occluded, it can still leverage the contextual information from the remaining visible areas for more robust global expression inference.

3.2.2. Local Stream: Fine-Grained Feature Perception Based on FOPM

The global stream excels at modeling overall structure and long-range dependencies; however, convolutional networks generally perform better in capturing micro-expressions or fine-grained texture cues such as wrinkles at the eye corners or subtle variations at the mouth corners. Therefore, this paper retains the lightweight local stream branch and adopts the FOPM architecture for fine-grained feature modeling. Because extracting fine-grained texture cues (such as micro-expressions and subtle variations around the eyes and mouth) inherently demands higher spatial resolutions, the standardized 112×112 system input is internally upsampled to 224×224 via bilinear interpolation before being processed by the FOPM local expert. This ensures sufficient spatial details are preserved for local discriminative learning.

The local stream employs a pre-trained ResNet-18 as the backbone network. Compared to deeper residual networks, ResNet-18 achieves a better balance between computational cost and texture feature representation. To preserve higher spatial resolution, this paper removes the last two downsampling stages of the original ResNet, thereby mitigating the loss of fine-grained expression cues caused by excessive downsampling.

Within the Local Perception Unit, the input image undergoes convolution and residual blocks to extract multi-channel feature maps, denoted as $F_{local} \in \mathbb{R}^{H \times W \times C}$, to enhance the response to key local regions and extract discriminative texture cues.

3.2.3. Adaptive Weighting Mixture-of-Experts Module

Conventional feature fusion methods typically assume that the importance of features from different branches is relatively fixed or weighted solely by static parameters. However, in facial expression recognition tasks under occlusion, the reliability of features varies significantly with changes in the visible facial information: when discriminative facial regions are highly visible, local texture features tend to provide more direct and discriminative cues; When facial visibility decreases and key information is progressively lost, relying solely on local details may become unstable. At this stage, global structure and contextual information provide more valuable references for expression inference.

To enable dynamic modeling of the aforementioned variations in feature reliability, this study constructs a MoE fusion module based on prior facial visibility information. Specifically, the dual-stream network is defined as two experts: (Global Expert, E_g) namely the VMamba branch, which outputs the feature vector $f_g \in \mathbb{R}^D$, and the Local Stream (Local Expert, E_l), namely the FOPM branch, which outputs the feature vector $f_l \in \mathbb{R}^D$.

The objective of the Mixture-of-Experts module is to learn a weight function $G(\cdot)$ that adaptively generates a set of normalized weights for any input sample $\mathbf{w} = [w_g, w_l]^T$, which are used to evaluate the reliability of different experts under the current occlusion conditions. The final fused feature is obtained through weighted summation, formally expressed as:

$$F_{final} = w_g \cdot f_g + w_l \cdot f_l, \tag{7}$$

where

$$w_g + w_l = 1 \tag{8}$$

Unlike traditional methods that estimate weights solely based on image features, this study introduces explicit yet automatically attainable occlusion prior information within the weighting decision process to enhance the sensitivity of weight allocation to occlusion variations, as illustrated in Figure 3.

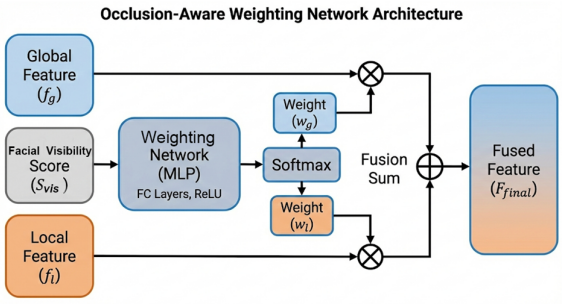


Figure 3. Architecture of the visibility-aware weighting network. The weighting multilayer perceptron (MLP) takes the facial visibility score s_{vis} as input and outputs normalized expert weights (w_g, w_l) via a Softmax layer $(w_g + w_l = 1)$. The final representation F_{final} is obtained by weighted fusion of the global feature f_g and local feature f_l. $\otimes$ denotes feature re-weighting by the visibility-aware gating weights.

Specifically, occlusion-related semantic features are initially extracted using a pre-trained occlusion detector f_{occ}, and simultaneously outputs the scalar facial visibility score s_{vis}, ranging from [0, 1], where a higher value indicates greater visibility of the discriminative facial regions.

The input vector of the weighting network is formed by concatenating the occlusion semantic features with the facial visibility score, and can be represented as:

$$v_{in} = \text{Concat}(f_{occ}, S_{vis}) \tag{9}$$

Subsequently, v_{in} is passed through a lightweight MLP, and the weights corresponding to the two experts are output via the Softmax function:

$$[w_g, w_l] = \text{Softmax}(\text{MLP}(v_{in})) \tag{10}$$

This design enables the weighting network to adaptively adjust the contribution ratio between the global and local experts based on variations in facial visible information within the input sample, thereby facilitating dynamic modeling of feature reliability and providing effective support for subsequent robust facial expression recognition.

3.2.4. Facial Visibility Estimation Module

To enable the Adaptive Weighting Network to explicitly perceive the availability of facial information in the input sample, this study introduces an independently trained facial visibility assessment module. This module is designed to quantify the visibility of discriminative facial regions within the input image and output a facial visibility score as a continuous scalar. This score functions as a critical prior signal for the weighting network, facilitating the dynamic allocation of weights between the global and local experts.

The module employs ResNet-18 as its backbone network, initialized with pretrained weights from ImageNet-1K. The global features extracted by the backbone are mapped to a single scalar through a lightweight regression head, which is constrained within the [0, 1] interval by a Sigmoid function. This output is defined herein as the facial visibility score, where a higher value indicates more complete effective facial information. Specifically, the module regresses the spatial occlusion ratio and mathematically transforms it into the visibility score to ensure consistency in semantic representation.

To ensure rigorous mathematical and notational consistency throughout the methodology, a clear distinction is established between the multidimensional image data and the quantitative occlusion metrics. The actual input to the facial visibility estimation module, as well as the main dual-stream network, is formally defined as the occluded facial image tensor denoted by the uppercase X_{aug}. This tensor contains the raw pixel-level information of the facial sample subjected to various synthetic occlusions. Conversely, the lowercase variable f_{occ} is strictly defined as the spatial occlusion ratio. It represents a continuous scalar value quantifying the exact proportion of the facial topological area covered by the synthetic mask during the data augmentation phase. By strictly decoupling the uppercase image tensor input from the lowercase scalar occlusion metric, the mathematical formulation of the proposed framework remains unambiguous across all analytical modules.

To ensure strict reproducibility and eliminate any algorithmic ambiguity, the ground-truth derivation of the continuous facial visibility score is mathematically and operationally formalized. The computation of this occlusion ratio is strictly executed at the pixel level following standard facial landmark detection, affine warping, and spatial alignment, which guarantees that all input facial images are uniformly normalized to a fixed spatial resolution of $H \times W$ pixels prior to any subsequent processing. During the progressive stochastic occlusion augmentation phase, a precise binary spatial mask $M \in \{0,1\}^{H \times W}$ is explicitly generated, where a pixel value of 1 denotes an artificially applied occlusion artifact, such as a synthetic mask or texture patch, and a value of 0 represents the pristine, unoccluded facial region. The exact ground-truth visibility score v is subsequently defined as the mathematical complement of the spatial occlusion ratio, formally computed as:

$$v = 1 - \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} M_{i,j} \tag{11}$$

Notably, this formalization calculates an unweighted, pure global pixel-wise proportion across the entire aligned facial bounding box without assigning heuristic importance

weights or attention biases to specific facial key regions such as the eyes or mouth. This deliberate unweighted design choice strictly preserves the task-agnostic nature of the proposed visibility prior, ensuring that the dynamic expert handover mechanism relies entirely on the objective spatial presence of physical occlusion rather than relying on subjective, pre-defined regional dependencies that frequently fail under extreme domain shifts.

To fundamentally prevent any data leakage during the subsequent evaluation phase, this module is trained strictly on the standard training split of the RAF-DB dataset, comprising 12,271 pristine facial images. The training process inherently relies on the proposed progressive stochastic occlusion synthesis pipeline. Pristine training images are dynamically augmented with diverse synthetic occlusions, including simulated medical masks, hands, and unstructured noise blocks. Utilizing the mathematically defined ground-truth visibility score v for regression supervision, the module completely eliminates the need for expensive manual annotations. By learning to regress this score via a Mean Squared Error (MSE) loss optimization, the module acquires the robust capability to accurately estimate facial visibility across diverse occlusion patterns.

Once the optimization is complete, the parameters of this visibility estimation module are strictly frozen. It subsequently serves as an off-the-shelf, independent visibility evaluator for the main dual-stream network, demonstrating strong zero-shot generalization when evaluated on other datasets such as RAF-DB and the Self-Constructed Occlusion Dataset (SCOD) benchmark without requiring any dataset-specific retraining. On the validation set, this module achieved a 2.54% Mean Absolute Error (MAE) specifically with respect to the predicted facial visibility score. Because the visibility score is mathematically defined as the exact complement of the spatial occlusion ratio, this low error magnitude inherently reflects the exceptional precision of the model in quantifying the unoccluded effective facial area. This quantitative performance demonstrates the ability of the module to accurately and consistently capture the variation trend of visible facial information, thereby providing a highly reliable prior input for the subsequent adaptive weighting network.

3.3. Dataset Construction

To address the limited coverage of occlusion types in existing datasets, which may induce domain shift, this paper proposes and implements a stochastic multi-modal occlusion augmentation strategy to generate diverse occluded samples during the training phase. Specifically, this study develops an online occlusion synthesis module: on one hand, it generates semantically consistent structured occlusion (such as masks covering the lower half of the face and sunglasses covering the eye region) based on facial keypoints and regional priors to simulate critical region loss in realistic wearing scenarios; On the other hand, geometric noise blocks with random positions, scales, and appearance parameters are generated to simulate conditions such as hand occlusion or occlusion by irregular objects. This online augmentation process is applied to training samples with a predefined probability, enabling the model to learn to exploit cross-regional contextual information under local information degradation, thus improving robustness in complex occlusion scenarios.

To ensure consistency in network input distribution and improve training stability, this study adopts a standardized data preprocessing pipeline. All images are first subjected to face detection and facial landmark alignment correction, followed by unified resampling to a resolution of 112×112, in order to mitigate the effects of scale variations. Subsequently, normalization based on ImageNet statistics is employed to reduce the impact of illumination and contrast variations on feature distributions, thus providing consistent inputs for downstream global modeling and local fine-grained feature extraction.

4. Experiments

In this section, we present a comprehensive empirical evaluation of the proposed DS-AW-MoE framework to substantiate its robustness and efficiency under complex facial occlusions. The experimental protocol is designed to validate the model's performance across both the synthetic SCOD benchmark and real-world occlusion scenarios. We begin by detailing the experimental configurations and dataset preparations, followed by a quantitative and qualitative analysis of the results relative to state-of-the-art baselines and modern visual recognition models.

4.1. Experimental Settings and Datasets

To evaluate the recognition performance and occlusion robustness of the proposed method across different scenarios, this study conducts experiments on two publicly available real-world facial expression recognition benchmark datasets (RAF-DB [10], FER+ [11]) as well as a self-constructed complex occlusion testing benchmark. A systematic comparison under both clean and occluded conditions is performed to thoroughly assess the model's stability in complex environments.

4.1.1. Benchmark Dataset

RAF-DB [10] (Real-world Affective Faces Database) contains 29,672 authentic facial images collected from the internet, exhibiting substantial diversity and pose variations. The official single-label split is used, including a training set of 12,271 images and a testing set of 3068 images, covering 7 basic expression categories.

FER+ [11] (Facial Expression Recognition Plus) is an extended version of FER2013, providing more consistent expression annotations through crowdsourcing. This dataset comprises 28,709 training images, 3589 validation images, and 3589 test images. In this work, face alignment and cropping are applied to the images, which are uniformly resized to 112×112 pixels for training and evaluation.

4.1.2. Self-Constructed Complex Occlusion Testing Benchmark

Most existing evaluation protocols rely on clean test sets, which fail to adequately reflect model robustness in real-world occlusion scenarios. Therefore, this study constructs, based on the RAF-DB [10] test set, a complex occlusion testing benchmark named SCOD. Unlike the online random occlusion augmentation applied during the training phase, SCOD is generated using an offline-fixed method and maintains consistent test samples across all methods to ensure fairness and reproducibility in comparative experiments. As shown in Table 2, SCOD includes three representative occlusion subsets: Subset A (Occ-Mask/Accessory) features daily life occlusions such as masks, sunglasses, and scarves overlaying the eye or mouth-nose regions, with the occlusion ratio controlled between 30% and 50%; Subset B involves hand or limb occlusions, with an occlusion ratio approximately between 20% and 40%; Subset C utilizes random occlusion covering 40–60% of the pixel area to simulate extreme information loss scenarios. The aforementioned test subsets are fixed and preserved for all subsequent comparative experiments.

Notably, the online random occlusion augmentation strategy applied during training differs significantly from SCOD's offline construction method in terms of occlusion distribution and parameter configuration. During the online augmentation process, occlusion position, scale, shape, appearance, and occlusion ratio vary randomly within a broad range and are dynamically generated across different training epochs, aiming to compel the model to adapt to highly uncertain occlusion patterns and thus enhance its generalization capability. In contrast, SCOD selects only several representative occlusion types and generates fixed test samples within relatively controlled parameter ranges, enabling fair

evaluation of different methods under unified conditions. This design of "broad coverage training augmentation combined with a fixed representative testing benchmark" helps mitigate the model's risk of overfitting to specific synthetic occlusion rules and allows for a more realistic assessment of its robustness in complex occlusion scenarios.

Table 2. SCOD benchmark built upon the official RAF-DB test set, including occlusion subsets, occlusion types, synthesis sources, and occlusion ratio ranges. The occlusion ratio denotes the approximate proportion of occluded pixels in an image.

Test Subset	Occlusion Type	Source	Occlusion Ratio	Difficulty
Original	None	RAF-DB Test Set	0%	Easy
Occ-Mask	Medical Masks, Sunglasses	Synthetic Overlay	30–50%	Medium
Occ-Hand	Hands, Arms	Texture Synthesis	20–40%	Hard
Occ-Noise	Random Black/White Blocks	Random Erasing	40–60%	Extreme

4.1.3. Implementation Details

All experiments are implemented using the PyTorch 2.0.0 deep learning framework and executed on a single NVIDIA RTX 4090 GPU for training and evaluation. The input image is initially processed using RetinaFace for face detection and keypoint alignment, followed by uniform resizing to a resolution of 112×112 pixels. To ensure feature standardization, the raw pixel intensities are normalized utilizing the universal ImageNet channel-wise mean and standard deviation metrics. During training, comprehensive data augmentation protocols are applied to enhance the spatial invariance of the model, which include standard random horizontal flipping alongside the progressive stochastic occlusion synthesis pipeline described in Section 3.

The global stream (VMamba) is initialized with ImageNet-1K pretrained weights, while the local stream (FOPM) uses ResNet-18 as the backbone network. The adaptive weighting network employs an orthogonal initialization strategy. The model is trained in an end-to-end manner using Adam as the optimizer. To provide robust structural regularization and prevent model overfitting, explicit weight decay is applied and set to 1×10^{-4}. A grouped learning rate strategy is adopted for different submodules, where the initial learning rate of the weighting network is set to 5×10^{-5}, while the learning rates for the global stream, local stream, and occlusion detector are all set to 5×10^{-6}. During training, a cosine annealing learning rate scheduling strategy is employed, with a batch size of 24 and training conducted over 300 epochs. Furthermore, to mitigate the severe optimization bias caused by the inherent long-tailed class distribution in facial expression datasets, a label-smoothed cross-entropy loss is employed as the primary objective function. This specific loss formulation deliberately softens the hard ground-truth annotations, thereby preventing the network from developing overconfident predictions on majority classes.

To rigorously guarantee the fairness and validity of all comparative evaluations presented in this study, a strictly unified experimental protocol is enforced across all baseline methodologies and the proposed framework. Specifically, every evaluated model, encompassing traditional convolutional networks, Vision Transformers, state-space models, and the generic YOLO architecture, processes input images subjected to the exact same preprocessing pipeline utilizing RetinaFace for spatial alignment followed by standardized spatial resizing to 112×112 pixels. Furthermore, to prevent any specific architecture from gaining an unfair data-level advantage, all compared methods are trained utilizing the identical progressive stochastic occlusion augmentation policy. The evaluation protocol is similarly strictly standardized, with all models being inherently tested on the exact same fixed offline splits of the established SCOD benchmark under identical hardware and software configurations. This comprehensive standardization completely eliminates

potential confounding variables introduced by inconsistent data processing or evaluation metrics, thereby ensuring that the observed performance discrepancies are exclusively attributable to the inherent architectural superiorities and dynamic weighting mechanisms of the respective models.

4.2. Comparative Experiments and Result Analysis

To comprehensively evaluate the effectiveness of the DS-AW-MoE, this study compares the proposed method with existing state-of-the-art methods as well as the baseline single-stream backbone within the framework.

4.2.1. Comprehensive Trade-Off Between Performance and Robustness

Table 2 presents a comprehensive comparison of the proposed method with other mainstream algorithms and internal baseline models on RAF-DB, FER+, and the SCOD. To quantify the model's defense capability against occlusion, this study introduces the Performance Drop Rate (Performance Drop Rate, $\Delta \downarrow$) metric:

$$\Delta = Acc_{clean} - Acc_{occluded} \tag{12}$$

where Acc_{clean} and $Acc_{occluded}$ denote the classification accuracies on the Clean Test Set and the occlusion test set, respectively. A smaller Δ value indicates less performance degradation under occlusion conditions.

Table 3 shows that the recognition accuracy of DS-AW-MoE on the Clean Test Set is comparable to that of certain single-stream models. Under complex occlusion conditions, the performance differences among various methods become more distinct. In contrast to the considerable performance degradation observed in single-stream models under occlusion scenarios, DS-AW-MoE exhibits more stable recognition performance across different occlusion types and intensities, with a Δ relatively low performance degradation rate. These results demonstrate that the dynamic adaptive weighting mechanism can modulate the contribution ratios of different feature branches based on the occlusion state, thereby alleviating changes in feature reliability induced by occlusion and enhancing robustness.

Table 3. Performance comparison of different methods on RAF-DB, FER+, and the SCOD occlusion benchmark. SCOD is constructed from the RAF-DB test set with fixed complex occlusions. "Drop rate ($\Delta\downarrow$)" denotes the relative performance degradation from the clean RAF-DB test set to SCOD (lower is better).

Method	Backbone	RAF-DB [10]	FER+ [11]	SCOD	Drop Rate ($\Delta\downarrow$)
RAN [7]	ResNet-18	86.90%	88.55%	-	
SCN [4]	ResNet-18	87.03%	88.01%	-	
FOPM [8] (Baseline)	ResNet-18	86.42%	-	75.30%	11.12%
VMamba [19]	Mamba	78.60%	-	71.06%	7.54%
DS-AW-MoE (Ours)	Dual-Stream	86.76%	88.95%	84.49%	2.27%

In addition to the overall accuracy on SCOD (Table 3), we further break down the robustness across different occlusion types. As shown in Figure 4, DS-AW-MoE consistently outperforms the single-stream baselines under mask, sunglasses, hand/object, and mixed/complex occlusions, indicating improved robustness across diverse occlusion patterns.

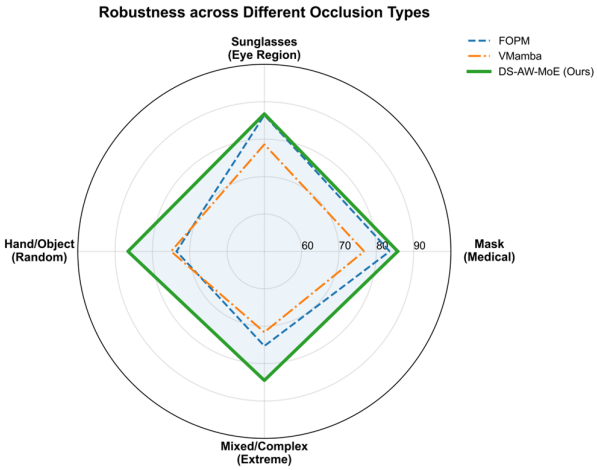


Figure 4. Radar chart comparing recognition accuracy (%) of FOPM, VMamba, and the proposed DS-AW-MoE across different occlusion types on the SCOD benchmark (mask, sunglasses, hand/object, and mixed/complex occlusions). Higher values indicate better robustness. This multi-dimensional evaluation highlights the models' generalization capabilities across heterogeneous real-world occlusion patterns. While single-stream baselines (FOPM and VMamba) exhibit pronounced vulnerabilities to specific regional information loss, the proposed DS-AW-MoE maintains a consistently superior and balanced performance boundary across all severe occlusion scenarios.

To further analyze the model's class discrimination behavior under complex occlusion conditions, Figure 5 presents the confusion matrix of DS-AW-MoE on the SCOD test set. It is observed that most expression categories maintain high diagonal responses under occlusion conditions, indicating that the model's discriminative capability for the primary categories remains relatively stable. Certain emotion categories (e.g., Fear, Disgust, and Sad) still exhibit some degree of confusion, which is associated with the similarity of their local facial cues under occlusion conditions. Overall, this result further substantiates the stability and validity of the proposed method under complex occlusion scenarios.

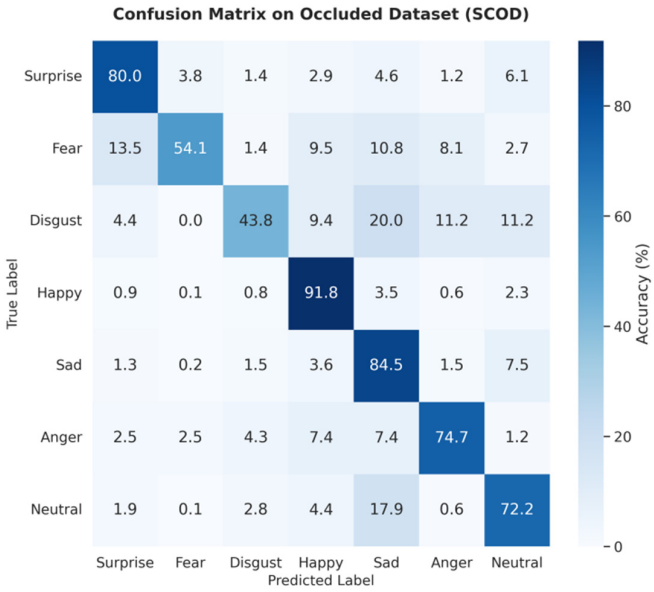


Figure 5. Confusion matrix of the proposed DS-AW-MoE on the SCOD occluded test set. Values are row-normalized percentages (each row sums to 100%), and darker colors indicate higher classification accuracy.

To provide a granular analysis of the classification performance across different emotion categories, a normalized confusion matrix is constructed. This confusion matrix

is entirely derived from the evaluation outcomes on the exactly 3068 standardized test samples of the SCOD benchmark. By visualizing the prediction distribution across this specific test split, the matrix effectively elucidates the precise misclassification tendencies of the model when confronting severe synthetic and real-world occlusions, particularly highlighting the performance bottlenecks associated with minority classes such as happy and sad.

The facial visibility assessment module is trained and validated on an independent dataset, with validation set performance serving as the basis for model selection. During validation, the module achieved a Mean Absolute Error (MAE) of 2.54% on the facial visibility regression task, demonstrating its ability to accurately estimate facial visibility in input samples, thus providing a stable and reliable prior signal for the subsequent Adaptive Weighting Network.

4.2.2. Computational Complexity and Inference Efficiency

To thoroughly evaluate the practical applicability of the proposed DS-AW-MoE framework, we conduct a comprehensive trade-off analysis between computational complexity and multi-dimensional recognition performance on the SCOD benchmark, as summarized in Table 4.

Table 4. Comprehensive trade-off analysis of computational complexity and multi-dimensional recognition performance for the single-stream baselines and the proposed DS-AW-MoE framework on the SCOD benchmark.

Method	Params (M)	Computation (GFLOPs)	Inference Speed (FPS)	Accuracy	Macro Precision	Macro Recall	Macro F1-Score
FOPM [8] (Baseline)	13.15	1.825	424.57	75.30%	63.21%	56.80%	59.03%
VMamba [19]	4.11	0.5764	31.15	71.06%	70.44%	57.16%	61.39%
YOLOv8n-cls [35]	1.44	3.30	~2500	78.00%	70.57%	67.10%	68.57%
DS-AW-MoE (Ours)	28.68	~2.90 *	27.55	84.49%	80.88%	73.79%	75.87%

* Note: The total GFLOPs of the integrated DS-AW-MoE system is theoretically estimated by aggregating the operations of the dual experts and the occlusion detector, accounting for the internal spatial interpolation (from 112×112 to 224×224) within the local stream.

We first measure the parameters (Params), floating-point operations (FLOPs), and inference frames per second (FPS) alongside the comparative baselines. The inference speed is measured on a single NVIDIA RTX 4090 GPU with a batch size of 1. As shown in Table 4, the linear complexity of the VMamba backbone ensures that the global expert remains highly lightweight, requiring only 4.11 M Params and 0.57 GFLOPs. Additionally, the general-purpose YOLOv8n-cls baseline exhibits exceptional efficiency with 1.44 M Params and an inference speed exceeding 2500 FPS. Although the proposed dual-stream architecture inherently introduces additional parameters, totaling 28.68 M, by integrating the local expert, global expert, and occlusion detector, the system maintains a highly competitive overall computational footprint. More importantly, the complete integrated system achieves a real-time inference speed of 27.55 FPS, demonstrating that the framework provides significant robustness improvements without sacrificing practical deployment capabilities.

Furthermore, given the inherent severe class imbalance in real-world facial expression datasets, relying solely on overall accuracy masks the performance discrepancies across minority categories. Therefore, we introduce Macro-averaged Precision, Recall, and F1-score to treat all emotion classes equally. As revealed in Table 4, while the baseline FOPM achieves an overall accuracy of 75.30%, its Macro F1-score severely drops to 59.03%, indicating catastrophic classification failure on hard and minority classes under occlusion. Similarly, the VMamba baseline only achieves a 61.39% Macro F1-score. Even the highly

efficient YOLOv8n-cls baseline, despite its generic feature extraction capabilities, achieves a Macro F1-score of only 68.57%, demonstrating its vulnerability to localized information loss. In stark contrast, the proposed DS-AW-MoE framework not only boosts the overall accuracy to 84.49% but also dramatically improves the Macro F1-score to 75.87%. This substantial improvement in Macro metrics definitively confirms that the visibility-prior guided dynamic weighting mechanism effectively mitigates occlusion-induced performance degradation, ensuring stable and robust recognition across all diverse emotion categories.

4.2.3. Generalization on Real-World Occlusions

While the SCOD benchmark provides a highly controlled environment for systematically evaluating model robustness against specific occlusion ratios, it inherently relies on synthetic perturbations. To rigorously verify the generalization capability and practical applicability of the proposed DS-AW-MoE framework in natural scenarios, additional evaluations are conducted on the Occlu-FER [36]. The Occlu-FER dataset comprises authentic in-the-wild images featuring natural occlusions such as real medical masks, authentic sunglasses, and natural hand-to-face interactions, presenting a severe domain shift from purely synthetic augmentations.

The performance comparisons on the Occlu-FER dataset validate the superior generalization of the proposed dual-stream architecture, as summarized in Table 5. The single-stream FOPM and VMamba baselines suffer from notable performance degradation when confronted with the unpredictable textures and irregular boundaries of real-world occlusions, achieving overall accuracies of 57.18% and 53.26%, respectively. In stark contrast, the DS-AW-MoE framework achieves an accuracy of 80.36% and maintains a highly competitive Macro F1-score of 76.28%. This substantial improvement confirms that the visibility-aware dynamic weighting mechanism successfully transfers the robustness learned from stochastic augmentation to real-world applications. The network effectively identifies severely corrupted facial regions in authentic images and adaptively shifts the inference reliance toward the global structural representations provided by the state-space model, ensuring stable recognition without requiring further dataset-specific fine-tuning.

Table 5. Performance comparison of the baseline models and the proposed DS-AW-MoE framework on the dataset containing real-world occlusions. This evaluation assesses the generalization capability of the methods when confronted with authentic masks and hand-to-face interactions.

Method	Backbone	Accuracy	Macro F1-Score
FOPM [8] (Baseline)	ResNet-18	57.18%	51.92%
VMamba [19]	Manba	53.26%	48.19%
DS-AW-MoE (Ours)	Dual-Stream	80.36%	76.28%

4.3. Ablation Experiments and Analysis

To verify the effectiveness of each core component of the proposed framework, a series of incremental ablation experiments was conducted on the SCOD, as summarized in Table 6. The experiments use the single-stream FOPM network as the baseline, sequentially incorporating defensive data augmentation (DA), an auxiliary global stream, and the Adaptive Weighting Mechanism.

Table 6. Incremental ablation study of defensive augmentation, VMamba global stream, and adaptive weighting on the SCOD benchmark. The symbol '√' denotes that the corresponding component is included in the model setup, while '×' denotes that it is excluded.

Model	Setup	Defensive Aug.	VMamba Stream	Adaptive Weighting	SCOD (Occluded)	Δ (Drop)
A	Baseline (FOPM)	×	×	×	75.30%	11.12%
B	+Augmentation	√	×	×	79.15%	6.75%
C	+MoE Weighting (Ours)	√	√	√	84.49%	2.27%

4.3.1. Contribution Analysis of Core Components

(1) After introducing the random occlusion augmentation strategy, the model's accuracy on SCOD improved from 75.30% to 79.15%. This result demonstrates that incorporating diverse occlusion patterns during training effectively mitigates the feature distribution mismatch encountered during testing. Simultaneously, the performance on the Clean Test Set exhibits slight variations, potentially attributable to the regularization effects introduced by augmentation.

(2) Furthermore, after integrating the VMamba global branch, recognition performance under occlusion conditions is further improved, indicating that global structural information and local texture features have complementary characteristics within the feature space, and the global branch can provide additional contextual support when local information is compromised.

(3) By substituting static fusion with a dynamic adaptive MoE weighting mechanism, the model's accuracy on the SCOD occlusion benchmark further increases to 84.49%, accompanied by a reduction in the performance degradation rate Δ. It is further reduced compared to the previously described configuration. The result demonstrates that, compared to static fusion, the dynamic weighting strategy can adaptively adjust the contributions of different experts based on occlusion conditions, thereby improving stability in complex occlusion scenarios.

Further analysis indicates that merely incorporating multi-branch structures or global features is insufficient to guarantee stable performance under severe occlusion, whereas the dynamic adaptive weighting mechanism is pivotal in suppressing corrupted features and enhancing the contribution of reliable features. This finding indicates that improvements in robustness derive not only from feature diversity but also crucially depend on the model's dynamic ability to assess feature reliability.

4.3.2. Dynamic Responsive Behavior Analysis of the Adaptive Weighting Mechanism

To analyze the behavioral characteristics of the adaptive weighting mechanism under varying facial information availability conditions, this study statistically analyzed and visualized the relationship between expert weights distribution and the Facial Visibility Score of samples in the test set, as shown in Figure 6. It should be noted that this score characterizes the visibility of discriminative facial regions in the input image; a higher value indicates more complete facial information, whereas a lower value corresponds to more severe information loss caused by occlusion.

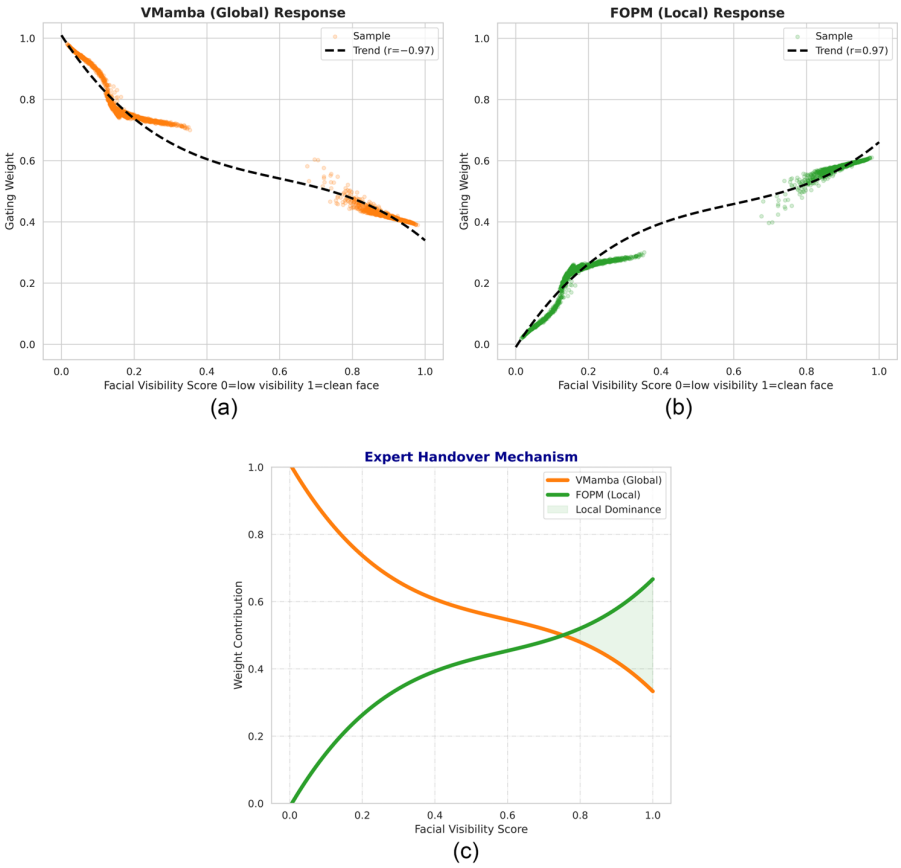


Figure 6. Dynamic responsive behavior analysis of the adaptive weighting mechanism under different facial visibility conditions. (**a**) The weight of the global expert (VMamba) exhibits a significant negative correlation with the Facial Visibility Score ($r = -0.97$); (**b**) The weight of the local expert (FOPM) increases as the Facial Visibility Score rises; (**c**) The expert handover mechanism illustrates the crossover region and smooth transition process of dynamic weights. The Facial Visibility Score ranges from 0 to 1, where 0 denotes minimal visible facial information (severe occlusion), and 1 denotes highly visible or nearly unobstructed facial information.

Experimental results demonstrate that with increasing Facial Visibility Score, the weight of the global expert (VMamba) generally decreases, while that of the local expert (FOPM) gradually increases, both showing consistent and stable response patterns. Correlation analysis further validated this trend: the global expert weights exhibit a significant negative correlation with the facial visibility scores (e.g., $r \approx -0.97$), whereas the local expert weights show a positive correlation with the visibility scores. This indicates that the weighting network can effectively perceive variations in facial visibility and accordingly adjust the contribution proportions of different experts. Around a facial visibility score of approximately 0.75, the weight curves of the two experts intersect, indicating a smooth handover of decision emphasis across different information-availability regimes. When facial visibility is high (i.e., facial cues are largely preserved), the weighting network assigns a larger contribution to the local expert, allowing the model to exploit fine-grained expression-related patterns. As visibility decreases and discriminative facial regions are progressively lost, the weighting network increases the reliance on the global expert, which is more capable of maintaining robust inference from coarse structure and broader contextual cues. This monotonic and interpretable response aligns with the correlation analysis (global weight negatively correlated with visibility; local weight positively correlated), suggesting that the visibility prior provides an effective signal for sample-wise expert re-allocation under complex occlusions.

Overall, the results demonstrate that the adaptive weighting mechanism does not generate expert weights randomly but produces consistent, continuous, and interpretable responses in accordance with changes in facial visible information. The smooth transition of expert weights reflects the weighting network's effective modeling of feature reliability variations, providing a clear mechanistic explanation and empirical evidence for feature fusion under complex occlusion conditions.

4.4. Qualitative Analysis and Failure Cases

To provide a deeper understanding of the practical behavior of the proposed DS-AW-MoE framework, qualitative visualizations of representative classification outcomes on the SCOD benchmark are presented in Figure 7. The top row of this figure illustrates the robustness of the framework by highlighting successful recognition cases. Even under conditions where the lower face is heavily obscured by masks or hands, the dynamic weighting mechanism effectively utilizes the global expert to extract holistic patterns from the remaining visible regions, such as the eyes and forehead. This capability enables the model to yield correct predictions for complex emotional states, including surprise and sadness.



Figure 7. Qualitative visualizations of representative classification outcomes by the proposed DS-AW-MoE framework on the SCOD benchmark. The top row illustrates success cases where the model accurately predicts the expression despite severe facial occlusions. The bottom row displays typical failure cases, reflecting the challenges posed by extreme information loss and inherent label ambiguity.

To critically evaluate the limitations of the proposed approach and its practical applicability, the incorrect predictions displayed in the bottom row of Figure 7 are analyzed across several primary dimensions encompassing methodological constraints, dataset issues, optimization settings, and other relevant factors. From a methodological perspective, the framework encounters feature representation limits when faced with extreme occlusions covering the vast majority of the facial area. Under such severe information loss, the model capacity is inherently insufficient to reconstruct the destroyed local fine-grained cues and global topological structures, which frequently causes the adaptive weighting mechanism to regress to a low-confidence neutral prediction. Furthermore, dataset-related issues significantly contribute to misclassifications. Human expressions are naturally nuanced, and the insufficient diversity of certain edge-case expressions combined with inherent label noise often leads to semantic ambiguity. In these instances, the inherent subjectivity present in real-world datasets emerges as a primary source of error.

Additionally, optimization and training settings pose inherent challenges under long-tailed data distributions. Consistent with the quantitative metric analysis indicating a lower

Macro F1-score compared to overall accuracy, the standard cross-entropy loss design struggles to maintain uniform convergence across all categories, causing the model to exhibit a slight optimization bias toward majority classes. Consequently, challenging minority class samples, such as those representing fear, are occasionally misclassified into dominant categories like happiness. Finally, considering other contributing factors, the subtle domain gap between the synthetic occlusion distributions generated during training and highly irregular real-world occlusions can also induce unexpected feature disruptions. Although the current training protocol incorporates label smoothing and stochastic augmentation, completely overcoming these optimization biases and representation limits under extreme occluded conditions remains an open challenge for future research.

5. Conclusions

This study addresses the robustness challenge of FER under complex occlusion conditions by proposing a DS-AW-MoE framework. The proposed method explicitly models feature reliability variations induced by occlusion and dynamically adjusts the contribution of global and local experts according to the estimated occlusion state. Experimental results on public datasets and the constructed SCOD benchmark demonstrate that the proposed framework maintains stable recognition performance across multiple occlusion types and exhibits reduced performance degradation under severe occlusion.

In terms of computational characteristics, the incorporation of the VMamba architecture with linear complexity enables global contextual modeling while controlling computational overhead relative to Transformer-based approaches. The adaptive weighting module is lightweight and introduces limited additional cost compared with the backbone networks, thereby maintaining an efficient inference structure within the dual-stream design.

The SCOD benchmark is constructed based on the official RAF-DB testing set for research evaluation purposes. Due to data licensing constraints, original images are not redistributed; instead, the complete occlusion synthesis pipeline and generation scripts will be released upon acceptance to support reproducibility and subsequent research.

Although the proposed framework improves robustness under complex occlusions, the dual-stream architecture and occlusion-aware weighting mechanism inevitably increase structural complexity. Future work will focus on architecture compression and efficiency optimization, such as multi-branch reduction through knowledge distillation or expert pruning. Further validation in more diverse real-world environments will also be conducted to examine the generalization capability of the proposed method.

Author Contributions: Conceptualization, S.M. and L.L.; methodology, S.M. and L.L.; software, S.M.; validation, M.C., P.Q. and C.C.; formal analysis, S.M., H.W. and M.C.; investigation, S.M., P.Q. and S.Y.; resources, L.L. and H.W.; data curation, S.M., P.Q. and Z.H.; writing—original draft preparation, S.M.; writing—review and editing, L.L., M.C. and H.W.; visualization, S.M.; supervision, L.L. and H.W.; project administration, L.L.; funding acquisition, L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Beijing Institute of Graphic Communication (Grant Nos. Eb202306, HXDK2024052, and HXDK2024133).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
DS-AW-MoE	Dual-Stream Adaptive Weighting Mixture-of-Experts
SCOD	Self-Constructed Occlusion Dataset
FOPM	Facial Occlusion Parsing Module
GAN	Generative Adversarial Network
MAE	Mean Absolute Error
MLP	Multilayer Perceptron
MSE	Mean Squared Error
RAF-DB	Real-world Affective Faces Database
ResNet	Residual Network
ViT	Vision Transformer
VMamba	Visual Mamba
YOLO	You Only Look Once

References

1. Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Gao, S.; Sun, Y.; Ge, W.; Zhang, W.; et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf. Fusion* **2022**, *83–84*, 19–52. [CrossRef]
2. Cheah, T.F.; Lee, C.P.; Lim, K.M.; Lim, J.Y. Facial expression classification with deep learning: A comparative study. In Proceedings of the 2023 IEEE 11th Conference on Systems, Process & Control (ICSPC), Malacca, Malaysia, 16 December 2023; pp. 56–59. [CrossRef]
3. Wang, K.; Yu, W.; Yamauchi, T. MVT-CEAM: A lightweight MobileViT with channel expansion and attention mechanism for facial expression recognition. *Signal Image Video Process.* **2024**, *18*, 6853–6865. [CrossRef]
4. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6896–6905. [CrossRef]
5. Cheheb, I. An investigation into the impact of occlusion on facial emotion recognition in the wild. In Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments, Crete, Greece, 26–28 June 2024; pp. 365–368. [CrossRef]
6. Ryumina, E.; Dresvyanskiy, D.; Karpov, A. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing* **2022**, *514*, 435–450. [CrossRef]
7. Wang, K.; Peng, X.; Yang, J.; Meng, D.; Qiao, Y. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Trans. Image Process.* **2020**, *29*, 4057–4069. [CrossRef]
8. Hou, H.; Sun, X. Joint learning for mask-aware facial expression recognition based on exposed feature analysis and occlusion feature enhancement. *Appl. Sci.* **2025**, *15*, 10433. [CrossRef]
9. Zhang, L.; Verma, B.; Tjondronegoro, D.; Chandran, V. Facial expression analysis under partial occlusion: A survey. *ACM Comput. Surv.* **2018**, *51*, 1–49. [CrossRef]
10. Li, S.; Deng, W.; Du, J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2584–2593. [CrossRef]
11. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 279–283. [CrossRef]
12. Li, Y.; Zeng, J.; Shan, S.; Chen, X. Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Trans. Image Process.* **2019**, *28*, 2439–2450. [CrossRef]
13. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.-C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the 15th European Conference on Computer Vision–ECCV 2018, Munich, Germany, 8–14 September 2018; pp. 89–105. [CrossRef]
14. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; Ebrahimi, M. EdgeConnect: Generative image inpainting with adversarial edge learning. *arXiv* **2019**, arXiv:1901.00212. [CrossRef]

15. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929. [CrossRef]
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010. [CrossRef]
17. Tay, Y.; Dehghani, M.; Bahri, D.; Metzler, D. Efficient Transformers: A survey. *ACM Comput. Surv.* **2023**, *55*, 1–28. [CrossRef]
18. Miao, Q.; Jia, L.; Xie, K.; Fu, K.; Yang, Z. A comprehensive survey and taxonomy of Mamba: Applications, challenges, and future directions. *Inf. Fusion* **2026**, *130*, 104094. [CrossRef]
19. Jiao, J.; Liu, Y.; Liu, Y.; Tian, Y.; Wang, Y.; Xie, L.; Ye, Q.; Yu, H.; Zhao, Y. VMamba: Visual state space model. In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024; pp. 103031–103063. [CrossRef]
20. Dixit, C.; Satapathy, S.M. Deep CNN with late fusion for real time multimodal emotion recognition. *Expert Syst. Appl.* **2024**, *240*, 122579. [CrossRef]
21. Liu, C.; Wang, Y.; Yang, J. A transformer-encoder-based multimodal multi-attention fusion network for sentiment analysis. *Appl. Intell.* **2024**, *54*, 8415–8441. [CrossRef]
22. Patel, R.; Ajoodha, R. Enhancing human expression classification using local binary patterns and support vector machines with RBF kernel in convolutional neural networks. *SSRN Electron. J.* **2023**. [CrossRef]
23. Lu, C.; Jiang, Y.; Fu, K.; Zhao, Q.; Yang, H. LSTPNet: Long short-term perception network for dynamic facial expression recognition in the wild. *Image Vision Comput.* **2024**, *142*, 104915. [CrossRef]
24. Hazmoune, S.; Bougamouza, F. Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108339. [CrossRef]
25. Ma, H.; Lei, S.; Li, H.; Celik, T. FER-VMamba: A robust facial expression recognition framework with global compact attention and hierarchical feature interaction. *Inf. Fusion* **2025**, *124*, 103371. [CrossRef]
26. Li, L.; Sun, Q.; Zhao, L.; Sun, H.; Zhao, F.; Gu, B. Face Mamba: A facial emotion analysis network based on VMamba*. In Proceedings of the 2024 7th International Conference on Machine Learning and Natural Language Processing (MLNLP), Chengdu, China, 18–20 October 2024; pp. 1–5. [CrossRef]
27. Wang, Z.; Wu, W.; Zeng, C.; Shen, J. PhysioFormer: Integrating multimodal physiological signals and symbolic regression for explainable affective state prediction. *PLoS ONE* **2025**, *20*, e0335221. [CrossRef]
28. Wang, H.; Ren, C.; Yu, Z. Multimodal sentiment analysis based on multiple attention. *Eng. Appl. Artif. Intell.* **2025**, *140*, 109731. [CrossRef]
29. Filho, G.P.R.; Meneguette, R.I.; Mendonça, F.L.L.d.; Enamoto, L.; Pessin, G.; Gonçalves, V.P. Toward an emotion efficient architecture based on the sound spectrum from the voice of Portuguese speakers. *Neural Comput. Appl.* **2024**, *36*, 19939–19950. [CrossRef]
30. Jekauc, D.; Burkart, D.; Fritsch, J.; Hesenius, M.; Meyer, O.; Sarfraz, S.; Stiefelhagen, R. Recognizing affective states from the expressive behavior of tennis players using convolutional neural networks. *Knowl.-Based Syst.* **2024**, *295*, 111856. [CrossRef]
31. Zhang, F.; Chai, L. A review of research on micro-expression recognition algorithms based on deep learning. *Neural Comput. Appl.* **2024**, *36*, 17787–17828. [CrossRef]
32. Pei, E.; Hu, Z.; He, L.; Ning, H.; Berenguer, A.D. An ensemble learning-enhanced multitask learning method for continuous affect recognition from facial images. *Expert Syst. Appl.* **2024**, *236*, 121290. [CrossRef]
33. Hu, Y.; Wang, J.; Wang, X.; Yu, J.; Zhang, J. Efficient virtual-to-real dataset synthesis for amodal instance segmentation of occlusion-aware rockfill material gradation detection. *Expert Syst. Appl.* **2024**, *238*, 122046. [CrossRef]
34. Wang, S. Domain-adaptive faster R-CNN for non-PPE identification on construction sites from body-worn and general images. *Sci. Rep.* **2026**, *16*, 4793. [CrossRef]
35. Jocher, G.; Qiu, J.; Chaurasia, A. Ultralytics YOLO. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 6 March 2026).
36. Chumachenko, K.; Iosifidis, A.; Gabbouj, M. MMA-DFER: MultiModal Adaptation of Unimodal Models for Dynamic Facial Expression Recognition In-the-wild. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–18 June 2024; pp. 4673–4682.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.