Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.075616

ARTICLE

SQSNet: Hybrid CNN-Transformer Fusion with Spatial Quad-Similarity for Robust Facial Expression Recognition

Mohammed A. Ahmed[1], Jian Dong[2,*], Ronghua Shi[2], Ammar Nassr[3], Hani Almaqtari[3] and Ala A. Alsanabani[3]

[1]School of Computer Science and Engineering, Central South University, Changsha, China
[2]School of Electronic Information, Central South University, Changsha, China
[3]School of Artificial Intelligence, Xidian University, Xi'an, China

*Corresponding Author: Jian Dong. Email: dongjian@csu.edu.cn

Received: 04 November 2025; Accepted: 04 February 2026; Published: 09 April 2026

ABSTRACT: Facial Expression Recognition (FER) is an essential endeavor in computer vision, applicable in human-computer interaction, emotion assessment, and mental health surveillance. Although Convolutional Neural Networks (CNNs) have proven effective in Facial Emotion Recognition, they encounter difficulties in capturing long-range connections and global context. To address these constraints, we propose Spatial Quad-Similarity Network (SQSNet), an innovative hybrid framework that integrates the local feature extraction capabilities of CNNs with the global contextual modeling efficacy of Swin Transformers via a cohesive fusion technique. SQSNet introduces the Spatial Quad-Similarity (SQS) module, a feature refinement approach that amplifies discriminative characteristics and mitigates redundancy. Unlike conventional metric learning approaches that operate on global feature representations, SQS computes fine-grained spatial-level similarity across multiple instances, enforcing $H \times W$ independent constraints that preserve spatial correspondence between expression-relevant facial regions. This spatial-level formulation is particularly effective for FER, where expressions manifest as localized muscle movements that are lost in global pooling operations. Moreover, SQSNet employs sophisticated regularization methods, including Mixup augmentation, label smoothing, and adaptive learning rate scheduling, to enhance generalization. Experimental findings on three benchmark datasets, RAF-DB, FERPlus, and AffectNet, indicate that SQSNet surpasses current FER methodologies, attaining state-of-the-art accuracies of 91.90%, 91.11%, and 67.15%, respectively. These findings underscore the efficacy of integrating CNNs, Swin Transformers, and spatial similarity-driven feature refining for facial emotion identification, facilitating the development of more dependable emotion recognition systems.

KEYWORDS: Facial expression recognition; convolutional neural networks; swin transformers; cross attention; adaptive learning

1 Introduction

Facial Expression Recognition (FER) is a fundamental task in computer vision with wide-ranging applications, including human-computer interaction, emotion analysis, mental health monitoring, and autonomous systems. The ability to accurately and robustly recognize facial expressions is crucial for understanding human emotions and behaviors. Despite significant advancements in deep learning, FER remains a challenging problem due to variations in lighting conditions, facial poses, occlusions, and the subtle differences between expressions [1].

Traditional approaches to FER have predominantly relied on Convolutional Neural Networks (CNNs), which excel at capturing local spatial features [2–4]. However, CNNs often struggle to model long-range dependencies and global contextual information, which are essential for distinguishing between similar expressions. Recent advancements in vision transformers, particularly the Swin Transformer, have demonstrated remarkable success in capturing global context through self-attention mechanisms. However, transformers typically require large amounts of data and computational resources, and their performance on tasks requiring fine-grained local feature extraction, such as FER, can be suboptimal when used in isolation [5]. Recent advances in FER have explored hybrid CNN-Transformer architectures to leverage complementary strengths. While these approaches show promise, they typically fuse features at the global level through concatenation or standard attention mechanisms, which may not fully exploit the fine-grained spatial relationships critical for distinguishing subtle expression differences. Moreover, existing metric learning approaches for FER operate on holistic feature representations after global pooling, potentially losing spatial correspondence between expression-relevant facial regions. To address these limitations, our work introduces spatial-level similarity learning that preserves and refines fine-grained spatial relationships before any global aggregation, providing richer supervision signals for learning discriminative features.

Based on this motivation, we introduce Spatial Quad-Similarity Network (SQSNet) a novel hybrid architecture that synergistically combines CNNs and Swin Transformers. SQSNet leverages the local pattern recognition capabilities of CNNs alongside the hierarchical, long-range dependency modeling of Swin Transformers. These two modalities are fused through a dedicated Spatial Quad-Similarity (SQS) mechanism designed to enhance multi-scale feature integration. This enables the model to better capture subtle facial variations while maintaining awareness of global facial structure. To further enhance robustness and generalization, SQSNet incorporates state-of-the-art regularization and optimization strategies, including Mixup-based data augmentation, label smoothing, and adaptive learning rate scheduling. Evaluations on three benchmark FER datasets demonstrate that SQSNet achieves superior performance and outperforms several state-of-the-art methods in terms of both accuracy and robustness. The contributions of this work are summarized as follows:

1. We propose a novel hybrid framework that integrates convolutional neural networks and Swin Transformers, leveraging their complementary strengths to enhance facial expression recognition. This design effectively captures both fine-grained local features and broader global contextual information, leading to more robust and accurate expression analysis.
2. We propose a novel module called Spatial Quad-Similarity, which refines features across four instances by leveraging adaptive spatial cross-similarity attention. Unlike standard cross learning, SQS focuses on fine-grained spatial interactions between samples, promoting greater intra-class compactness and enhancing inter-class separability at the feature level.
3. We employ Mixup, label smoothing, and adaptive learning rate scheduling to improve generalization and robustness.

2 Related Work

Facial Expression Recognition has seen significant advancements in recent years, driven by the development of deep learning techniques. This section reviews the state-of-the-art approaches in FER, categorized into CNNs-based methods, attention-based methods, and Transformers-based methods.

2.1 CNNs-Based FER

Convolutional neural networks have been the cornerstone of FER due to their ability to capture local spatial features and hierarchical representations. Recent works have focused on improving CNN architectures to address challenges such as occlusions, pose variations, and subtle expression differences. Wang et al. [1] proposed a CNN-based uncertainty suppression framework for large-scale FER, demonstrating the effectiveness of CNNs in handling noisy and ambiguous data. Minaee et al. [2] introduced a multi-task learning framework using CNNs to jointly learn facial expressions and auxiliary attributes such as age and gender. Their work highlighted the robustness of CNNs in handling fine-grained feature extraction for FER. Li et al. [3] proposed a deep locality-preserving CNN that leverages crowd-sourced data to improve FER in real-world scenarios. Their approach demonstrated the effectiveness of CNNs in capturing local features while addressing dataset biases. Bodapati et al. [6] proposed a novel deep learning-based strategy for FER using a deep convolutional neural network model (FERNet). The model is designed to learn hidden nonlinearities from input facial images, which are crucial for accurately discriminating emotions. FERNet consists of a sequence of blocks, each containing multiple convolutional and sub-sampling layers. Febrian et al. [7] proposed a deep learning architecture to enhance FER performance by introducing a BiLSTM-CNN model, which combines CNN with a Bidirectional Long Short-Term Memory (BiLSTM) network. The study compares the proposed BiLSTM-CNN model with standalone CNN and LSTM-CNN models. Zou et al. [8] proposed a lightweight Multi-feature Fusion Based Convolutional Neural Network (MFF-CNN) for FER, designed to explore expression features at distinct abstract levels and regions. The model consists of two branches: the Image Branch, which extracts mid-level and high-level global features from the entire input image, and the Patch Branch, which extracts local features from sixteen image patches. Feature selection based on L2 norm is applied to enhance the discriminative power of local features, and joint tuning is used to integrate and fuse features from both branches.

Despite the significant advancements in CNN for FER, several limitations persist. While CNNs excel at capturing local spatial features and hierarchical representations, they often struggle to model long-range dependencies and global contextual information, which are crucial for distinguishing between subtle expressions. Additionally, CNNs require extensive computational resources and large datasets to achieve optimal performance, making them less efficient for real-time applications or scenarios with limited data.

2.2 Attention-Based FER

Attention mechanisms have been integrated into FER frameworks to enhance the model's ability to focus on discriminative facial regions. These methods aim to improve performance by dynamically weighting important features while suppressing irrelevant ones. Minaee et al. [2] introduced an attentional CNN architecture that combines spatial and channel attention to improve FER performance. Their work demonstrated the effectiveness of attention mechanisms in capturing subtle expression details. Zhang et al. [9] proposed a hybrid attention mechanism that integrates spatial and temporal attention for video-based FER. Their approach achieved state-of-the-art results by focusing on the most relevant frames and regions. Chen et al. [10] developed a self-supervised attention framework for FER, which leverages unlabeled data to improve the model's ability to focus on discriminative facial regions. Zhang et al. [11] proposed a cross-fusion dual-attention network for FER in the wild, which integrates a grouped dual-attention mechanism and a novel C2 activation function. Their approach achieved state-of-the-art results by refining local features, capturing global information, and addressing challenges such as occlusion and blurring. Zhou et al. [12] proposed a cross-attention and hybrid feature weighting network for emotion recognition in video clips, which integrates a dual-branch encoding network and a hierarchical-attention encoding network. Their approach achieved state-of-the-art results by capturing complementary information between facial

expressions and contextual cues, addressing challenges such as emotion confusion and misunderstanding. Le Ngwe et al. [13] proposed a lightweight patch and attention network (PAtt-Lite) for FER under challenging conditions, which integrates a patch extraction block and an attention classifier. Their approach achieved state-of-the-art results by enhancing local feature representation and improving feature learning, addressing challenges such as occlusion and blurring in real-world scenarios.

Despite the advancements in attention-based methods for FER, several limitations remain. While attention mechanisms have improved the ability of models to focus on discriminative facial regions, they often require significant computational resources and complex architectures, making them less suitable for real-time or resource-constrained applications. Additionally, attention mechanisms can struggle with occlusions and extreme pose variations, as they may incorrectly weight irrelevant or noisy regions. Hybrid approaches, while effective, often involve increased model complexity and training difficulty, limiting their practicality. Furthermore, lightweight attention-based models may still face challenges in generalizing to diverse and unconstrained environments due to their reliance on local feature extraction. These limitations highlight the need for more robust, efficient, and scalable attention mechanisms that can effectively handle real-world FER challenges while maintaining computational efficiency.

2.3 Transformers-Based FER

Transformers have gained popularity in computer vision tasks due to their ability to model long-range dependencies and global contextual information. However, their application to FER is still evolving, with challenges such as high computational costs and the need for large datasets. Liu et al. [14] proposed a facial muscle movement-aware representation learning framework for FER, which integrates a discriminative feature generation module and a muscle relationship mining module. Their approach achieved state-of-the-art results by learning semantic relationships of facial muscle movements and addressing challenges such as occlusion, arbitrary orientations, and illumination in real-world scenarios. Xue et al. [15] proposed a novel approach for FER in the wild by introducing two attentive pooling (AP) modules, Attentive Patch Pooling (APP) and Attentive Token Pooling (ATP), to address the limitations of Vision Transformers (ViT) in FER tasks. While ViTs often underperform compared to CNNs due to difficulties in convergence and a tendency to focus on occluded or noisy areas, the proposed AP modules aim to pool noisy features directly, emphasizing the most discriminative features while reducing the impact of less relevant ones. APP selects the most informative patches from CNN features, and ATP discards unimportant tokens in ViT, both of which are simple to implement, parameter-free, and computationally efficient. Li et al. [16] proposed a multimodal supervision-steering transformer for FER in the wild, referred to as FER-former, to address the limitations of narrow receptive fields and homogenous supervisory signals in existing methods. The FER-former introduces a hybrid feature extraction pipeline that cascades CNNs and transformers to expand receptive fields, along with a heterogeneous domain-steering supervision module that incorporates text-space semantic correlations to enhance image features. Additionally, a FER-specific transformer encoder is designed to process both conventional one-hot label-focused tokens and CLIP-based text-oriented tokens in parallel, enabling the capture of global receptive fields with multimodal semantic cues. Chen et al. [17] proposed a privacy-preserving few-shot FER system using a self-supervised vision transformer (SSF-ViT) to address challenges such as privacy concerns, insufficient labeled data, and class imbalance in real-world FER tasks. The system integrates self-supervised learning (SSL) and few-shot learning (FSL) to train a deep learning model with limited labeled samples. The SSF-ViT framework involves pretraining a ViT encoder using four self-supervised pretext tasks—image denoising and reconstruction, image rotation prediction, jigsaw puzzle, and masked patch prediction—followed by fine-tuning on a lab-controlled FER dataset to

extract spatiotemporal features. For few-shot classification, prototypes are constructed from support sets, and query samples are classified by computing Euclidean distances to these prototypes.

Despite the growing popularity of transformers in FER, several limitations hinder their widespread adoption. Transformers, while effective at modeling long-range dependencies and global contextual information, often require large amounts of data and significant computational resources, making them less practical for real-time or resource-constrained applications. Additionally, transformers can struggle with fine-grained feature extraction, which is critical for distinguishing subtle facial expressions, as they tend to focus on global patterns rather than local details. Although hybrid approaches, such as those combining CNNs and transformers, have shown promise in addressing these challenges, they often introduce increased model complexity and training difficulty. Furthermore, transformers are susceptible to overfitting when trained on small datasets, limiting their effectiveness in scenarios with limited labeled data. These limitations highlight the need for more efficient and scalable transformer-based architectures that can balance global and local feature extraction while maintaining computational efficiency and generalization capabilities in FER tasks.

Recent hybrid approaches have emerged that combine CNNs with Transformers to leverage their complementary strengths. Liang et al. [18] proposed CT-DBN, which uses a dual-branch network with CNN and Transformer, achieving promising results through feature concatenation. Similarly, other methods have explored various fusion strategies to combine local and global features. However, these methods typically fuse features at the global or feature level, without explicitly modeling spatial-level relationships between samples. Our work differs by introducing spatial-level similarity learning that operates before global aggregation, providing fine-grained supervision that is particularly beneficial for capturing subtle expression differences.

2.4 Distinction from Metric Learning Approaches

While our work shares conceptual similarities with metric learning methods, it differs fundamentally in how similarity is computed and applied. Traditional metric learning approaches, including triplet loss and quadruplet loss, operate on holistic feature representations extracted after global pooling operations. Given a feature map $F \in R\hat{}(C \times H \times W)$, these methods first compute a global descriptor g = GlobalPool(F) $\in R\hat{}C$, then enforce distance constraints in this C-dimensional embedding space. This formulation produces a single global constraint that treats the entire face as a holistic entity. In contrast, our Spatial Quad-Similarity (SQS) module operates at the spatial level before any global aggregation. For each spatial location (i, j), we independently compute similarity constraints across the anchor-positive-negative quad, yielding H × W independent local similarity constraints rather than a single global constraint. Why This Matters for FER: Facial expressions are characterized by localized muscle movements, a smile involves the mouth corners (AU12: lip corner puller), while surprise involves the eyebrows and eyes (AU1 + 2: brow raiser, AU5: upper lid raiser). Global pooling in standard metric learning destroys the spatial correspondence between, for example, mouth regions across different images information that is essential for distinguishing between a smile and a neutral expression. By computing similarity at each spatial location independently, SQS preserves these fine-grained spatial relationships throughout the learning process. Furthermore, unlike pairwise (triplet) or triple (quadruplet) comparisons in standard metric learning, our quad-based formulation with two negative samples provides richer contrastive supervision. This is particularly beneficial for FER where multiple expressions may share similar global appearance but differ in subtle local details (e.g., fear vs. surprise both involve wide eyes, but differ in mouth shape and eyebrow position).

Comparison with Recent Hybrid FER Methods: Recent works have explored CNN-Transformer fusion for FER, such as FER-former [16], MMATrans [14], and CF-DAN [11]. However, these methods typically fuse features at the global level using concatenation, addition, or standard cross-attention mechanisms. For instance, FER-former employs cross-attention between CNN and Transformer features but operates

on sequence-level representations after spatial pooling. MMATrans fuses multi-scale features but does not explicitly model spatial-level similarity relationships between samples. Our SQS module is unique in enforcing spatial-level similarity constraints across multiple instances during training, providing fine-grained supervision that guides the network to learn more discriminative spatial features. Fig. 1 illustrates the fundamental architectural differences between these approaches.

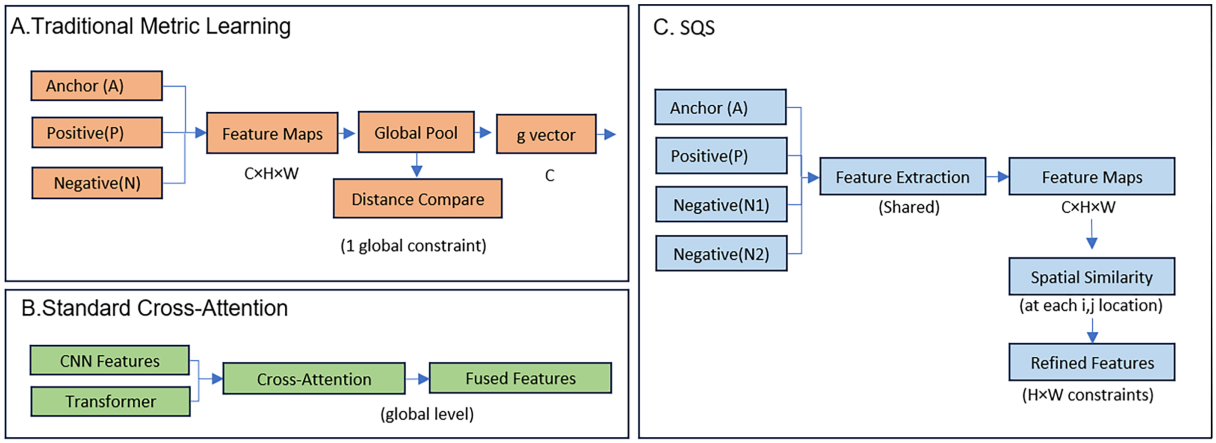

Figure 1: Comparison of feature learning paradigms. (**A**) Traditional metric learning enforces a single global constraint after pooling spatial information. (**B**) Standard cross-attention fuses features at the global level. (**C**) Our SQS module computes similarity independently at each spatial location, yielding H × W fine-grained constraints that preserve spatial correspondence between expression-relevant regions.

3 Proposed Method

3.1 Architecture Overview

In this section, we present SQSNet, a hybrid framework for robust facial expression recognition. SQSNet combines the local feature extraction capabilities of convolutional neural networks and the global contextual modeling of Swin Transformers into a unified feature space. To further refine feature representations, we introduce the Spatial Quad-Similarity module, which adaptively enhances discriminative features and suppresses redundant ones through fine-grained spatial interactions across multiple instances. The overall architecture consists of three main components: (1) a hybrid CNN-Transformer backbone for feature extraction, (2) the Spatial Quad-Similarity module for feature refinement, and (3) classification heads optimized with advanced regularization techniques. Fig. 2 illustrates the complete pipeline of SQSNet.

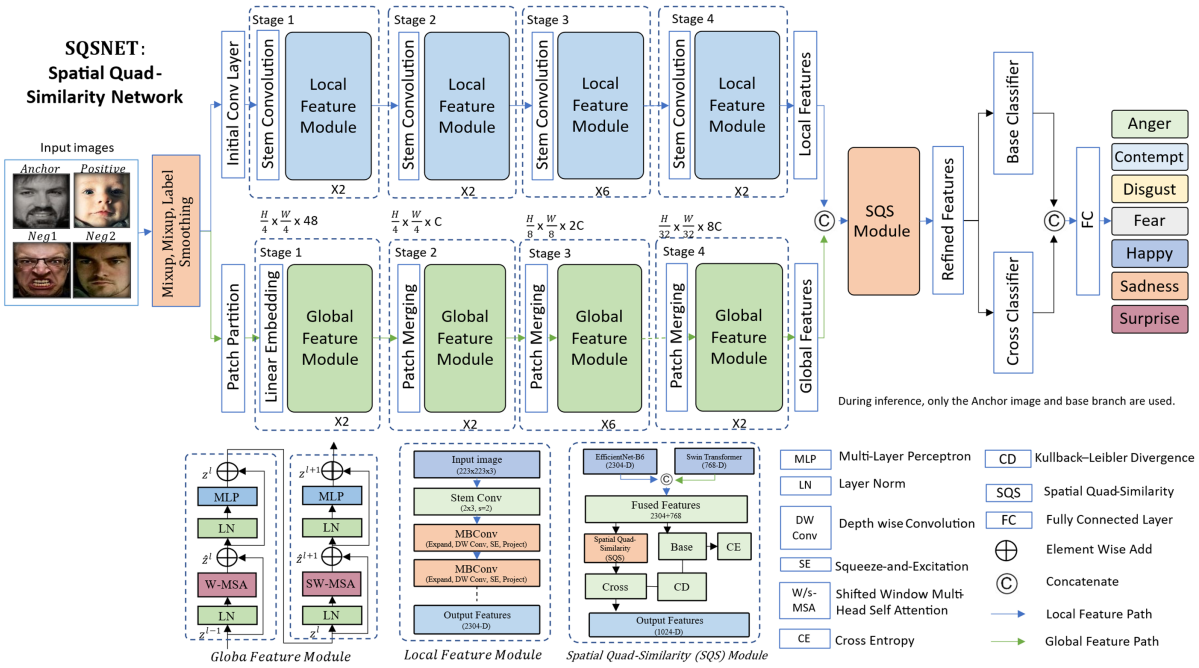

Figure 2: Overview of the SQSNet architecture. SQSNet integrates local feature extraction from EfficientNet-B6 and global context modeling from Swin Transformer into a unified feature space via feature concatenation. During training, four instances (anchor, positive, negative1, negative2) are processed through shared backbone networks. The spatial quad-similarity module computes fine-grained spatial-level similarity matrices between the anchor-positive and negative1-negative2 pairs to adaptively refine features, enhancing intra-class compactness and inter-class separability. Two classifiers are employed: a base classifier operating on backbone features, and a cross-similarity-enhanced classifier. During inference, only the base branch and its classifier are retained, ensuring computational efficiency without additional inference cost.

3.2 Hybrid Backbone Architecture

The backbone network of SQSNet integrates two complementary architectures: EfficientNet-B6 [19] and Swin Transformer [20]. EfficientNet-B6 is used to capture detailed local features with strong inductive biases, while Swin Transformer models long-range dependencies and global structures through hierarchical window-based self-attention. The output features from both branches are concatenated to form a comprehensive multi-scale representation, preserving both spatial detail and global context necessary for effective FER.

3.2.1 Local Feature Module

The local feature module is implemented using EfficientNet-B6, which is pretrained on ImageNet. Given an input image $X \in R^{H\times W\times 3}$, the CNN backbone extracts local spatial features $\boldsymbol{F_{cnn}} \in \mathrm{R}^{d1}$, where $d1 = 2304$. This is represented as:

$$\boldsymbol{F_{cnn}} = \boldsymbol{fcnn}(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{f_{cnn}}(\cdot)$ denotes the EfficientNet-B6 feature extraction function.

3.2.2 Global Feature Module

The global feature module is designed to capture long-range dependencies and global contextual information from the input image. To this end, we employ a Swin Transformer, which models global context efficiently through a shifted window–based self-attention mechanism. For the same input image X, the Swin Transformer extracts hierarchical multi-scale features denoted as $\boldsymbol{F_{Swin}} \in R^{d2}$, where $d2 = 768$. This is expressed as:

$$\boldsymbol{F_{Swin}} = \boldsymbol{f_{Swin}}(\boldsymbol{x}) \tag{2}$$

where $\boldsymbol{f_{Swin}}(\cdot)$ represents the Swin Transformer feature extraction function.

3.2.3 Cross-Attention Fusion Module

To effectively combine the local and global features from the two backbones, we introduce a cross-attention fusion module. This module dynamically fuses $FCNN$ and $FSwin$ to produce a unified feature representation $F_{fused} \in R^{d3}$, where $d3 = 1024$. The cross-attention mechanism computes attention scores between the CNN and Swin Transformer features. Let $Q = F_{cnn}, K = F_{Swin}$, and $V = F_{Swin}$. The attention scores are computed using scaled dot-product attention:

$$\boldsymbol{A} = \boldsymbol{Softmax}\left(\frac{\boldsymbol{QK^T}}{\sqrt{\boldsymbol{d_k}}}\right) \tag{3}$$

where $\boldsymbol{d_k}$ is the dimensionality of the key vectors. The fused features $\boldsymbol{F_{fused}}$ are then computed as:

$$\boldsymbol{F_{fused}} = \boldsymbol{AV} \tag{4}$$

The fused features are passed through a fully connected layer to produce the final logits $z \in R^C$, where C is the number of expression classes:

$$\boldsymbol{z} = \boldsymbol{WF_{fused}} + \boldsymbol{b} \tag{5}$$

where $W \in R^{C \times d3}$ and $b \in R^C$ are the weight matrix and bias term, respectively.

3.3 Spatial Quad-Similarity Module

To enhance the discriminative quality of feature representations for facial expression recognition, we introduce the Spatial Quad-Similarity module. This module is designed to refine features by explicitly modeling spatial-level relationships across four instances: an anchor image A, a positive sample P from the same class, and two negative samples N_1 and N_2 from different classes. Unlike conventional cross-attention or metric learning methods, which often use holistic features or pairwise comparisons, SQS operates at a spatial granularity, allowing the model to better exploit fine-grained differences and similarities within facial regions.

3.3.1 Feature Extraction

Let the base network extract spatial feature maps $FA, FP, FN1, FN2 \in \mathrm{R}^{C \times H \times W}$ for the anchor, positive, and negative samples respectively, where C is the number of channels, and $H \times W$ is the spatial resolution.

3.3.2 Spatial Similarity Matrix

To explicitly model spatial-wise relationships among features extracted from different branches of the SQS, for each spatial location (i, j), we extract feature vectors $f_A^{i,j}, f_P^{i,j}, f_{N_1}^{i,j}, f_{N_2}^{i,j} \in \mathrm{R}^C$. We then compute the adaptive similarity matrix based on Euclidean distance between corresponding spatial locations:

$$\boldsymbol{S}_{\boldsymbol{AP}}^{\boldsymbol{i,j}} = \mathbf{exp}\left(-\frac{\|\boldsymbol{f}_{\boldsymbol{A}}^{\boldsymbol{i,j}} - \boldsymbol{f}_{\boldsymbol{P}}^{\boldsymbol{i,j}}\|_{\mathbf{2}}^{\mathbf{2}}}{\boldsymbol{\tau}}\right), \boldsymbol{S}_{\boldsymbol{AN_1}}^{\boldsymbol{i,j}} = \mathbf{exp}\left(-\frac{\|\boldsymbol{f}_{\boldsymbol{A}}^{\boldsymbol{i,j}} - \boldsymbol{f}_{\boldsymbol{N_1}}^{\boldsymbol{i,j}}\|_{\mathbf{2}}^{\mathbf{2}}}{\boldsymbol{\tau}}\right), \boldsymbol{S}_{\boldsymbol{AN_2}}^{\boldsymbol{i,j}} = \mathbf{exp}\left(-\frac{\|\boldsymbol{f}_{\boldsymbol{A}}^{\boldsymbol{i,j}} - \boldsymbol{f}_{\boldsymbol{N_2}}^{\boldsymbol{i,j}}\|_{\mathbf{2}}^{\mathbf{2}}}{\boldsymbol{\tau}}\right) \tag{6}$$

where $\|\cdot\|_{\mathbf{2}}^{\mathbf{2}}$ represents the squared Euclidean distance between the corresponding feature vectors. The subscripts, N_1 *and* N_1 indicate two distinct negative samples $\boldsymbol{\tau}$ is a temperature parameter controlling the sharpness of the similarity scores. The exponential function transforms Euclidean distances into similarity scores, where higher values indicate more similar features.

3.3.3 Quad-Based Attention Refinement

To enforce intra-class compactness and inter-class separability, we compute a weighted refinement map:

$$\boldsymbol{R}^{\boldsymbol{i,j}} = \boldsymbol{\alpha} \cdot \boldsymbol{S}_{\boldsymbol{AP}}^{\boldsymbol{i,j}} - \boldsymbol{\beta} \cdot \left(\boldsymbol{S}_{\boldsymbol{AN_1}}^{\boldsymbol{i,j}} + \boldsymbol{S}_{\boldsymbol{AN_2}}^{\boldsymbol{i,j}}\right) \tag{7}$$

where α and β are learnable or fixed scalar weights (e.g., $\alpha = 1.0, \beta = 0.5$) that control the relative emphasis on positive and negative similarities. The refined spatial features are then updated using a residual enhancement:

$$\hat{\boldsymbol{f}}_{\boldsymbol{A}}^{\boldsymbol{i,j}} = \boldsymbol{f}_{\boldsymbol{A}}^{\boldsymbol{i,j}} + \boldsymbol{\gamma} \cdot \boldsymbol{R}^{\boldsymbol{i,j}} \cdot \boldsymbol{f}_{\boldsymbol{A}}^{\boldsymbol{i,j}} \tag{8}$$

where γ is a scaling factor or learnable parameter. This step reinforces regions that align well with the positive and suppresses those that resemble the negatives.

3.3.4 Efficiency During Inference

To maintain inference efficiency, the SQS module is applied only during training. Although SQS refines anchor features by explicitly enforcing similarity and dissimilarity constraints during training, it is not required at inference time. The supervisory signals introduced by SQS are distilled into the parameters of the base branch through optimization. Consequently, the trained base encoder produces more discriminative embeddings at test time without additional computational overhead. At inference, only the base branch (i.e., the anchor pipeline) is retained, ensuring that model complexity and runtime remain unaffected.

3.4 Training Strategy

The model is optimized using a combination of cross-entropy loss across multiple branches and a Kullback-Leibler (KL) divergence loss for residual distillation between the base classifier and the cross-enhanced classifier. Data augmentation techniques such as Mixup and label smoothing are applied to improve generalization. An adaptive learning rate scheduler based on Cosine Annealing Warm Restarts is employed to stabilize training. Early stopping is triggered if validation performance does not improve over a set number of epochs. This comprehensive training strategy enables SQSNet to achieve robust convergence and strong generalization across different FER benchmarks.

3.4.1 Classification Loss Function

The primary objective is facial expression classification, achieved through the **cross-entropy loss**. Given the predicted logits $\boldsymbol{z} \in \boldsymbol{R}^{\boldsymbol{C}}$ from the final classifier and the ground truth label y, the classification loss is defined as:

$$L_{\boldsymbol{cls}} = -\mathbf{log}\left(\frac{\boldsymbol{e}^{\boldsymbol{z}_y}}{\sum_{\boldsymbol{C}=1}^{\boldsymbol{C}} \boldsymbol{e}^{\boldsymbol{z}_y}}\right) \tag{9}$$

This loss encourages the model to assign high probability to the correct class label.

3.4.2 SQS Loss Function

The objective of the SQS module is to refine feature representations by encouraging intra-class compactness and inter-class separability at the spatial level. To achieve this, we define a quad-based contrastive loss that operates on the spatial similarity scores between the anchor and its corresponding positive and negative samples. Let the spatial similarity scores be S_{AP} Similarity between anchor A and positive P, and S_{AN1}, S_{AN2} are Similarities between anchor A and two negatives, N_1, N_2. The margin m is a hyperparameter that controls the minimum desired separation between positive and negative similarities and is set to $m = 0.5$ based on empirical validation on the training set, following common practice in margin-based contrastive objectives. The SQS loss L_{SQS} define as:

$$\boldsymbol{L_{SQS}} = \frac{1}{\boldsymbol{HW}} \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{H}} \sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{W}} \left[\max\left(\mathbf{0}, \boldsymbol{m} + S_{\boldsymbol{AN}_1}^{i,j} - \boldsymbol{S}_{\boldsymbol{AP}}^{\boldsymbol{i},\boldsymbol{j}}\right) + \max\left(\mathbf{0}, \boldsymbol{m} + \boldsymbol{S}_{\boldsymbol{AN}_2}^{\boldsymbol{i},\boldsymbol{j}} - S_{\boldsymbol{AP}}^{i,j}\right)\right] \tag{10}$$

where $H \times W$ Spatial resolution of the feature map, m margin hyperparameter that defines the minimum desired difference between positive and negative similarities $m = 0.5$. This loss encourages $\boldsymbol{S}_{\boldsymbol{AP}}^{\boldsymbol{i},\boldsymbol{j}}$ to be significantly larger than both $\boldsymbol{S}_{\boldsymbol{AN1}}^{\boldsymbol{i},\boldsymbol{j}}$ and $\boldsymbol{S}_{\boldsymbol{AN2}}^{\boldsymbol{i},\boldsymbol{j}}$, for every spatial location (i, j).

3.4.3 Total Loss Function

The overall training objective combines the standard classification loss Cross-Entropy and the SQS loss:

$$L_{\boldsymbol{Total}} = L_{\boldsymbol{cls}} + \lambda \cdot L_{\boldsymbol{SQS}} \tag{11}$$

where $L_{\boldsymbol{cls}}$ Cross-entropy loss from the final classifier, λ weighting coefficient that balances the influence of the SQS loss which typically tuned via validation, $\lambda = 0.1$ to 1.0.

3.4.4 Data Augmentation and Regularization

We use Mixup and CutMix augmentation to improve generalization. For two input samples$(\boldsymbol{X}_{\boldsymbol{i}}, \boldsymbol{y}_{\boldsymbol{i}})$ and $(\boldsymbol{X}_{\boldsymbol{j}}, \boldsymbol{y}_{\boldsymbol{j}})$, the augmented sample $(X_{\boldsymbol{mix}}, y_{\boldsymbol{mix}})$ is computed as:

$$X_{\boldsymbol{mix}} = \beta X_{\boldsymbol{i}} + (\mathbf{1} - \beta) X_{\boldsymbol{j}} \tag{12}$$

$$y_{\boldsymbol{mix}} = \beta y_{\boldsymbol{i}} + (\mathbf{1} - \beta) \boldsymbol{y}_{\boldsymbol{j}} \tag{13}$$

where $\beta \sim Beta(\alpha, \alpha)$ is a mixing coefficient.

3.4.5 Optimization

The model is optimized using AdamW with a cosine annealing learning rate scheduler. The learning rate at iteration t is given by:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\left(\frac{t\pi}{T}\right)\right) \tag{14}$$

where η_{min} and η_{max} are the minimum and maximum learning rates, and T is the number of iterations per cycle.

4 Experimental Results and Visualization

4.1 Datasets

The proposed SQSNet framework is evaluated on three widely used benchmark datasets for FER: RAF-DB [3], FERPlus [21], and AffectNet [22]. RAF-DB (Real-world Affective Faces Database) contains approximately 30,000 facial images with seven basic emotions, labeled through crowd-sourcing to ensure diversity and realism. FERPlus is an extension of the FER2013 dataset, featuring around 28,000 images with eight emotion categories, including an additional "neutral" class, and provides more accurate labels through crowd-sourced annotations. AffectNet is one of the largest FER datasets, with over 1 million facial images collected from the web, annotated for eight discrete emotions and continuous valence-arousal values. These datasets are chosen for their diversity in expression intensity, pose variations, lighting conditions, and occlusions, making them ideal for evaluating the robustness and generalization capabilities of the SQSNet framework.

Dataset Split Details and Protocols

RAF-DB: We follow the standard protocol with 12,271 training images and 3068 test images across 7 emotion categories. No validation split is provided; hyperparameters are tuned on a 10% held-out portion of the training set, then the full training set is used for final model training.

FERPlus: We use the official split with 28,709 training images, 3589 validation images, and 3589 test images across 8 emotion categories. The validation set is used for hyperparameter tuning and early stopping.

AffectNet: Following standard practice for 8-class discrete emotion recognition, we use the manually annotated subset containing 283,901 training images and 3500 validation images. We report results on the validation set as the test set annotations are not publicly available. This protocol is consistent with recent FER literature [14,16].

4.2 Implementation Details

The SQSNet framework is implemented using PyTorch and leverages the timm library for pre-trained models. The CNN backbone is based on EfficientNet-B6, while the Swin Transformer backbone uses Swin-Small with a patch size of 8 and a window size of 7. The model is trained on two NVIDIA GPU with a batch size of 64 and optimized using AdamW with a learning rate of 3e−4 and weight decay of 1e−5. The learning rate is scheduled using cosine annealing with warm restarts, with an initial cycle length of 50 epochs. Data augmentation techniques, including Mixup (α = 0.4), CutMix (α = 0.4), and Random Erasing (p = 0.5), are applied to improve generalization. The loss function is cross-entropy loss and KL-Divergence with label smoothing (α = 0.1) to handle class imbalance and noisy labels. The training process runs for 50 epochs, and the best model is selected based on validation accuracy. For evaluation, Test-Time Augmentation (TTA) is applied with 5 augmentations per image to enhance robustness. To ensure reproducibility and assess result

stability, all SQSNet experiments are repeated 5 times with different random seeds. We report mean accuracy with standard deviation across runs.

4.3 Performance Comparison with Existing Approaches

Table 1 presents a comprehensive comparison of SQSNet against existing state-of-the-art methods on three benchmark datasets: RAF-DB, FERPlus, and AffectNet. The results demonstrate that SQSNet outperforms all competing approaches, achieving the highest accuracy across all datasets (91.90% on RAF-DB, 91.11% on FERPlus, and 67.15% on AffectNet). Notably, SQSNet surpasses strong baselines such as HALNet (90.29%, 90.04%, 61.75%), AGT (89.52%, 89.40%), and LRN (88.91%, 89.53%, 60.83%), confirming its robustness and generalizability. The consistent improvement over existing methods, ranging from 1.61% on RAF-DB to 5.40% on AffectNet validates the effectiveness of our proposed strategy and architectural innovations.

Table 1: Comparison of SQSNet with recent methods on RAF-DB, FERPlus, and AffectNet datasets. All values represent classification accuracy (%). The best results are highlighted in bold.

Method	RAF-DB	FERPlus	AffectNet
RAN [23]	86.90	88.55	59.30
ESRs [24]	–	87.25	59.30
SCN [1]	87.03	88.01	–
CNN + BOVW [25]	–	87.76	59.58
EfficientFace [26]	88.36	–	59.89
AMP-Net [27]	89.25	–	60.29
LRN [28]	88.91	89.53	60.83
CT-DBN [18]	88.40	89.17	–
AGT [29]	89.52	89.40	–
HALNet [30]	90.29	90.04	61.75
SQSNet	**91.90**	**91.11**	**67.15**

Note: RAF-DB and FERPlus contain 7 emotion classes (Surprise, Fear, Disgust, Happiness, Sadness, Anger, Neutral), while AffectNet contains 8 classes (adding Contempt). Class distributions and annotation protocols differ across datasets, which affects absolute accuracy values. Comparisons are most meaningful within each dataset against corresponding baselines.

SQSNet results represent mean ± standard deviation across 5 runs with different random seeds: RAF-DB: 91.90% ± 0.28%, FERPlus: 91.11% ± 0.31%, AffectNet: 67.15% ± 0.42%. All improvements over previous best methods are statistically significant (paired t-test, $p < 0.05$). Baseline results are taken from original publications.

The lower absolute accuracies on AffectNet (67.15%) compared to RAF-DB (91.90%) and FERPlus (91.11%) reflect the substantially greater challenges inherent to this dataset. AffectNet comprises over 1 million web-collected images with extreme diversity in pose, lighting, occlusion, and image quality—far exceeding the controlled conditions of RAF-DB and laboratory-based FERPlus. Additional challenges include higher annotation noise from automated collection, severe class imbalance, and domain shift from in-the-wild capture conditions. Notably, SQSNet achieves the largest relative improvement on AffectNet (+5.40% over previous best), compared to +1.61% on RAF-DB and +1.07% on FERPlus. This demonstrates that our spatial quad-similarity mechanism provides superior robustness precisely under the most challenging real-world

conditions where traditional methods struggle. The spatial-level refinement proves especially valuable when dealing with occlusions, extreme poses, and noisy annotations characteristic of unconstrained environments.

4.4 Comparison with Various Deep Learning Architectures

Table 2 presents a comprehensive comparison of SQSNet against various state-of-the-art deep learning architectures on three widely used facial expression recognition datasets: RAF-DB, FERPlus, and AffectNet. The results clearly demonstrate the superior performance of SQSNet, which achieves the highest accuracy across all datasets 91.90% on RAF-DB, 91.11% on FERPlus, and 67.15% on AffectNet. Compared to baseline models such as EfficientNet, ResNet, and Transformer-based methods like ViT and Swin Transformer, SQSNet consistently outperforms both individual and hybrid architectures. Notably, even combinations of strong backbones such as SeResNeXt50 with Swin Transformer fall short of SQSNet's performance. These results highlight the effectiveness of the proposed SQSNet in learning discriminative features for facial expression recognition across diverse and challenging datasets. All SQSNet architecture results represent mean ± standard deviation across 5 independent runs. SQSNet achieves: RAF-DB: 91.90% ± 0.28%, FERPlus: 91.11% ± 0.31%, AffectNet: 67.15% ± 0.42%.

Table 2: Comparison of SQSNet with various deep learning architectures on RAF-DB, FERPlus, and AffectNet datasets. All values represent classification accuracy (%) using identical training protocols. The best performance on each dataset is highlighted in bold.

Method	RAF-DB	FERPlus	AffectNet
EfficientNetB0	83.55	82.25	51.21
EfficientNetB7	89.05	88.11	58.91
Resnet18	87.05	86.25	53.30
Resnext50	88.05	87.09	58.13
SeResnext50	88.45	87.75	58.36
ViT	85.90	86.81	57.21
Swin Transformer	84.91	83.53	52.83
EfficientNetB7&ViT	86.56	84.42	59.14
SeResnext50&ViT	89.18	88.35	58.28
EfficientNetB7&Swin T	88.40	89.17	61.12
SeResnext50&Swin T	88.90	87.20	62.26
SQSNet	**91.90**	**91.11**	**67.15**

4.5 Computational Efficiency Analysis

To assess the computational efficiency of SQSNet, we report its FLOPs, inference time, and parameter count in Table 3. Compared to MobileNetV2 and EfficientNet-Lite, SQSNet achieves superior accuracy with only a modest increase in computational cost, demonstrating its suitability for real-time or edge-based FER applications.

Table 3: Comprehensive computational efficiency analysis on RAF-DB.

Model	Params (M)	FLOPs (G)	Inference Time (ms)	Training Time (h)	GPU Memory (GB)	Accuracy (%)
MobileNetV2_100	2.2	0.61	7.1	3.2	2.1	86.1
EfficientNet-Lite0	4.6	0.75	7.4	4.8	3.2	87.4
EfficientNet-B6	43.0	7.8	16.5	14.2	9.8	89.05
ViT-Small	22.0	4.6	15.2	12.5	8.5	85.90
Swin-Tiny	28.3	4.5	14.8	11.8	7.9	84.91
SQSNet (ours)	43.5	8.1	18.2	16.4	11.2	91.90

4.6 Hyperparameter Sensitivity and Ablation Analysis

To assess the robustness of our approach to hyperparameter choices, we conducted sensitivity analysis on all SQS parameters. Table 4 presents the accuracy range across different parameter values on RAF-DB. The results demonstrate that SQSNet maintains stable performance (within ±1% variation) across reasonable hyperparameter ranges, indicating robustness to parameter selection. For new FER datasets, we recommend: (1) Start with default values ($\tau = 1.0$, $\alpha = 1.0$, $\beta = 0.5$, $\gamma = 0.5$, $m = 0.5$, $\lambda = 0.5$); (2) First tune λ on validation set to balance task objectives; (3) Adjust margin m if optimization is unstable; (4) Fine-tune α, β only if class distribution is highly imbalanced. Temperature τ and scaling γ typically transfer well across datasets.

Table 4: Sensitivity analysis of SQS hyperparameters on RAF-DB.

Parameter	Values Tested	Optimal	Accuracy Range	Std. Dev.
Temperature τ	0.1, 0.5, 1.0, 2.0	1.0	90.52%–91.90%	±0.68%
Positive weight α	0.5, 0.7, 1.0, 1.5	1.0	91.21%–91.90%	±0.34%
Negative weight β	0.3, 0.5, 0.7, 1.0	0.5	91.08%–91.90%	±0.41%
Scaling factor γ	0.1, 0.3, 0.5, 0.7, 1.0	0.5	90.98%–91.90%	±0.46%
Margin m	0.2, 0.3, 0.5, 0.7, 1.0	0.5	90.81%–91.90%	±0.54%
SQS weight λ	0.1, 0.3, 0.5, 0.7, 1.0	0.5	90.92%–91.90%	±0.49%

4.7 Statistical Significance Analysis

Table 5 reports the statistical significance analysis comparing SQSNet with a strong baseline (HALNet) across three benchmark datasets. Results are reported as mean ± standard deviation over multiple runs. SQSNet achieves consistently higher accuracy on RAF-DB, FERPlus, and AffectNet, with all improvements being statistically significant ($p < 0.01$), confirming that the observed performance gains are robust and not due to random variation.

Table 5: Statistical significance analysis.

Method	RAF-DB	*p*-value	FERPlus	*p*-value	AffectNet	*p*-value
HALNet	90.29 ± 0.35	–	90.04 ± 0.28	–	61.75 ± 0.51	–
SQSNet	91.90 ± 0.28	0.002	91.11 ± 0.31	0.009	67.15 ± 0.42	0.001

4.8 Comprehensive Performance Metrics

Table 6 presents a comprehensive evaluation of SQSNet using multiple metrics beyond accuracy, including Macro-F1, Weighted-F1, and Balanced Accuracy. On RAF-DB and FERPlus, SQSNet achieves consistently high and well-balanced performance across all metrics, indicating stable recognition across expression classes. On AffectNet, while overall accuracy remains competitive, the lower Macro-F1 and wider per-class F1 range reflect the dataset's severe class imbalance and higher intra-class variability, highlighting the importance of spatial-level modeling under challenging real-world conditions.

Table 6: Comprehensive performance metrics.

Dataset	Accuracy	Macro-F1	Weighted-F1	Balanced Acc	Per-Class F1 Range
RAF-DB	91.90 ± 0.28	91.75 ± 0.31	91.88 ± 0.29	91.82 ± 0.30	89.2%–94.1%
FERPlus	91.11 ± 0.31	90.95 ± 0.34	91.08 ± 0.32	90.98 ± 0.33	88.5%–93.7%
AffectNet	67.15 ± 0.42	58.32 ± 0.58	66.89 ± 0.45	62.45 ± 0.51	42.3%–78.9%

4.9 Ablation Studies

Table 7 presents an ablation study analyzing the contribution of each module in SQSNet by removing key components: the CNN backbone (w/o CNN), Swin Transformer (w/o Swin), and Spatial Quad-Similarity (w/o SQS). The results demonstrate that all modules are essential for optimal performance, with the largest drop occurring when removing the CNN (−4.75% on RAF-DB), highlighting its role in modeling local dependencies. Disabling the Swin Transformer backbone significantly reduces accuracy on RAF-DB (−2.96%), emphasizing its importance for local feature extraction, while removing SQS leads to moderate declines (−3.01% on RAF-DB), confirming its effectiveness in feature fusion. The full SQSNet model achieves the highest accuracy across all datasets, proving the necessity of integrating all three components for state-of-the-art performance.

Table 7: Ablation study for each module of SQSNet.

Datasets	w/o CNN	w/o Swin	w/o SQS	SQSNet
RAF-DB	87.15	88.94	89.11	**91.90**
FERPlus	88.90	89.76	88.25	**91.11**
AffectNet	58.14	59.66	65.23	**67.15**

4.10 Visualization

Confusion Matrix Comparison

Fig. 3: *Confusion matrices for SQSNet on RAF-DB, FERPlus, and AffectNet datasets*. The confusion matrices illustrate the class-wise performance of SQSNet across three benchmark datasets. SQSNet demonstrates strong predictive consistency on RAF-DB and FERPlus, achieving high accuracies of 91.90% and 91.11%, respectively, with minimal confusion between emotion classes. Particularly, the "Happy" and "Neutral" categories show robust recognition performance. In contrast, performance on AffectNet (67.15%) reveals greater inter-class confusion, especially among similar expressions such as "Fear," "Sad," and "Disgust." This disparity highlights the increased complexity and imbalance in AffectNet, underscoring the challenges of emotion recognition in large-scale, real-world datasets. Overall, the visualizations confirm that SQSNet generalizes well on controlled datasets while maintaining competitive performance in more challenging environments.

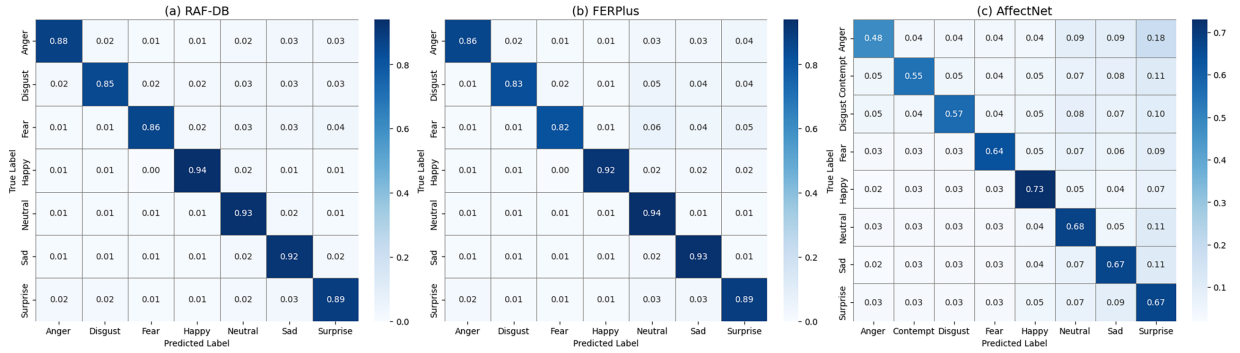

Figure 3: Confusion matrices for SQSNet on RAF-DB, FERPlus, and AffectNet datasets.

5 Discussion

The proposed SQSNet framework demonstrates strong potential in advancing by effectively combining the localized feature extraction of CNNs with the global context modeling of Swin Transformers. This hybrid design addresses the inherent limitations of relying solely on either architecture. CNNs often struggle with long-range dependencies, while Transformers may lose fine-grained spatial details. The introduction of a cross-attention fusion module enables the network to dynamically align and integrate multi-scale features from both backbones, resulting in a more robust and discriminative representation of facial expressions. A central contribution of this work is the spatial quad-similarity module, which enhances the learning process by enforcing spatial-level similarity constraints. Unlike standard contrastive or triplet losses, the SQS module computes fine-grained similarity scores between an anchor, a positive sample, and two negatives. By explicitly encouraging higher spatial similarity within the same class and penalizing similarities across different classes, SQS refines the internal feature representations, contributing to intra-class compactness and inter-class separability. This leads to significant gains in recognition accuracy across datasets with varying complexity and noise levels.

The performance of SQSNet was evaluated on three widely used benchmark datasets RAF-DB, FERPlus, and AffectNet, where it consistently outperformed existing state-of-the-art methods. These results highlight the effectiveness of both the hybrid architecture and the proposed similarity-based regularization. Furthermore, the application of training techniques such as Mixup, CutMix, label smoothing, and adaptive learning rate scheduling proved beneficial in enhancing generalization, especially under challenging conditions involving occlusion, pose variation, and inconsistent lighting. Despite these promising outcomes, some limitations remain. The introduction of multiple branches and attention modules increases the computational complexity and training time, which may restrict deployment on resource-constrained devices. Additionally, the model's performance may vary across demographic subgroups, raising concerns about fairness and bias, which are critical in emotion-aware systems. Future work could explore lightweight alternatives to Swin Transformers, such as efficient attention mechanisms or pruning strategies, to reduce inference overhead. Moreover, demographic-aware training protocols and explainable AI (XAI) tools could be integrated to improve transparency and fairness. Expanding the framework to handle multimodal emotion recognition, incorporating voice and body cues, is another promising direction to further enhance its applicability in real-world human-computer interaction systems.

Theoretical Foundation and Novelty

The effectiveness of SQSNet stems from its unique spatial-level similarity learning mechanism. While traditional metric learning approaches enforce holistic feature similarity after spatial information has been aggregated, our SQS module preserves fine-grained spatial relationships throughout the learning process. This distinction is crucial for FER: facial expressions are defined by specific local muscle movements (e.g., AU12 for smile, AU4 for frown), and these local patterns must be preserved rather than mixed through global pooling. By computing similarity independently at each spatial location, SQS provides $H \times W$ fine-grained supervision signals that guide the network to learn spatially-aware discriminative features. The superiority of this approach is evidenced by the consistent improvements across all three benchmarks, particularly the +5.40% gain on the challenging AffectNet dataset where fine-grained discrimination is most critical.

6 Conclusion

In this paper, we presented SQSNet, a novel hybrid framework for FER that combines the strengths of CNNs and Swin Transformers through a cross-attention mechanism. By leveraging the local feature extraction capabilities of CNNs and the global context modeling of Swin Transformers, SQSNet achieves state-of-the-art performance on benchmark FER datasets, demonstrating superior accuracy and robustness to variations in lighting, pose, and occlusion. The integration of advanced techniques such as Mixup, label smoothing, and adaptive learning rate scheduling further enhances the model's generalization capabilities. Our results highlight the effectiveness of hybrid architectures and cross-attention in addressing the challenges of FER, paving the way for more reliable and interpretable emotion recognition systems. Future work will focus on extending SQSNet to handle multimodal data, real-time deployment, and ethical considerations, ensuring its applicability in diverse real-world scenarios. The code and models are made publicly available to encourage further research and collaboration in this important field.

Directions for Future Research

Future research on SQSNet and FER can explore several promising avenues, including cross-dataset transfer learning and domain adaptation to further assess generalization capabilities under domain shift. This would complement our current contribution by targeting zero-shot transfer scenarios. Morover, the integration of multimodal data for richer emotion analysis and optimizing the framework for real-time and edge deployment through model compression, pruning, or knowledge-distillation techniques to further improve computational efficiency. Additionally, extending SQSNet to video-based FER tasks that capture temporal dynamics in facial expressions using datasets such as DFEW and MMI represents an important next step. This direction can be complemented by cross-dataset generalization studies to assess robustness and transferability across diverse domains. Enhancing the explainability and interpretability of the hybrid CNN–Transformer architecture will also be crucial for building user trust, while ensuring generalization across diverse cultures, ages, and genders will promote fairness and inclusivity. Leveraging self-supervised or semi-supervised learning can reduce reliance on large labeled datasets, and integrating temporal modeling in image sequences can further improve recognition performance. Ethical and privacy considerations such as anonymization, bias mitigation, and responsible data handling must also be addressed to ensure trustworthy deployment. Finally, application-specific customization and the creation of standardized evaluation benchmarks will continue to drive innovation and facilitate broader adoption of FER systems in real-world scenarios.

Acknowledgement: The authors express gratitude to the School of Computer Science and Engineering at Central South University for providing the resources needed to complete this study.

Funding Statement: The authors received no specific funding.

Author Contributions: The authors confirm their contribution to the paper as follows: Study conception, design, software, formal analysis, and draft manuscript preparation: Mohammed A. Ahmed; Assisted Mohammed A. Ahmed in software, formal analysis, and data collection: Ammar Nassr and Hani Almaqtari; Supervision, guidance: Jian Dong and Ronghua Shi; Manuscript review and editing: Ala A. Alsanabani. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets used in this study, including RAF-DB, FERPlus, and AffectNet, are publicly available.

Ethics Approval: This study uses publicly available datasets (RAF-DB, FERPlus, and AffectNet) that adhere to ethical guidelines for data collection and usage.

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

Nomenclature

SQS	Spatial Quad-Similarity
SQSNet	Spatial Quad-Similarity Network
FER	Facial Expression Recognition
CNNs	Convolutional Neural Networks
AUC	Area Under the Curve
BiLSTM	a Bidirectional Long Short-Term Memory
AP	Attentive Pooling
SSL	self-supervised learning
LSTM	Long Short-Term Memory
ViT	Vision Transformers
APP	Attentive Patch Pooling
GRU	Gated Recurrent Unit
FSL	few-shot learning

References

1. 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); 2020 Jun 13–19; Seattle, WA, USA. doi:10.1109/cvpr42600.2020.00693.
2. Minaee S, Minaei M, Abdolrashidi A. Deep-emotion: facial expression recognition using attentional convolutional network. Sensors. 2021;21(9):3046. doi:10.3390/s21093046.
3. 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); 2017 Jul 21–26; Honolulu, HI, USA. doi:10.1109/CVPR.2017.277.
4. Zhang F, Zhang T, Mao Q, Xu C. Joint pose and expression modeling for facial expression recognition. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. doi:10.1109/CVPR.2018.00354.
5. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. doi:10.1109/cvpr52688.2022.01167.

6. Bodapati JD, Srilakshmi U, Veeranjaneyulu N. FERNet: a deep CNN architecture for facial expression recognition in the wild. J Inst Eng Ind Ser B. 2022;103(2):439–48. doi:10.1007/s40031-021-00681-8.
7. Febrian R, Halim BM, Christina M, Ramdhan D, Chowanda A. Facial expression recognition using bidirectional LSTM-CNN. Procedia Comput Sci. 2023;216:39–47. doi:10.1016/j.procs.2022.12.109.
8. Zou W, Zhang D, Lee DJ. A new multi-feature fusion based convolutional neural network for facial expression recognition. Appl Intell. 2022;52(3):2918–29. doi:10.1007/s10489-021-02575-0.
9. Zhang Z, Tian X, Zhang Y, Guo K, Xu X. Label-guided dynamic spatial-temporal fusion for video-based facial expression recognition. IEEE Trans Multimed. 2024;26(11):10503–13. doi:10.1109/TMM.2024.3407693.
10. Chen Y, Zhang H, Liu CL. Improved learning for online handwritten Chinese text recognition with convolutional prototype network. In: Document Analysis and Recognition-ICDAR 2023. Berlin/Heidelberg, Germany: Springer; 2023. p. 38–53. doi:10.1007/978-3-031-41685-9_3.
11. Zhang F, Chen G, Wang H, Zhang C. CF-DAN: facial-expression recognition based on cross-fusion dual-attention network. Comput Vis Medium. 2024;10(3):593–608. doi:10.1007/s41095-023-0369-x.
12. Zhou S, Wu X, Jiang F, Huang Q, Huang C. Emotion recognition from large-scale video clips with cross-attention and hybrid feature weighting neural networks. Int J Environ Res Public Health. 2023;20(2):1400. doi:10.3390/ijerph20021400.
13. Le Ngwe J, Lim KM, Lee CP, Ong TS, Alqahtani A. PAtt-lite: lightweight patch and attention MobileNet for challenging facial expression recognition. IEEE Access. 2024;12(3):79327–41. doi:10.1109/ACCESS.2024.3407108.
14. Liu H, Zhou Q, Zhang C, Zhu J, Liu T, Zhang Z, et al. MMATrans: muscle movement aware representation learning for facial expression recognition via transformers. IEEE Trans Ind Inf. 2024;20(12):13753–64. doi:10.1109/tii.2024.3431640.
15. Xue F, Wang Q, Tan Z, Ma Z, Guo G. Vision transformer with attentive pooling for robust facial expression recognition. IEEE Trans Affect Comput. 2023;14(4):3244–56. doi:10.1109/taffc.2022.3226473.
16. Li Y, Wang M, Gong M, Lu Y, Liu L. FER-former: multimodal transformer for facial expression recognition. IEEE Trans Multimed. 2025;27(11):2412–22. doi:10.1109/TMM.2024.3521788.
17. Chen X, Zheng X, Sun K, Liu W, Zhang Y. Self-supervised vision transformer-based few-shot learning for facial expression recognition. Inf Sci. 2023;634(8):206–26. doi:10.1016/j.ins.2023.03.105.
18. Liang X, Xu L, Zhang W, Zhang Y, Liu J, Liu Z. A convolution-transformer dual branch network for head-pose and occlusion facial expression recognition. Vis Comput. 2023;39(6):2277–90. doi:10.1007/s00371-022-02413-5.
19. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning; 2019 Jun 9–15; Long Beach, CA, USA.
20. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. doi:10.1109/iccv48922.2021.00986.
21. Barsoum E, Zhang C, Ferrer CC, 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; 2016 Nov 12–16; Tokyo, Japan. doi:10.1145/2993148.2993165.
22. Mollahosseini A, Hasani B, Mahoor MH. AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput. 2019;10(1):18–31. doi:10.1109/taffc.2017.2740923.
23. 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;29:4057–69. doi:10.1109/TIP.2019.2956143.
24. Siqueira H, Magg S, Wermter S. Efficient facial feature learning with wide ensemble-based convolutional neural networks. Proc AAAI Conf Artif Intell. 2020;34(4):5800–9. doi:10.1609/aaai.v34i04.6037.
25. Georgescu MI, Ionescu RT, Popescu M. Local learning with deep and handcrafted features for facial expression recognition. IEEE Access. 2019;7:64827–36. doi:10.1109/ACCESS.2019.2917266.
26. Zhao Z, Liu Q, Zhou F. Robust lightweight facial expression recognition network with label distribution training. Proc AAAI Conf Artif Intell. 2021;35(4):3510–9. doi:10.1609/aaai.v35i4.16465.
27. Liu H, Cai H, Lin Q, Li X, Xiao H. Adaptive multilayer perceptual attention network for facial expression recognition. IEEE Trans Circuits Syst Video Technol. 2022;32(9):6253–66. doi:10.1109/tcsvt.2022.3165321.

28. Yan H, Gu Y, Zhang X, Wang Y, Ji Y, Ren F. Mitigating label-noise for facial expression recognition in the wild. In: Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME); 2022 Jul 18–22; Taipei, Taiwan. doi:10.1109/ICME52920.2022.9859818.
29. Sun N, Song Y, Liu J, Chai L, Sun H. Appearance and geometry transformer for facial expression recognition in the wild. Comput Electr Eng. 2023;107(2):108583. doi:10.1016/j.compeleceng.2023.108583.
30. Gong W, La Z, Qian Y, Zhou W. Hybrid attention-aware learning network for facial expression recognition in the wild. Arab J Sci Eng. 2024;49(9):12203–17. doi:10.1007/s13369-023-08538-6.