scientific reports

Check for updates

OPEN

Spectral-spatial feature fusion for real-time facial expression recognition

Jinjing Ma[1]✉, Yongcheng Lin[1], Lanmei Qian[1], Hongfeng You[2] & Tingyu Gao[1]✉

Facial expression recognition (FER), as a critical task in computer vision and affective computing, has gained considerable attention in recent years. However, current methods often suffer from high computational costs and limited capability in extracting key discriminative features. To address these issues, this paper proposes SPAYOLO (Spectral-aware Perception and Aggregation YOLOv8), a novel FER network based on the YOLOv8 architecture. We introduce a new Spectral-aware Perception and Aggregation Module (SPAM), designed to enhance expression recognition performance by systematically modeling spatial and frequency features. SPAM comprises three components: a Hierarchical Receptive Modeling (HRM) path that uses multi-scale convolutional branches to capture fine-grained and mid-level spatial variations; a Frequency Enhancement Path (FEP) that leverages Fast Fourier Transform (FFT) to extract high-frequency texture and micro-expression features; and a Gated Attention Mechanism (GAM) that adaptively fuses spatial and frequency features to mitigate feature distribution inconsistency and improve discriminative stability. Experimental results show that the proposed model achieves an accuracy of 70.74% on the FER2013 dataset and 67.88% on the AffectNet dataset, while maintaining high computational efficiency. These results highlight its suitability for real-time facial expression recognition tasks.Our findings validate the effectiveness of hierarchical feature fusion and frequency-domain enhancement in FER tasks, offering valuable insights for future research in computer vision.The custom code for this study is available at GitHub repository: https://github.com/YociLam/Spectral-Spatial-Feature-Fusion-for-Real-Time-Facial-Expression-Recognition.

Keywords Facial expression recognition, YOLOv8, Hierarchical Multi-Scale feature fusion, Spectral-Domain representation, Gated Cross-Domain attention, Frequency-Aware model

Recent advances in computer vision, driven by deep learning, have significantly improved facial analysis tasks. Facial expression recognition (FER), a key component in areas such as identity verification, affective computing, and human-computer interaction, involves detecting facial expressions, extracting relevant features, and performing classification.

Early approaches to FER relied on handcrafted features and landmark-based methods, such as Principal Component Analysis (PCA)[1], Linear Discriminant Analysis (LDA)[2], and Local Binary Patterns (LBP)[3]. While these techniques achieved moderate success under controlled conditions, they struggle with illumination changes, occlusions, and pose variations in real-world scenarios. The emergence of convolutional neural networks (CNNs)[4], such as FaceNet[5] and ArcFace[6], has marked a paradigm shift in facial recognition, enabling the learning of high-dimensional representations that improve intra-class compactness and inter-class separability. Nevertheless, these models remain suboptimal for FER, which demands finer discrimination of subtle muscular movements and faces the additional challenge of ambiguous annotations, as noted by Jin et al. who modeled visual sentiment classification with low-rank subspace learning and label relaxation[7].

Facial expression recognition differs from standard identity or object classification in that it requires precise modeling of localized facial muscle changes. To address this, recent studies have incorporated attention mechanisms[8] into CNNs to enhance feature focus on critical facial regions. For instance, SENet[9] introduced channel-wise attention, CBAM[10] integrated spatial and channel attention, and ECA[11] proposed lightweight attention with local 1D convolutions. These approaches inspired the design of the multi-scale attention modules in this work.

[1]Nantong Institute of Technology, Nantong 226001, Jiangsu, China. [2]Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223005, Jiangsu, China. ✉email: 843731114@qq.com; 846206605@qq.com

YOLO[12] (You Only Look Once), as an efficient single-stage object detection framework, achieves a favorable balance between computational efficiency and detection accuracy. Since the introduction of YOLOv1, subsequent versions have incorporated architectural advancements such as Cross Stage Partial Networks (CSPNet)[13], Path Aggregation Networks (PANet)[14], and dynamic label assignment strategies, resulting in consistent performance improvements across a wide range of tasks. Compared with two-stage approaches like Faster R-CNN[15], YOLO exhibits significant advantages in terms of end-to-end inference speed and structural efficiency.In parallel, recent advances such as DFD-NAS[16] further highlight how neural architecture search can automatically discover lightweight yet accurate backbones for multimedia forensics, underscoring the importance of efficiency–accuracy trade-offs also pursued in our FER framework.

Although YOLO was originally designed for object detection, recent studies have repurposed its architecture for classification tasks by leveraging its strong feature extraction capabilities. Applications have extended to domains such as medical imaging and traffic sign recognition. However, facial expression recognition poses distinct challenges due to its sensitivity to fine-grained facial variations, such as subtle movements around the eyes, eyebrows, and mouth. YOLO's conventional architecture struggles to effectively capture these local variations. Moreover, facial expressions often involve multi-scale and high-frequency details, which are not adequately modeled by traditional CNNs operating purely in the spatial domain.

To address these challenges, we propose a Spectral-aware Perception and Aggregation Module (SPAM), which is integrated into a classification-adapted version of YOLOv8. SPAM enhances the modeling of facial expression features by incorporating multi-scale convolutional operations to improve feature extraction and employing Fast Fourier Transform (FFT)[17] to emphasize frequency-domain details. Additionally, a Gated Attention Mechanism (GAM)[18] is introduced to optimize feature interaction across spatial and frequency domains. This design improves expression recognition accuracy while maintaining computational efficiency.

The main contributions of this work are as follows:

1) Hierarchical Receptive Modeling (HRM): To address the spatial-scale heterogeneity of facial expression regions, we design a multi-branch convolutional structure with varying receptive fields. This enhances the model's spatial awareness of both fine-grained local variations and mid-scale dynamic regions.
2) Frequency Enhancement Path (FEP): We introduce Fast Fourier Transform (FFT) for frequency-domain modeling of spatial feature maps. Combined with a channel attention mechanism, this reweights spectral components to emphasize high-frequency details, improving the model's responsiveness and discriminative consistency to subtle facial expression changes.
3) Gated Attention Mechanism (GAM): A cross-domain attention mechanism is developed to dynamically fuse spatial and frequency features under global semantic guidance. This mitigates domain inconsistency and facilitates adaptive integration of heterogeneous feature representations.
4) Efficient Integration of Lightweight Attention into a Detection Framework: The proposed SPAM module is embedded into a classification-adapted YOLOv8 backbone. Without introducing significant computational overhead, this integration improves semantic awareness and structural adaptability in FER tasks, achieving a balance between recognition accuracy and deployment feasibility.

The remainder of this paper is organized as follows: Sect. 2 reviews related work; Sect. 3 presents the detailed architecture and methodology; Sect. 4 outlines the experimental setup and results; and Sect. 5 concludes with a summary and discussion of future directions.

Related work

Facial expression recognition is a critical task in affective computing and human-computer interaction. Recent methods have made significant progress by leveraging multi-scale modeling, attention mechanisms, and frequency-domain features. However, these approaches often face challenges such as structural complexity, limited adaptability to real-world conditions, and high computational costs, which restrict their deployment in resource-constrained scenarios. To address these limitations, the proposed SPA module integrates multi-scale spatial perception, frequency enhancement, and lightweight gated attention into a unified framework. By dynamically combining spatial and frequency-domain features, SPA not only improves recognition accuracy but also ensures practical efficiency, making it a robust solution for real-time FER applications.

Recent studies have explored frequency-domain modeling and cross-modality fusion techniques to enhance FER performance. For example, MoADNet[19] introduced a lightweight dual-stream network for RGB-D salient object detection, showcasing the efficiency of compact architectures for multimodal feature processing. Similarly, FCMNet[20] leveraged frequency-aware attention mechanisms to effectively capture complementary information between modalities. Furthermore, TriPINet[21]proposed dynamic cross-modality fusion strategies to integrate multistage features, improving localization performance in image forensics. Liang et al. further demonstrated that progressive cross-modality integration can boost fine-grained localization[22], a principle echoed in our gated spectral–spatial fusion. These approaches inspired the design of the Frequency Enhancement Path (FEP) and Gated Attention Mechanism (GAM) in this work, enabling dynamic spatial-frequency alignment to overcome limitations in prior methods.

Multi-scale feature modelingnn

To enhance the expressiveness of facial features, multi-scale architectures have become a prevalent modeling strategy. For example, POSTER + +[23] employs a dual-branch design to integrate image and landmark features, using a hierarchical pyramid structure to enhance perception at different spatial scales and reduce intra-class variability. However, it heavily depends on external landmark detectors and suffers from low modular coupling and a complex training pipeline. EMA-Net introduces cross-scale attention modules and compact aggregation

structures to improve context modeling while maintaining high efficiency. Nonetheless, its fusion mechanism mainly relies on shallow-level concatenation, limiting its ability to dynamically emphasize important regions at varying scales.

In contrast, the proposed Spectral-aware Perception and Aggregation (SPA) module adopts structurally distinct convolutional paths tailored for different spatial semantics. A dynamic gating mechanism is introduced during the fusion stage to enable weighted selection of cross-scale features. This design allows the model to capture both global context and local details while adaptively enhancing the discriminability of key facial regions.

Attention mechanisms in FER

Attention mechanisms have emerged as critical components in deep models for improving region selectivity and feature expressiveness in FER. For instance, FerNeXt[24] enhances the quality of expression features by incorporating channel attention, effectively suppressing interference from irrelevant regions. However, traditional attention methods—such as Squeeze-and-Excitation (SE)—often rely on static receptive fields or global pooling, which limits their responsiveness to local dynamic changes and can lead to redundancy or missing information in complex scenes.

To address long-range dependencies, Transformer-based architectures have recently been introduced in FER. FER-Former[25] combines multimodal information with hybrid self-attention to achieve state-of-the-art performance in expression classification. Similarly, Face-MLLM[26] employs multi-stage pretraining and semantic space alignment to enhance cross-task generalizability. Other studies have proposed Transformer-based aggregation modules such as EMA[27], which aim to balance model performance and parameter efficiency. Additionally, some approaches integrate Transformer modules with efficient detection frameworks such as YOLOv5[28,29], balancing model expressiveness with structural efficiency. Despite the global modeling capabilities of Transformers, they are often computationally intensive and structurally complex, limiting their deployment in resource-constrained environments. Meanwhile, CNN-based attention mechanisms continue to struggle with scale adaptiveness and fine-grained modeling.

To this end, we propose a Gated Attention Mechanism (GAM) that unifies spatial, frequency, and scale dimensions. With learnable gating functions, GAM selectively emphasizes discriminative features while combining the efficiency of CNNs with the flexibility of Transformers, yielding a lightweight solution for joint spatial–frequency modeling.

Frequency-domain modeling

Recently, frequency-domain information has been increasingly explored as a valuable supplement to spatial-domain modeling in vision tasks such as facial expression recognition and real-world image denoising. Zhou et al.[30] introduced Fourier transforms into pose estimation to enhance high-frequency detail representation, while Zhuang et al.[31] proposed a Frequency-Regulated Channel-Spatial Attention (FCSA) mechanism to improve image classification via frequency-based reweighting. Similarly, recent studies in image denoising[32,33] have demonstrated the utility of frequency-domain modeling in preserving subtle textures and suppressing irrelevant noise. However, many of these approaches apply global frequency operations and lack spatial or semantic alignment, which can limit their integration with attention mechanisms and reduce effectiveness in expression-specific modeling tasks.

Building on these insights, we embed the frequency enhancement process within the value vector generation path of our attention mechanism. Conventional CNN-based attention mechanisms (e.g., SE, CBAM) often struggle with scale adaptiveness and fine-grained feature modeling, as also evidenced in recent studies[34] on real-world image denoising. These limitations are particularly critical for FER tasks, where subtle, localized muscle movements require both high spatial precision and flexible feature emphasis.

Our Gated Attention Mechanism (GAM) adaptively controls the fusion of spatial and frequency features, enabling joint modeling within the attention framework. Unlike FCSA's global channel reweighting, SPAM aligns frequency cues with spatial semantics at multiple scales, selectively emphasizing expression-relevant details (e.g., wrinkles, micro-movements) while reducing inconsistencies from static global operations.

Model efficiency and deployability

Lightweight design and inference efficiency are vital for real-world FER. The YOLO series, with its single-stage architecture and high speed, has been widely used in recognition tasks. FER-YOLO-Mamba[36] extends YOLOv5 with a state-space attention mechanism, but such methods remain limited to shallow attention or minor structural tuning, lacking deep modeling of fine-grained expressions or cross-domain enhancement.

In this work, we integrate the SPA module into the YOLOv8 backbone in a lightweight manner, combining YOLOv8's efficient feature extraction with enhanced multi-scale and frequency-aware representation. This enables a better trade-off between accuracy and real-time performance.

Summary

Recent FER methods have made significant progress in multi-scale modeling, attention mechanisms, and frequency-domain features. However, they often suffer from structural complexity, limited adaptability, and challenges in real-world deployment. The proposed SPA module addresses these limitations by unifying multi-scale spatial perception, frequency enhancement, and lightweight gated attention into a cohesive design. This approach not only improves expression recognition accuracy but also ensures practical efficiency, making it suitable for real-time and resource-constrained applications.

Methodology

This section provides a detailed description of the proposed SPAYOLO network architecture and its core component—the Spectral-aware Perception and Aggregation Module (SPAM). The model is specifically designed to meet the demands of facial expression recognition, which requires precise modeling of fine-grained features. As illustrated in Fig. 1, SPAYOLO integrates spatial and frequency feature modeling through a targeted and efficient architectural design, enabling both deep perception and adaptive fusion of heterogeneous information.

Overall architecture of SPAYOLO

SPAYOLO builds upon the lightweight and efficient backbone of YOLOv8, enhancing its ability to model local detail variations that are crucial in FER tasks. In particular, SPAM is introduced at the feature extraction stage to strengthen the model's sensitivity to high-frequency micro-patterns and to enable robust multi-scale spatial feature learning.

The overall network comprises three functionally complementary paths:

1) Hierarchical Receptive Modeling (HRM): This branch captures spatial structures at multiple receptive field scales, enabling the modeling of diverse facial regions with varying granularity.
2) Frequency Enhancement Path (FEP): This branch focuses on frequency-domain representation, enhancing the model's ability to capture high-frequency changes such as subtle muscle movements and fine textures.
3) Gated Attention Mechanism (GAM): This module adaptively fuses spatial and frequency features via a learnable gating mechanism, mitigating domain inconsistency and enabling dynamic weighting across heterogeneous representations.

This cooperative architecture—combining hierarchical spatial perception, frequency enhancement, and gated cross-domain fusion—maximizes the advantages of YOLOv8 in efficient feature extraction, while addressing its limitations in capturing subtle facial expressions. By incorporating targeted feature modeling across both spatial and frequency dimensions, SPAYOLO achieves superior recognition accuracy and generalization capability under complex emotional conditions.

Spectral-aware perception and aggregation module

The Spectral-aware Perception and Aggregation Module (SPAM) is the core innovative component of the SPAYOLO architecture. It is designed to achieve deep integration of spatial-scale heterogeneity modeling and frequency component enhancement. As illustrated in Fig. 2, in contrast to conventional convolutional networks that rely solely on spatial-domain operations for local feature extraction, SPAM employs a multi-branch and multi-domain collaborative framework to systematically improve the richness and discriminative capacity of the learned features.

HRM

The Hierarchical Receptive Modeling (HRM) path is designed to capture the spatial-scale heterogeneity of expression-activated facial regions. Facial expressions often involve the coordinated movement of multiple anatomical areas, each exhibiting distinct activation patterns at varying spatial scales. For instance, eyebrow raising typically occurs in small, localized regions, whereas downward movement of the mouth corners or global facial contractions affects broader areas.

Conventional convolutional networks, constrained by fixed receptive field sizes, struggle to simultaneously model both localized micro-changes and broad structural variations within a unified feature space. This limitation hinders multi-scale feature integration and impairs the model's ability to interpret complex emotional

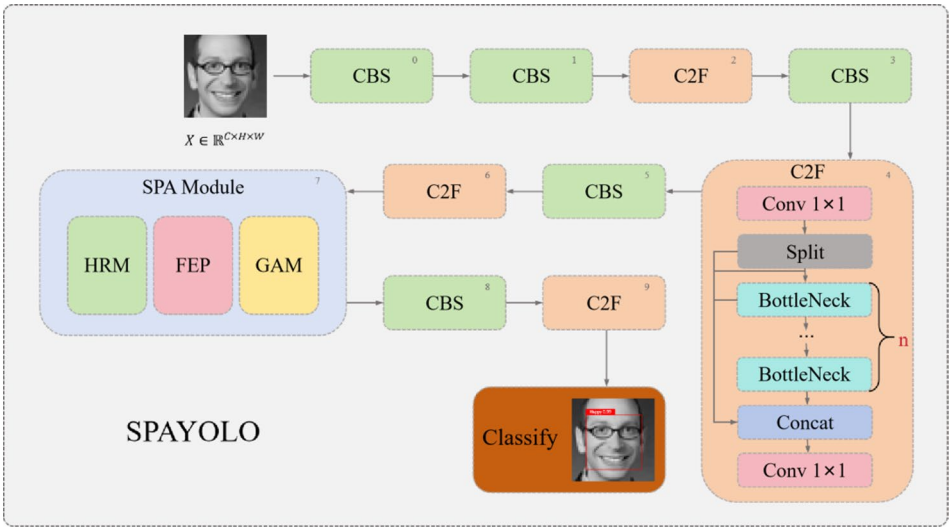

Fig. 1. The overall architecture of the SPAYOLO network.

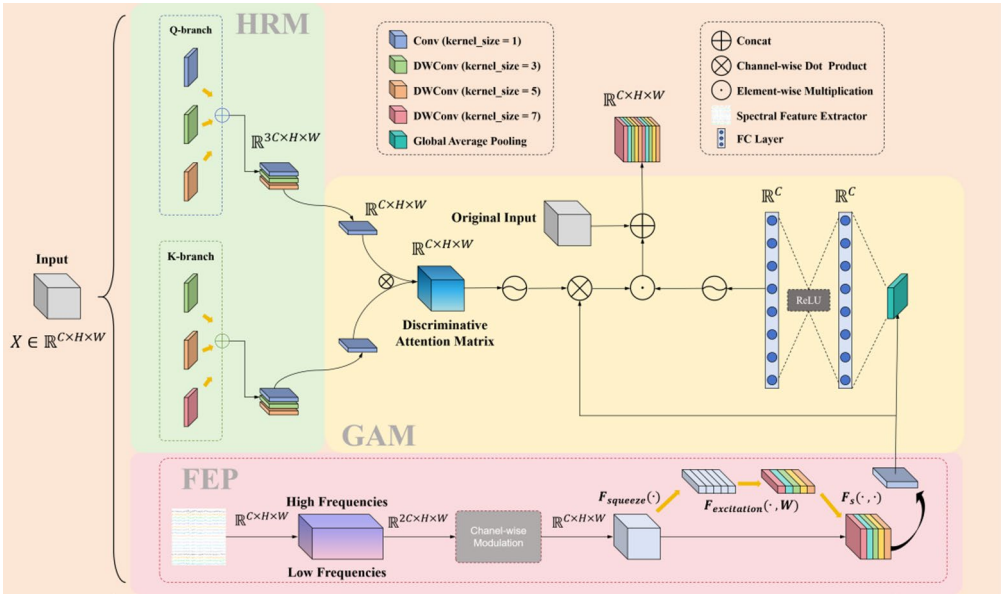

Fig. 2. The SPA module consists of three components: HRM (green) for multi-scale spatial feature extraction, FEP (pink) for frequency enhancement using FFT, and GAM (yellow) for adaptive spatial-frequency fusion via gated attention. Key operations are illustrated in the legend.

expressions. To overcome this, HRM introduces a parallel multi-branch architecture with varied receptive fields to explicitly enable multi-scale spatial modeling.

Specifically, HRM comprises four convolutional branches employing kernel sizes of 1×1, 3×3, 5×5, and 7×7, respectively, to extract features across different spatial granularities—from fine to mid-scale responses. Each branch is tailored to its target scale and combines standard and depthwise separable convolutions to balance modeling capacity with computational efficiency. After extracting features at multiple scales, outputs from all branches are passed through a shared 1×1 convolution for channel compression, resulting in a compact and expressive spatial representation.

To further support the dynamic modeling needs of the subsequent attention mechanism, HRM incorporates a differentiated pathway design during channel compression:

1) **Query vectors (Q)** are primarily derived from the 1×1, 3×3, and 5×5 branches to capture localized details and rapidly changing features, enhancing sensitivity to fine-grained facial expressions.
2) **Key vectors (K)**, while incorporating 3×3 features for fine-scale context, place greater emphasis on 5×5 and 7×7 branches to support global and mid-scale structural modeling, thereby improving stability in perceiving broader facial layouts and emotional transitions.

To formally represent the multi-branch structure of HRM, we define the scale-aware spatial feature aggregation process as a weighted combination of features extracted from different receptive fields. Each branch output is modulated by a learnable scalar and subsequently fused via a shared 1×1 convolution. The resulting unified feature map is then used to generate the query and key vectors for the subsequent attention mechanism, as formulated below:

$$\boldsymbol{F}_s = \mathrm{Conv}_{1\times 1}\left(\sum\nolimits_{i=1}^{n} \alpha_i \cdot f_i(\boldsymbol{X})\right) \tag{1}$$

$$\boldsymbol{Q} = \Psi_Q(\boldsymbol{F}_s),\ \boldsymbol{K} = \Psi_K(\boldsymbol{F}_s) \tag{2}$$

where: $\mathrm{X} \in \mathbb{R}^{\mathrm{C}\times \mathrm{H}\times \mathrm{W}}$ denotes the input feature map to the HRM module,

$\mathrm{f_i}$denotes the output of the i-th convolutional branch with a specific,

receptive field, where $\mathrm{i} \in \{1,2,3,4\}$corresponds to kernel sizes 1×1, 3×3, 5×5 and 7×7, respectively,

$\alpha_{\mathrm{i}} \in \mathbb{R}$ is a learnable scalar weight assigned to the i-th branch, used to adaptively balance the contribution of each scale,

$\Psi_{\mathrm{Q}}(\cdot)$and $\Psi_{\mathrm{K}}(\cdot)$denote learnable mappings used to project features into the query/key spaces.

This structured division of labor enables complementary modeling of local sensitivity and global consistency, empowering the attention mechanism with more discriminative capability for spatial alignment and feature selection.

Moreover, compared to dilated convolutions, HRM's continuous scale convolution avoids the gridding effect inherent in sparse sampling, thus preserving the integrity and coherence of local structures. Unlike pyramid pooling modules, HRM maintains an all-convolutional design, eliminating semantic fragmentation introduced by pooling and better retaining contextual coherence in critical facial regions. Additionally, the widespread

use of depthwise separable convolutions across all branches ensures that the module maintains low parameter overhead and computational cost, significantly enhancing its suitability for lightweight deployment, especially in resource-constrained environments such as mobile or edge devices.

In summary, the HRM module enhances the model's capability to decode cross-scale expression variations by leveraging diverse receptive fields and carefully organized branch structures. It provides a semantically rich yet compact feature foundation for subsequent frequency enhancement and.

dynamic fusion stages, achieving a dual optimization of structural efficiency and expressive performance in FER tasks.

FEP

Facial expression recognition often requires the identification of subtle local variations, such as fine wrinkles, micro-expressions, and muscle tremors, which tend to be encoded in the high-frequency components of facial images. However, traditional convolutional neural networks (CNNs), operating solely in the spatial domain, have inherent limitations in capturing these high-frequency cues due to their limited receptive fields and tendency toward spatial smoothing. This restricts their ability to distinguish between visually similar emotion categories, such as fear and disgust.

To address these challenges, the Frequency Enhancement Path (FEP) introduces a frequency-domain modeling strategy based on Fourier analysis. Specifically, it leverages the Fast Fourier Transform (FFT) to extract and emphasize high-frequency components that correspond to expression-critical texture cues. As shown in Fig. 3, incorporating frequency-aware attention allows the model to focus more precisely on semantically discriminative facial regions, such as the eyes and mouth.

The Fourier Transform decomposes complex spatial structures into orthogonal components across different frequency levels, thereby improving both feature separability and sparsity. In the context of FER, high-frequency components typically reflect fine-grained motion patterns and contour variations, while low-frequency components describe broader facial structure or illumination. Mapping features to the frequency domain therefore enhances the model's sensitivity to localized expression dynamics while suppressing irrelevant background noise.

Concretely, let the input feature map be denoted as $\mathbf{X} \in \mathbb{R}^{\mathrm{C} \times \mathrm{H} \times \mathrm{W}}$. A two-dimensional Fast Fourier Transform (FFT) is applied to project the spatial feature into the frequency domain:

$$\boldsymbol{F}_f = \mathcal{F}(\boldsymbol{X}) \tag{3}$$

where:

$\boldsymbol{F}_f$ represents the frequency-domain representation,

$\mathcal{F}(\bullet)$denotes the Fourier Transform operation, which is applied to project the spatial-domain features into the frequency domain.

To emphasize frequency components that are most relevant to facial expression recognition and suppress irrelevant or noisy responses, we introduce a channel-wise modulation mechanism within the frequency enhancement path. Specifically, we apply a learnable frequency reweighting strategy that incorporates channel-specific scaling and bias, followed by a lightweight attention module to generate refined frequency representations. The attention-enhanced feature is computed as:

$$\widetilde{\boldsymbol{F}}_f = \varphi_{\,\mathrm{att}}\left(\boldsymbol{F}_f \odot \boldsymbol{\gamma} + \boldsymbol{\beta}\right) \tag{4}$$

where:

$\widetilde{\boldsymbol{F}}_{\mathrm{f}}$ denotes the frequency feature map after channel-wise modulation and attention weighting,

$\varphi_{\,\mathrm{att}}$denotes the channel attention function, which selectively emphasizes informative channels in the frequency domain,

$\odot$ represents element-wise multiplication,

$\boldsymbol{\gamma} \in \mathbb{R}^{\mathrm{C}}$represents a learnable channel-wise scaling vector used to modulate frequency responses,

$\boldsymbol{\beta} \in \mathbb{R}^{\mathrm{C}}$represents a learnable channel-wise bias vector added to the modulated features.

To further compress the channel dimension and integrate semantic information within the frequency space, a 1×1 convolution is applied:

$$\boldsymbol{F}_c = \mathrm{Conv}_{1\times 1}\left(\widetilde{\boldsymbol{F}}_f\right) \tag{5}$$

where $\boldsymbol{F}_c$ denotes the compressed frequency-domain feature map.

To emphasize the frequency components most relevant to facial expression recognition and suppress irrelevant or noisy frequencies, FEP incorporates a channel attention mechanism for explicit weighting. The attention weights are computed as follows:

$$\varphi_{\,\mathrm{att}}(\boldsymbol{z}) = \sigma\left(\boldsymbol{W}_2 \cdot \delta\left(\boldsymbol{W}_1 \cdot \mathrm{GAP}(\boldsymbol{z})\right)\right) \tag{6}$$

where:

$\mathbf{z}$ denotes the frequency-domain feature map obtained from the FEP branch prior to attention computation.

$\mathrm{GAP}(\cdot)$ denotes global average pooling,

$\mathbf{W}_1$and $\mathbf{W}_2$ are learnable weights of two fully connected layers,

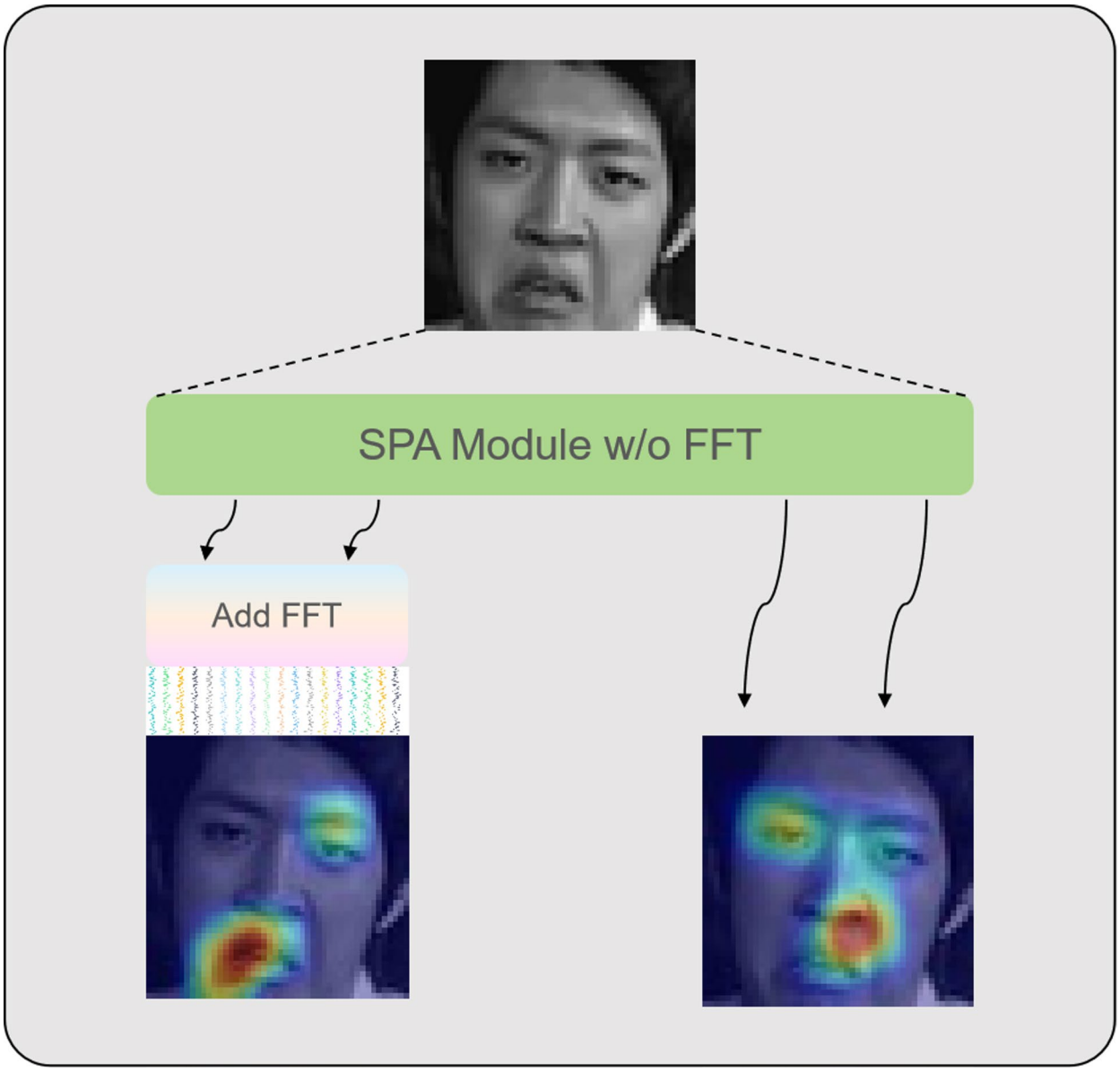

Fig. 3. Effect of integrating spectral-domain information via FFT in the SPA module. The image on the right shows the attention distribution from the original SPA module without frequency modeling. After incorporating the FFT-based spectral extractor (left), the model focuses more precisely on semantically discriminative facial regions such as the mouth and eyes. This highlights the importance of frequency-aware attention in enhancing subtle expression cues.

$\delta(\cdot)$ is a non-linear activation function,
$\sigma(\cdot)$ is a sigmoid activation function.
The final enhanced frequency feature is obtained by channel-wise multiplication:

$$F_{\text{enhanced}} = F_c \odot \varphi_{\text{att}}(z) \tag{7}$$

where $\odot$ denotes element-wise multiplication across channels.

The enhanced frequency features F_{enhanced} are subsequently used as value vectors (V) in the downstream attention fusion mechanism, where they are jointly modeled with the spatial features from the HRM path to generate a complete spatial-frequency representation.

Through this design, FEP effectively addresses the spatial pathway's deficiency in modeling high-frequency details. It enhances feature sparsity and saliency, selectively amplifies expression-relevant frequency components, and suppresses background and redundant frequencies. This leads to improved discriminability and robustness in feature representation.

In summary, FEP provides SPAYOLO with an efficient, frequency-aware modeling path grounded in signal processing theory. It offers a compelling trade-off between expressive power and computational cost, demonstrating strong performance and application potential in complex facial expression recognition scenarios.

GAM

The Gated Attention Mechanism (GAM) serves as the core component for spatial–frequency feature fusion in the SPAYOLO architecture. Its primary goal is to enable efficient interaction and adaptive integration between heterogeneous feature representations originating from distinct perceptual domains.

Specifically, the HRM branch encodes local and mid-level spatial semantics, whereas the FEP branch captures high-frequency textures and fine-grained temporal dynamics. Due to significant differences in scale, sparsity, and semantic focus, direct fusion strategies—such as simple concatenation or fixed weighting—can lead to semantic mismatch, misaligned attention responses, and diminished discriminative power caused by information imbalance.

To address these issues, GAM adopts a gated modulation strategy that facilitates cross-domain attention fusion in a dynamic and context-aware manner. By adaptively regulating the contribution of each feature stream, GAM enhances fusion accuracy and stability, while preserving spatial–frequency alignment throughout the network.

Let the query vector $\mathrm{Q} \in \mathbb{R}^{\mathrm{C}\times \mathrm{H}\times \mathrm{W}}$ be derived from the HRM output, and the value vector $\mathrm{V} \in \mathbb{R}^{\mathrm{C}\times \mathrm{H}\times \mathrm{W}}$ from the FEP output. GAM first computes a spatial attention map A by applying a sigmoid activation to the query and performing element-wise modulation with the value:

$$\boldsymbol{A} = \sigma\ (\boldsymbol{Q} \otimes\ \boldsymbol{V}) \tag{8}$$

where$\otimes$denotes element-wise multiplication across corresponding spatial locations. This formulation offers a lightweight alternative to standard attention by enabling dynamic gating of frequency-enhanced features while preserving spatial alignment.

To further enhance semantic coherence during fusion, GAM incorporates a global context modulation path. Specifically, a global average pooling operation is applied to the fused feature map, followed by two lightweight fully connected (MLP) layers to generate a soft gating signal:

$$\boldsymbol{g} = \sigma\ (\boldsymbol{W}_2\, \delta\ (\boldsymbol{W}_1 \cdot\ \mathrm{GAP}\,(\boldsymbol{V}))) \tag{9}$$

where:

W_1 and W_2 are learnable weight matrices, $\delta\,(\cdot)$ is a non-linear activation function.

This gate vector g captures global-level feature importance and acts as a dynamic controller to modulate local attention fusion.

The final fused feature output is computed by integrating attention weighting, gated modulation, and residual information as follows:

$$\boldsymbol{F}_{\mathrm{out}} = (\boldsymbol{A} \odot\ \boldsymbol{V}) \times\ \boldsymbol{g} + \boldsymbol{X} \tag{10}$$

where:

X denotes the residual feature from the input branch,

$\odot$represents element-wise multiplication between the attention map A and value feature V.

This fusion strategy captures key responses at the local spatial level while dynamically reweighting features globally, effectively mitigating common biases in local attention mechanisms such as overfitting to dominant regions.

From a design perspective, GAM enhances adaptive integration of multi-source features by mitigating mutual interference through gating, while global context guidance promotes semantic consistency and regional alignment. This ultimately strengthens the network's capacity for fine-grained discrimination and robust feature representation.

In summary, GAM not only unifies spatial and frequency-domain features structurally, but also introduces a lightweight, dynamically tunable, and generalizable fusion paradigm. It plays a vital role in enabling SPAYOLO to achieve strong performance in complex facial expression recognition tasks.

Module collaboration and overall integration strategy

To balance computational efficiency with representational power, SPAYOLO adopts a modular collaboration and hierarchical integration strategy that fully exploits the complementary strengths of its three key pathways—HRM, FEP, and GAM. This design systematically enhances the network's depth of feature perception and its discriminative capacity across both spatial and frequency domains.

Specifically, the SPAM module serves as a unified feature processing unit that integrates three sub-paths: multi-receptive field spatial modeling (HRM), frequency component enhancement (FEP), and gated dynamic fusion (GAM). During the feature extraction stage:

1) The HRM path first encodes spatial features at varying granularities, capturing local to mid-scale structural responses;
2) The FEP path concurrently models frequency-domain representations using Fourier Transform, reinforcing high-frequency detail and compensating for the spatial path's limitations in modeling rapid variations;
3) The GAM module then acts as a fusion hub, dynamically aligning and weighting features from both domains via joint spatial-frequency attention mechanisms.

This modular collaboration process can be formally represented as:

$$\boldsymbol{F}_{\mathrm{HRM}} = \mathrm{HRM}\left(X\right),\ \boldsymbol{F}_{\mathrm{FEP}} = \mathrm{FEP}\left(X\right) \tag{11}$$

$$\boldsymbol{F}_{\mathrm{SPAM}} = \mathrm{GAM}\left(\boldsymbol{F}_{\mathrm{HRM}}, \boldsymbol{F}_{\mathrm{FEP}}\right) \tag{12}$$

where:

$\mathrm{F}_{\mathrm{HRM}}$ represents the spatial features extracted by the Hierarchical Receptive Modeling (HRM) module, representing multi-scale spatial responses,

$\mathrm{F}_{\mathrm{FEP}}$ represents the frequency-domain features extracted by the Frequency Enhancement Path (FEP), representing high-frequency responses critical for facial expression recognition.

$\mathrm{F}_{\mathrm{SPAM}}$ represents the fused features obtained by the Gated Attention Mechanism (GAM), integrating spatial and frequency-domain features adaptively.

Within the overall SPAYOLO architecture, the SPAM module is embedded in a lightweight fashion into the mid-to-high-level semantic stages of the YOLOv8 backbone, typically at feature layers C3 or C4. These integration points strike a balance between local detail richness and high-level semantic context. The resulting feature maps are then merged with the original backbone outputs via residual connections, ensuring stable information flow and consistent gradient propagation throughout training. Finally, the fused features are passed to the classification and detection heads, allowing the model to maintain both fine-grained expression resolution and overall inference efficiency.

Importantly, SPAM is explicitly designed to be computationally efficient, with lightweight implementation achieved through three core strategies:

1) The HRM path employs depthwise separable convolutions to reduce the parameter overhead associated with large receptive fields;
2) The FEP path retains only the real component of the Fourier spectrum and incorporates a compression projection to minimize redundancy;
3) The GAM module uses a shallow fully connected gating branch, avoiding excessive computation in the fusion process.

In conclusion, SPAYOLO's modular and structural integration forms a unified feature processing system that combines multi-scale spatial modeling, frequency-enhanced perception, and cross-domain dynamic fusion. This design achieves substantial improvements in recognition accuracy, sensitivity to fine details, and system robustness—all while maintaining fast inference speeds, making the model suitable for practical deployment in real-time facial expression recognition scenarios.

Code availability

The custom code used in this study, including the implementation of the SPAYOLO network and its components (Spectral-aware Perception and Aggregation Module), is available at the following GitHub repository.

The repository contains all the necessary files to reproduce the results presented in this paper, including pre-trained models, training scripts, and evaluation code.

Experiments and results

Datasets

This study utilizes two widely adopted benchmark datasets for facial expression recognition: FER2013 and AffectNet.

The FER2013 dataset contains a total of 35,887 labeled facial images, spanning seven emotion categories: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprised. The dataset is divided according to the standard protocol into a training set of 28,709 images, and validation and test sets each consisting of 7,178 images.

The AffectNet dataset was originally annotated with eight emotion classes. However, for consistency with FER2013, we select only the seven categories shared across both datasets, resulting in a subset of 41,553 images. To ensure balanced distribution and consistent evaluation, we randomly divide this subset into a training set with 37,553 samples, and validation and test sets each containing 4,000 images.

All experiments in this study are conducted based on the above-defined data splits.

Experimental setup

All experiments are conducted under consistent hardware and software environments. The implementation is based on PyTorch 2.5.1, and training is performed on a workstation equipped with an Intel i5-12400 F CPU and NVIDIA GeForce RTX 4060 Ti GPU.

For model optimization, we adopt the Adan optimizer (details of the optimizer comparison are provided later) for SPAYOLO, with a learning rate of 1e-3, batch size of 64, and training for 200 epochs. For all baseline models, we follow the optimization settings and hyperparameters recommended in their respective original papers, using their official implementations whenever available.

To ensure fairness, all models are trained and evaluated on the same dataset splits. The loss function used is cross-entropy loss, and accuracy serves as the primary evaluation metric throughout our experiments.

Model	Accuracy (%)	Final loss	Epoch to plateau	Remarks
Adan[37]	70.74	0.00934	~ 151	Best performance overall
SGD[38]	70.10	0.01113	~ 171	Stable, slower convergence
AdaBoB[39]	70.13	0.01044	~ 148	Competitive
Adam[40]	68.42	0.01208	~ 146	Fast convergence, less stable
Adamax	68.83	0.01167	~ 144	Similar to Adam
Lion[41]	68.95	0.01153	~ 131	Fastest but lower accuracy

Table 1. Optimizer benchmarking for SPAYOLO on FER2013.

Method	FER2013	AffectNet
SPAYOLO(ours)	70.74	67.88
YOLOv8(original)	67.15	62.15
ISONet[42]	70.32	67.94
ResEmoteNet[43]	69.02	67.90
Mini-ResEmoteNet[44]	68.63	66.05
EmoNeXt-Tiny-22k[45]	64.54	65.30
LHC-Net[46]	63.17	64.72
QCS[47]	68.33	66.03
POSTER++	65.27	65.95

Table 2. Classification accuracy (%) of various models on the FER2013 and AffectNet dataset.

Optimizer comparison

We compared the performance of six optimizers—Adan, SGD, AdaBoB, Adam, Adamax, and Lion—on SPAYOLO using the FER2013 dataset. All models were trained under the same conditions (batch size of 64, learning rate of 1e-3, and 200 epochs).

As shown in Table 1, Adan achieved the highest accuracy (70.74%) and the lowest final loss (0.00934), reaching convergence around epoch 151. SGD was slightly lower in accuracy (70.10%) but stable. AdaBoB demonstrated excellent potential, with 69.8% accuracy and strong convergence, showing it to be a competitive optimizer for FER tasks.

Adam and Adamax had moderate accuracy (68.42% and 68.83%, respectively) but took longer to converge and exhibited higher final loss. Lion converged the fastest (~ 102 epochs) but yielded the lowest accuracy (68.95%).

Experimental results

Model performance evaluation and comparison

We evaluate the proposed SPAYOLO framework on the FER2013 and AffectNet datasets, and conduct comparative experiments against several recent state-of-the-art methods under identical hardware and software conditions. Due to variations in experimental setups—such as input resolution, preprocessing strategies, and training schedules—the reproduced results of some baseline models may slightly differ from those reported in their original publications. Nevertheless, all models were retrained using unified FER-specific settings to ensure fair and consistent comparison.

As shown in Table 2, SPAYOLO achieves classification accuracies of 70.74% on FER2013 and 67.88% on AffectNet. While its performance on AffectNet is marginally lower than that of ISONet (67.94%) and ResEmoteNet (67.90%), it consistently outperforms them on FER2013, indicating superior generalization on low-resolution, challenging datasets. This suggests that SPAYOLO strikes a favorable trade-off between accuracy and efficiency, validating the robustness and effectiveness of the proposed architecture across diverse FER benchmarks.

As shown in Fig. 4 (left), we visualize the training and validation loss curves along with the Top-1 and Top-5 accuracy trends throughout the training process. The steady decline in training loss indicates that the model is progressively learning discriminative features for facial expression recognition. The consistent rise in both Top-1 and Top-5 accuracy suggests improved performance and convergence stability over time.

Furthermore, Fig. 4 (right) presents the confusion matrix of the trained model on the FER2013 test set. The model achieves the highest classification accuracy in the "Happy" and "Surprised" categories, reflecting strong capability in identifying high-salience emotions. However, moderate confusion is observed between "Fear" and "Disgust", indicating some challenges in distinguishing fine-grained expressions with subtle differences. Despite this, the overall classification performance surpasses that of the baseline and other mainstream models, demonstrating the robustness and generalizability of the proposed method.

Model efficiency and PQS-FP analysis

In real-world applications, model efficiency plays a crucial role in determining the feasibility of deployment. To evaluate the computational performance of our proposed method, we compare SPAYOLO with two representative

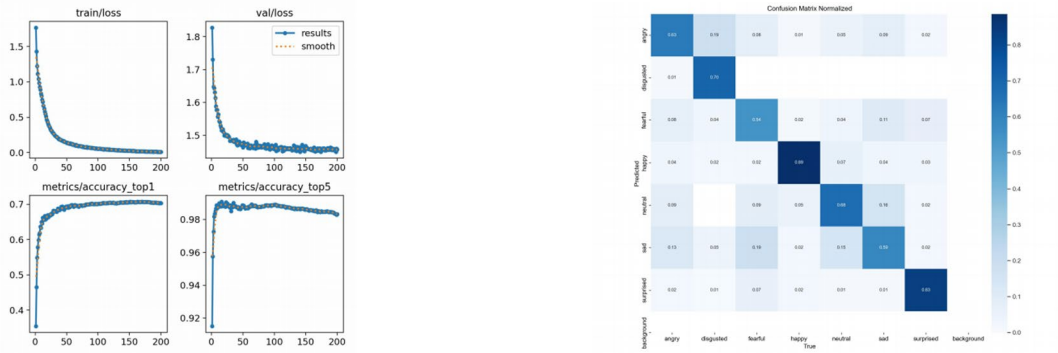

Fig. 4. Training and validation performance of SPAYOLO on the FER2013 dataset. Left: Loss and accuracy curves over training epochs. Right: Normalized confusion matrix illustrating classification performance across emotion categories.

Model	Computation (GFLOPs)	Parameters (M)	Training Time per Epoch (s)	Inference Time per Image (ms)
SPAYOLO(ours)	14.4	5.7	13	2.1
ResEmoteNet	4.35	80.24	43	5.8
EmoNeXt-Tiny-22k	4.57	30.56	150	8.4

Table 3. Computational efficiency comparison of SPAYOLO and baseline models in terms of FLOPs, parameter count, training time per epoch, and average inference time per image.

baselines: ResEmoteNet and EmoNeXt-Tiny-22k. The comparison includes floating-point operations (FLOPs), number of parameters, training time per epoch, and inference time per image, highlighting differences in computational resource utilization.

Table 3 summarizes the computational statistics of the three models. Notably, SPAYOLO demonstrates high efficiency while maintaining strong accuracy. Under the same batch size conditions, it achieves a training time of just 13 s per epoch, significantly faster than ResEmoteNet (43 s) and EmoNeXt-Tiny-22k (150 s).

In terms of FLOPs, SPAYOLO requires approximately 3.3× more computation than ResEmoteNet, yet it trains 3.3× faster, indicating superior computational efficiency. This is largely attributed to the highly optimized YOLOv8 backbone, which leverages depthwise separable convolutions and CSPNet to reduce redundancy in feature extraction, resulting in higher throughput despite increased theoretical complexity.

By contrast, ResEmoteNet—though lower in FLOPs—contains over 80 million parameters, primarily due to SE modules and fully connected layers. This leads to increased memory consumption and slower training. EmoNeXt-Tiny-22k, based on a ConvNeXt-Transformer hybrid architecture, exhibits slightly higher FLOPs than SPAYOLO but suffers from high training latency due to the computational complexity of self-attention operations.

Regarding inference speed, SPAYOLO achieves 2.1 ms per image, making it approximately 2.8× faster than ResEmoteNet and nearly 4× faster than EmoNeXt-Tiny-22k. The reduced parameter count and streamlined backbone make SPAYOLO well-suited for real-time facial expression recognition tasks in latency-sensitive environments.

To further interpret the comparative results, we adopt the PQS-FP (Parameter Quantity Shift–Fitting Performance) coordinate system[48] as a theoretical tool to analyze the trade-off between model complexity and performance. In this framework, the X-axis denotes the deviation in parameter quantity from an ideal reference, while the Y-axis reflects the corresponding deviation in model accuracy. Based on the direction and magnitude of these shifts, models are categorized into four quadrants, each representing a distinct fitting behavior.

We empirically set the ideal parameter count $O = 30$ million, which aligns with the scale of compact yet expressive architectures such as EmoNeXt-Tiny-22k. The performance baseline $P^* = 70.74\%$ corresponds to the highest FER2013 accuracy achieved by SPAYOLO. Together, these define the coordinate origin $(X = 0, Y = 0)$, representing a model that achieves an optimal balance between complexity and fitting capacity.

Under this framework (illustrated in Fig. 5): SPAYOLO falls in Quadrant IV (UAR), where reduced complexity correlates with improved performance, suggesting that the model effectively alleviates underfitting through its spectral-spatial design. ResEmoteNet appears in Quadrant I (OER), indicating that its excessive parameter count may lead to overfitting and redundant computation. EmoNeXt-Tiny-22k trends toward Quadrant III (UER), reflecting stagnant performance despite a reasonable parameter scale—implying ineffective utilization of its capacity.

This analysis reinforces that SPAYOLO not only achieves competitive accuracy and efficiency, but also occupies a theoretically favorable position in the PQS-FP space, confirming the practical value of its lightweight design in real-world FER applications.

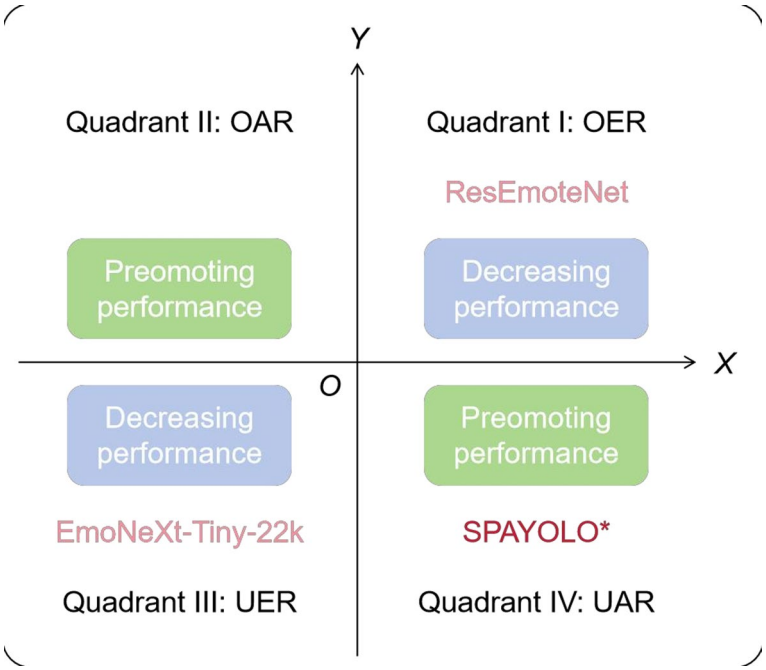

Fig. 5. The PQS-FP coordinate system: Y-axis: larger values indicate a higher existing parameter quantity of models; X-axis: larger values indicate an increase in the parameter quantity of models.

No.	Model variant	FER2013	AffectNet
1	YOLOv8 (Baseline)	67.45	62.55
2	SPAYOLO (ours)	70.74	68.28
3	SPAYOLO w/o FFT	69.95	65.42
4	SPAYOLO w/o multi-scale	67.66	54.13

Table 4. Ablation study: accuracy (%) of SPAYOLO variants on FER2013 and AffectNet.

Ablation study

To further validate the individual contributions of components within the SPA module, we conduct a series of ablation experiments based on the full SPAYOLO model. The experiments selectively modify key subcomponents of SPA to assess their impact on final performance. The following configurations are evaluated:

1) YOLOv8 Baseline: The original YOLOv8 model without any SPA integration.
2) SPAYOLO (Full Model): The proposed model with the complete SPA module.
3) SPAYOLO w/o FFT: The model with the frequency enhancement path removed—i.e., no FFT is applied, but the multi-scale convolution and Gated Attention Mechanism (GAM) remain.
4) SPAYOLO w/o Multi-scale Convolution: The model with multi-scale convolution removed—query and key vectors are generated from a unified transformation, while FFT and GAM are retained.

The results in Table 4 demonstrate that both the frequency enhancement (FFT) and multi-scale spatial modeling significantly contribute to the overall performance of SPAYOLO. Removing either component leads to a noticeable drop in classification accuracy. Notably, the exclusion of the FFT component reduces the model's ability to capture high-frequency details—such as subtle texture changes—thereby lowering the expression recognition accuracy. The removal of multi-scale convolution leads to even sharper performance degradation, especially on AffectNet, suggesting that modeling expressions at multiple receptive fields is essential for robust FER in real-world conditions.

Figure 6 illustrates the attention heatmaps under different SPA configurations, offering visual insights into the model's attention focus: (A) shows the original input facial image. (B) displays the attention map generated by the full SPA module. The model focuses accurately on critical facial regions such as the eyes, eyebrows, and mouth, indicating precise localization of expression-related cues. (C) shows the attention map when FFT is removed. Compared to (B), attention is more dispersed, and key regions—especially for classes like Happy and Sad—receive less emphasis on mouth corners and eye edges, reflecting weakened sensitivity to fine-grained details. (D) depicts the attention map without multi-scale convolution. The spatial coverage is significantly

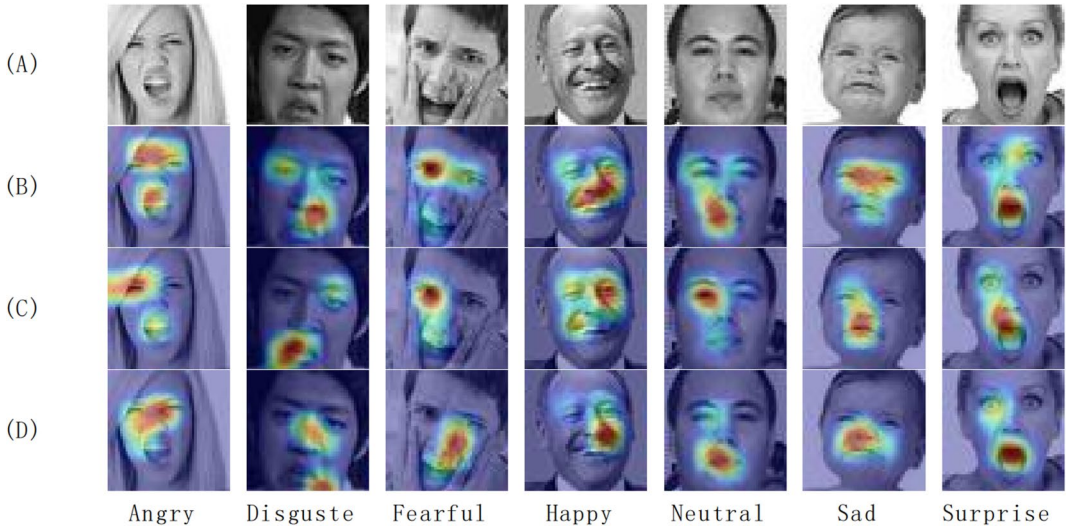

Fig. 6. Ablation study: attention heatmaps generated under different SPA configurations across various expression categories. (**A**) Original input images, (**B**) Attention maps produced by the full SPA module, (**C**) Attention maps after removing the frequency enhancement component (FFT), (**D**) Attention maps after removing the multi-scale convolution component.The full SPA module enables precise focus on critical facial regions such as the eyes, eyebrows, and mouth. In contrast, removing frequency or multi-scale components results in more dispersed or incomplete attention, indicating their importance for fine-grained expression modeling.

Emotion Label	1	2	3	4
Angry	60	63	60	59
Disgusted	64	70	68	67
Fearful	51	54	52	51
Happy	87	89	89	87
Neutral	63	68	66	65
Sad	58	59	61	61
Surprised	82	83	83	81

Table 5. Ablation study: per-class accuracy (%) of different SPAYOLO configurations on FER2013.

narrowed. For categories like Surprised and Fearful, the model fails to attend to the full extent of facial expression regions, underscoring the importance of multi-scale feature extraction.

Overall, the full SPA module ensures effective attention distribution and rich feature extraction across spatial and frequency domains. Removing either FFT or multi-scale convolution degrades the model's representational power and hinders expression classification performance.

Table 5 further presents the per-class accuracy across different methods. The full SPAYOLO model, with the complete SPA module, achieves consistent performance improvements across all emotion categories, with particularly strong results in the "Happy" and "Surprised" classes. This indicates that multi-scale feature fusion and frequency-domain enhancement effectively capture key facial regions, improving the discriminability of expression-related features.

When the FFT component is removed, the recognition accuracy for "Sad" and "Neutral" drops significantly. This highlights the critical role of frequency-domain modeling in representing subtle facial details. For these expression types, high-frequency components carry essential information about micro-level variations; their absence makes it difficult for the model to accurately capture emotional nuances.

On the other hand, removing multi-scale convolution results in the most pronounced accuracy drop for the "Fearful" category. This suggests that scale-aware feature integration plays a vital role in distinguishing between visually similar emotions. Without this mechanism, the model struggles to effectively combine local and global features, ultimately impairing its ability to recognize complex expressions.

Conclusion

This paper presents SPAYOLO, a novel facial expression recognition framework built upon the YOLOv8 architecture and enhanced by the proposed Spectral-aware Perception and Aggregation Module (SPAM). The SPAM module strengthens the model's ability to capture fine-grained emotional cues by integrating multi-scale spatial features and frequency-domain information. Extensive experiments on the FER2013 and

AffectNet datasets demonstrate that SPAYOLO achieves superior recognition accuracy while maintaining high computational efficiency. Ablation studies further verify that both the frequency-domain modeling and the multi-scale receptive mechanism are critical to boosting expression discriminability.

Thanks to its lightweight architecture and low-latency performance, SPAYOLO shows strong potential for deployment in real-time and resource-constrained environments such as mobile platforms or edge devices. Nevertheless, the relatively high GFLOPs compared to ultra-efficient networks highlight opportunities for further improvement. In this regard, future work will explore model compression strategies to reduce computational complexity while retaining accuracy. Recent advances in low-bit quantization, binarized networks, and adaptive learning-rate compression—such as BiVM[49], AdaLoRA[50], and QuantSR[51] —provide valuable insights for developing even more compact FER models.

In parallel, the recognition of visually similar categories such as "Fear" and "Disgust" may benefit from class-specific adaptive modeling or temporal cue integration. Incorporating multimodal signals like speech or physiological data could also improve robustness in real-world scenarios. Finally, hybridizing YOLOv8 with Transformer-based modules may further enhance global context modeling without significantly increasing computational cost. Overall, SPAYOLO strikes a compelling balance between accuracy, efficiency, and extensibility, making it a practical and scalable solution for modern facial expression recognition.`.

Data availability

The datasets used in this study, FER2013 and AffectNet, are publicly available. FER2013 can be accessed at https://paperswithcode.com/dataset/fer2013, and AffectNet can be accessed at https://paperswithcode.com/dataset/affectnet. The processed subsets used for training and evaluation in this study are available from the corresponding author upon reasonable request.

Received: 13 June 2025; Accepted: 5 November 2025
Published online: 17 December 2025

References

1. Holland, S. M. Principal components analysis. Department of Geology, University of Georgia, Athens, GA 30602 – 2501 (2008). https://doi.org/10.1038/s43586-022-00184-w
2. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
3. Ojala, T., Maenpaa, T. & Pietikainen, M. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 971–987. https://doi.org/10.1109/TPAMI.2002.1017623 (2002).
4. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444. https://doi.org/10.1038/nature14539 (2015).
5. Schroff, F., Kalenichenko, D., Philbin, J. & FaceNet A unified embedding for face recognition and clustering. In Proc. *IEEE Conf. Comput. Vis. Pattern Recognit.* 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682researchgate.net+1scirp.org+1
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S. & ArcFace Additive angular margin loss for deep face recognition. In Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 4690–4699 (2019). https://doi.org/10.1109/CVPR.2019.00482
7. Jin, X., Jing, P., Wu, J., Xu, J. & Su, Y. Visual sentiment classification via low-rank regularization and label relaxation. *IEEE Trans. Cogn. Dev. Syst.* https://doi.org/10.1109/TCDS.2021.3135948 (2021).
8. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762 (2017).
9. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745
10. Woo, S., Park, J., Lee, J. Y. & Kweon, I. S. CBAM: Convolutional block attention module. In *Proc. Eur. Conf. Comput. Vis.* 3–19 (2018). https://doi.org/10.1007/978-3-030-01234-2_1
11. Wang, Q. et al. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 11534–11542 (2020). https://doi.org/10.1109/CVPR42600.2020.01155
12. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91
13. Wang, C. Y. et al. CSPNet: A new backbone that can enhance learning capability of CNN. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops* 390–391 (2020). https://doi.org/10.48550/arXiv.1911.11929
14. Hanchao, L. et al. Pyramid attention network for semantic segmentation. arXiv preprint arXiv:1805.10180 (2018).
15. Girshick, R. & Fast, R-C-N-N. In *Proc. IEEE Int. Conf. Comput. Vis.* 1440–1448 (2015). https://doi.org/10.1109/ICCV.2015.169
16. Jin, X., Yu, W., Chen, D. W. & Shi, W. DFD-NAS: general deepfake detection via efficient neural architecture search. *Neurocomputing* https://doi.org/10.1016/j.neucom.2024.129129 (2024).
17. Liu, S., Wang, Q. & Liu, G. A versatile method of discrete Convolution and FFT (DC-FFT) for contact analyses. *Wear* **243**, 101–111. https://doi.org/10.1016/S0043-1648(00)00427-0 (2000).
18. Xue, L., Li, X. & Zhang, N. L. Not all attention is needed: gated attention network for sequence data. In *Proc. AAAI Conf. Artif. Intell.* 34, 6550–6557 (2020). https://doi.org/10.48550/arXiv.1912.00349
19. Jin, X., Yi, K., Xu, J. & MoADNet Mobile asymmetric dual-stream networks for real-time and lightweight RGB-D salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* https://doi.org/10.1109/TCSVT.2022.3180274 (2022).
20. Wang, Y. (ed Sartoretti, G.) FCMNet: full communication memory net for team-level Cooperation in multi-agent systems. *ArXiv Preprint arXiv:2201 11994* https://doi.org/10.48550/arXiv.2201.11994 (2022).
21. Liang, W. Y., Xu, J., Jin, X. & TriPINet Tripartite progressive integration network for image manipulation localization. *ArXiv Preprint arXiv*. https://doi.org/10.48550/arXiv.2212.12841 (2022). :2212.12841.
22. Liang, W. Y., Xu, J. & Jin, X. Image manipulation localization via dynamic cross-modality fusion and progressive integration. *Neurocomputing* https://doi.org/10.1016/j.neucom.2024.128607 (2024).
23. Mao, J. et al. Poster++: A simpler and stronger facial expression recognition network. *Pattern Recognit.* **146**, 110951. https://doi.org/10.1016/j.patcog.2024.110951 (2024).
24. El-Khashab, O., Hamdy, A., Mahmoud, A. & FerNeXt Facial expression recognition using ConvNeXt with channel attention. In *Proc. Int. Mobile Intell. Ubiquitous Comput. Conf.* 1–8 (2023). https://doi.org/10.1109/MIUCC58832.2023.10278345
25. Li, Y. et al. FER-former: multimodal transformer for facial expression recognition. *IEEE Trans. Multimedia*. https://doi.org/10.1109/TMM.2024.3521788 (2024).
26. Sun, H. et al. Face-MLLM: A large face perception model. *ArXiv Preprint arXiv*. https://doi.org/10.48550/arXiv.2410.20717 (2024). :2410.20717.

27. Ouyang, D. et al. Efficient multi-scale attention module with cross-spatial learning. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10096516
28. Hao, W. et al. Cattle body detection based on YOLOv5-EMA for precision livestock farming. *Animals* **13**, 3535. https://doi.org/10.3390/ani13223535 (2023).
29. Zhang, X. et al. Loader bucket working angle identification method based on YOLOv5s and EMA attention mechanism. *IEEE Access*. https://doi.org/10.24433/CO.3584989.v1 (2024).
30. Zhou, S., Duan, X. & Zhou, J. Human pose Estimation based on frequency domain and attention module. *Neurocomputing* **604**, 128318. https://doi.org/10.1016/j.neucom.2024.128318 (2024).
31. Zhuang, C. et al. Frequency regulated channel-spatial attention module for improved image classification. *Expert Syst. Appl.* **260**, 125463. https://doi.org/10.1016/j.eswa.2024.125463 (2025).
32. Ma, R. et al. Learning attention in the frequency domain for flexible real photograph denoising. *IEEE Trans. Image Process.* https://doi.org/10.1109/TIP.2024.3404253 (2024).
33. Ma, R., Li, S., Zhang, B. & Li, Z. Generative adaptive convolutions for real-world noisy image denoising. In Proc. AAAI Conf. Artif. Intell. 36(2), 1609–1617 (2022). https://doi.org/10.1609/aaai.v36i2.20088
34. Ma, R., Li, S., Zhang, B., Fang, L. & Li, Z. Flexible and generalized real photograph denoising exploiting dual meta attention. *IEEE Trans. Cybern*. https://doi.org/10.1109/TCYB.2022.3170472 (2022).
35. Ma, R., Li, S., Zhang, B. & Hu, H. Meta PID attention network for flexible and efficient real-world noisy image denoising. *IEEE Trans. Image Process.* https://doi.org/10.1109/TIP.2022.3150294 (2022).
36. Ma, H. et al. Fer-YOLO-Mamba: facial expression detection and classification based on selective state space. *ArXiv* Preprint ArXiv:2401.12345 (2024). https://doi.org/10.48550/arXiv.2405.01828
37. Xie, X., Zhou, P., Li, H., Lin, Z. & Yan, S. Adan: adaptive Nesterov momentum algorithm for faster optimizing deep models. *ArXiv Preprint arXiv*. 220806677. https://doi.org/10.48550/arXiv.2208.06677 (2022).
38. Robbins, H. & Monro, S. A stochastic approximation method. *Ann. Math. Stat.* **22** (3), 400–407. https://doi.org/10.1214/aoms/1177729586 (1951).
39. Xiang, Q., Wang, X., Lei, L. & Song, Y. Dynamic bound adaptive gradient methods with belief in observed gradients. *Pattern Recognit*. https://doi.org/10.1016/j.patcog.2025.111819 (2025).
40. P Kingma, D. & Ba, J. Adam: a method for stochastic optimization. *ArXiv Preprint arXiv:1412 6980*. https://doi.org/10.48550/arXiv.1412.6980 (2014).
41. Chen, X. et al. Symbolic discovery of optimization algorithms. *ArXiv Preprint arXiv:2302 06675*. https://doi.org/10.48550/arXiv.2302.06675 (2023).
42. Xiang, Q., Wang, X., Song, Y., Lei, L. & ISONet Reforming 1DCNN for aero-engine system inter-shaft bearing fault diagnosis via input Spatial over-parameterization. *Expert Syst. Appl.* https://doi.org/10.1016/j.eswa.2025.127248 (2025).
43. Roy, A. K. et al. ResEmoteNet: bridging accuracy and loss reduction in facial emotion recognition. *IEEE Signal. Process. Lett.* https://doi.org/10.1109/LSP.2024.3521321 (2024).
44. Murtada, A., Abdelrhman, O. & Attia, T. A. Mini-ResEmoteNet: leveraging knowledge distillation for human-centered design. arXiv preprint arXiv:2501.18538 (2025).
45. El Boudouri, Y., Bohi, A. & Emonext An adapted ConvNeXt for facial emotion recognition. In *Proc. IEEE Int. Workshop Multimedia Signal* Process. 1–6 (2023). https://doi.org/10.1109/MMSP59012.2023.10337732
46. Pecoraro, R., Basile, V. & Bono, V. Local multi-head channel self-attention for facial expression recognition. *Information* **13**, 419. https://doi.org/10.3390/info13090419 (2022).
47. Wang, C. et al. QCS: Feature refining from quadruplet cross similarity for facial expression recognition. *arXiv preprint* arXiv:2411.01988 (2024).
48. Xiang, Q. et al. Quadruplet depth-wise separable fusion Convolution neural network for ballistic target recognition with limited samples. *Expert Syst. Appl.* https://doi.org/10.1016/j.eswa.2023.121182 (2023).
49. Qin, H. et al. BiVM: accurate binarized neural network for efficient video matting. *ArXiv Preprint arXiv:2507 04456*. https://doi.org/10.48550/arXiv.2507.04456 (2025).
50. Qin, H. et al. Accurate LoRA-Finetuning quantization of LLMs via information retention. *ArXiv Preprint arXiv:2402 05445*. https://doi.org/10.48550/arXiv.2402.05445 (2024).
51. Qin, H. et al. QuantSR: accurate low-bit quantization for efficient image super-resolution. In *Proc. NeurIPS* (2023).

Acknowledgements

This work was funded by 2024 Annual Research Projects of Nantong Institute of Technology [2024XK(Z)30]; Natural Science Foundation of University in Jiangsu Province[23KJD520010]; Directive Projects of Nantong municipal science and technology plan[MS2023061]; 2024 Annual Research Projects of Nantong Institute of Technology[2024XK(Z)27].

Author contributions

Ma conceived and designed the model, led the overall research, conducted key experiments, analyzed the results, and drafted the manuscript. Lin assisted in methodology development, performed supplementary experiments, and contributed to data analysis. Qian was responsible for data collection and pre-processing. You contributed to algorithm implementation and optimization. Gao assisted in manuscript revision. All authors reviewed and approved the final version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.M. or T.G.