informatics

MDPI

Article

SE-POSTER: Channel-Enhanced Landmark Guided Transformer for Facial Emotion Recognition

Alpamis Kutlimuratov [1], Kongratbay Sharipov [2], Piratdin Allayarov [3], Sayyora Iskandarova [4], Ruslan Latyfskiy [5], Gulchehra Tolibaeva [6] and Fazliddin Makhmudov [7,8,*]

[1] Department of Applied Informatics, Kimyo International University in Tashkent, Tashkent 100121, Uzbekistan; kutlimuratov.aj@kiut.uz
[2] Tashkent Institute of Irrigation and Agricultural Mechanization Engineers, Tashkent 100000, Uzbekistan; k.sharipov@edu.uz
[3] Department of Econometrics, Tashkent State University of Economics, Tashkent 100066, Uzbekistan; p.allayarov@tsue.uz
[4] Department of Computer Systems, Tashkent University of Information Technologies Named After Muhammad al-Khwarizmi, Tashkent 100084, Uzbekistan; sayyora5@mail.ru
[5] Department of Information Technology and Automation of Technological Processes and Production, National University of Science and Technology MISIS in Almalyk, Almalyk 110100, Uzbekistan; latrused@gmail.com
[6] Department of General Education Disciplines and Distance Education, Nukus State Pedagogical Institute Named After Ajiniyaz, Nukus 230100, Uzbekistan; gulchexra.tolibaeva@ndpi.uz
[7] Department of Computer Engineering, Gachon University, Seongnam 13120, Republic of Korea
[8] School of Digital Technologies, American University of Technology, Tashkent 100060, Uzbekistan
* Correspondence: fazliddin12@gachon.ac.kr

Abstract

Recognizing facial emotions automatically from images/videos (FER) still represents a difficult problem for emotion computing, mainly due to variations in the face pose, lighting, occlusion, facial features, and expression intensity in the wild. Recent CNN–Transformer-based hybrid models like POSTER have leveraged local feature learning, landmark guidance, and global dependency modeling to achieve strong performance. Yet these methods give the main focus to spatial and contextual representations while not really going deep into adaptive channel-wise feature importance over multi-scale representations. As different feature channels represent emotions in varying degrees, it is likely that by treating all feature channels equally, one would limit the ability of the learned features to discriminate effectively. To overcome this weakness, this article presents a ResNet-18–Transformer landmark-guided module called SE-POSTER that fuses lightweight Squeeze-and-Excitation (SE) attention modules into the multi-scale feature pyramid of the baseline POSTER architecture. The proposed method carries out feature channel recalibration adaptively at the level of features before Transformer-based global attention modeling, thus allowing the network to focus on emotionally informative feature channels and suppress less relevant responses. The inclusion of SE attention in the network enhances fine, mid, and global levels of feature representations at a very low cost in terms of computation. On the basis of the RAF-DB, FERPlus, and AffectNet datasets, enormous experiments prove that the SE-POSTER framework proposed is capable of steadily boosting recognition accuracy relative to the baseline POSTER and several state-of-the-art FER methods. Especially, the proposed model delivers 92.78% accuracy on RAF-DB while it also shows better robustness and generalization capability under difficult real-world conditions. Moreover, additional ablation studies reveal that multi-level channel recalibration is effective in improving discriminative emotional feature learning.

Check for updates

Academic Editors: Roberto Theron, Patricia Anthony and Jing Zhou

Received: 6 May 2026
Revised: 6 July 2026
Accepted: 28 July 2026
Published: 30 July 2026

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

Keywords: facial emotion recognition; channel-wise attention; Transformer encoder; facial landmark

1. Introduction

FER is an essential aspect of affective computing, which helps machines to understand human emotional state based on the visual facial cues [1]. Thorough FER is indispensable to a wide range of real-life applications, such as human, computer interaction, mental health monitoring, intelligent tutoring system, driven safety analysis, and social robotics [2]. In contrast to tightly controlled laboratory environments, real-life FER systems are expected to work under very difficult conditions where there are changes in illumination, pose, occlusion, facial appearance [3], and background clutter. All these factors cause a significant rise in intra-class variability and a decrease in inter-class separability, thus making the task of emotion recognition a highly challenging computer vision problem [4]. Handcrafted features such as Local Binary Pattern (LBP), Histogram of Oriented Gradient (HOG) and Gabor filters with conventional classifiers such as support vector machine and k-nearest neighbors were the main components of the traditional FER system [5,6]. Although these approaches yielded interpretable representation, they were not expressive enough to represent features and were incapable of capturing intricate emotional patterns which led to their performance being limited in the wild [7]. With the coming of deep learning, especially CNNs, the facial expression recognition task has undergone a paradigm shift as CNNs can learn multi-level features automatically from raw images [8]. Facial expression recognition models based on CNNs outperformed other methods as they effectively captured the local texture, edge and structural detail of the facial expression [9]. But CNNs, by nature, involve only local receptive fields and convolution operations, which makes it difficult for them to understand long-range dependencies and global contextual relationships between spatially remote facial regions [10].

Recently, researchers have introduced Transformer architectures into computer vision to solve these problems. Transformers use self-attention mechanisms to recognize global relations of spatial features and thus gain an even better understanding of the context than pure convolutional approaches [11]. In fact, Vision Transformers (ViTs) and hybrid CNN–Transformer models have performed exceptionally well in a number of computer vision tasks, including facial emotion recognition [12]. Moreover, the combination of CNN for extracting local features and Transformer for capturing global dependencies in hybrid models allows each of the paradigms to compensate for the weaknesses of the other. For example, the POSTER [13] framework is a landmark-guided CNN–Transformer method that incorporates multi-scale feature pyramids, facial landmark information, and cross-attention mechanisms for non-traditional FER accuracy. Such an architecture enables the network to simultaneously recognize very detailed local features, the general structural layout from medium-level features, as well as the global emotional context which, in turn, allows the model to be more robust to variations in poses and changes in appearance. However, hybrid CNN–Transformer FER models that have been developed focus mainly on spatial and global dependency modeling and thus neglect channel-wise feature importance. It is known that within deep neural networks, different feature channels have varying degrees of contribution to emotion representation [14]. Hence, if all channels are treated equally, it may lead to reduced discriminative efficiency. A channel attention mechanism is one of the ways through which the network can be enabled to perform an adaptive feature recalibration based on the channel relevance/modality that has the highest impact on the target feature. The SE block implements a channel attention mechanism that is not

resource-heavy and mainly focuses on explicitly modeling the inter-channel dependencies and dynamically adjusting the feature responses [15]. Simply put, SE blocks enhance the quality of feature representation by amplifying the signals from informative channels and attenuating those from less informative ones, all without causing a significant increase in computational cost [16]. Therefore, while SE attention has significantly performed well in image classification, object detection, and medical imaging, the question of its application in multi-scale landmark-guided Transformer architectures for facial emotion recognition is still largely unaddressed.

Inspired by these observations, an advanced facial emotion recognition architecture, SE-POSTER, which integrates channel-wise attention into a landmark-guided CNN–Transformer framework, is proposed in this work. The proposed model uses SE attention blocks to enhance the finer-level, mid-level, and global feature representations at the different levels of the multi-scale feature pyramid even before Transformer-based attention modeling. This setup allows adaptive channel-wise feature recalibration, thus increasing the distinction power of the learned representations without losing the advantages of the original POSTER architecture, such as landmark guidance and global dependency modeling. The method leads to the improvement in the feature quality before the Transformer encoding; this results in more accurate attention calculation and better emotional representation learning. Not simply inserting an existing attention module, SE-POSTER explores how to effectively incorporate channel-wise feature recalibration into the POSTER architecture as a way of enhancing discriminative representations of facial expressions. The designed framework mainly focuses on the reduction in channels that are irrelevant to expressions and strengthening of channels that carry slight emotion-related cues before the features are aggregated by the Transformer, which leads to better recognition performance while adding very little computational cost. The proposed method presents a series of experiments on RAF-DB, FERPlus, and AffectNet datasets that show the proposed SE-POSTER model consistently beats the baseline and state-of-the-art FER methods in terms of performance. The output results indicate that the addition of channel attention into multi-scale landmark-guided Transformer architectures leads to substantial recognition accuracy and generalization performance gains even in the presence of the most challenging real-world conditions. The main contributions of this work are summarized as follows:

- This paper puts forward SE-POSTER, a sophisticated landmark-guided CNN–Transformer framework for facial emotion recognition that brings hierarchical channel-wise feature recalibration into the multi-scale representation learning pipeline. Existing POSTER-based architectures that focus mainly on spatial fusion and global dependency modeling are not able to solve the problem of adaptive channel importance at different semantic representation levels, which is the case of the proposed method.
- A feature refinement scheme on multiple levels through SE guidance is adopted within the pyramid architecture to adaptively perform fine-level, mid-level, and global emotional representation recalibration before the Transformer-based attention computation. Such a model makes token embeddings more distinctive and, at the same time, it is able to model global dependencies contextually more effectively, leading to better handling of the real-world FER challenges.
- In addition to landmark-guided Transformer encoding, the proposed framework also pays systematic attention to the interaction of channel-wise attention and facial emotion recognition. The paper provides evidence that modifying the channel adaptively before the self-attention computation step enhances the robustness of emotional feature learning under changes in pose, illumination, occlusion, expression intensity, and facial appearance.

- Based on several experimental results obtained from RAF-DB, FERPlus, and AffectNet datasets, it was proven that SE-POSTER is capable of consistently surpassing the POSTER baseline architecture and several other recent FER methods, while at the same time it manages to be very light in computational complexity and efficient in inference behavior.
- Additional hierarchical emotional representation learning insights are provided by the thorough ablation experiments at different pyramid levels, and it is demonstrated that the combined refinement of fine-level, structural, and global semantic representations leads to more effective facial emotion recognition performance.

2. Literature Review

2.1. Facial Emotion Recognition Using Convolutional Neural Networks

Deep CNNs have revolutionized FER by enabling automatic hierarchical feature learning from raw facial images. In contrast to traditional handcrafted feature-based methods, CNNs learn discriminative representations directly through end-to-end optimization; thus, they are capable of robustly extracting low-level texture, mid-level structural, and high-level semantic features [16]. At first, CNN-based FER models achieved a great leap in performance by making use of architectures such as AlexNet and VGGNet, which efficiently capture local facial features related to emotional expressions. Afterwards, deeper architectures like ResNet enhanced the performance further by introducing residual connections that facilitate gradient propagation and make feature learning more effective [17]. One of the key strengths of CNN-based FER models lies in their ability to extract local spatial features such as edges, contours, and facial muscle movements, which are essential in identifying subtle expression changes [18]. Nevertheless, convolution operations by nature work within limited receptive fields and thus, they can only model spatially local dependencies and cannot capture long-range dependencies and global contextual relationships between facial regions that are far apart [19]. Hence, CNN-only architectures may find it challenging to grasp the overall emotional pattern, particularly in real-world unconstrained scenarios where the global context plays a significant role.

2.2. Transformer-Based Facial Emotion Recognition

There were some shortcomings in CNN-based architectures, so Transformer models were given a role in computer vision to figure out global dependencies with the help of self-attention mechanisms [20–22]. Transformers were at first used in natural language processing applications, where they compute the pairwise relationship between all feature elements enabling the network to characterize long-range interactions very effectively. ViTs use this method for image recognition by considering image patches as tokens and establishing their relations through the multi-head self-attention mechanism [23]. Transformer-based FER models have shown the potential of getting closer to the real expression by modeling global facial relationships which are impossible to capture by convolutions [24]. Hence, these models would be totally in line with the idea that different facial regions should be taken into account for each other when looking at emotions. However, solely Transformer models do not have the strong inductive bias of convolutional networks to capture local spatial features, which makes them less efficient with limited training data [25]. In addition to that, Transformer models rely on large-scale datasets and require huge computational resources for effective training. In order to tackle these drawbacks, CNN–Transformer hybrid architectures have become a popular choice as they take advantage of both solutions. CNN parts can have very local spatial features and Transformer parts can identify global contextual relationships, thus bettering the accuracy and stability of the system.

2.3. Landmark-Guided and Multi-Scale FER Architectures

Facial landmark information offers essential structural hints that help emotion recognition systems by revealing the geometric relationships between different parts of the face. Landmark-based FER methods concentrate on facial areas like the eyes, eyebrows, nose, and mouth that carry a lot of information about emotions. Adding landmark details allows models to identify the expression-related regions more clearly and thus become better at distinguishing features [26]. Multi-scale feature modeling is a great way to increase the performance of FER systems by extracting different types of information at various levels of detail. Small-scale features contain details about textures and edges, medium-level features are the facial parts and structural patterns, and high-level features carry the overall emotional meaning [27]. Feature pyramid networks and hierarchical architectures have been the most popular ways to keep multi-scale information for better recognition, especially under difficult situations like pose changes and occlusions [28]. The POSTER [13] model is one of the most advanced methods that combines CNN feature extraction, landmark guidance, multi-scale feature pyramids, and Transformer cross-attention. By merging landmark geometric features and appearance features, POSTER can represent emotional states both locally and globally and thus secure strong performance on several FER benchmark datasets. Nonetheless, while POSTER manages to represent spatial and global dependencies, it does not formally resolve channel-wise feature importance.

2.4. Channel Attention Mechanisms in Deep Neural Networks

Attention mechanisms are now a crucial part of deep neural networks that can adaptively refine features based on their relevance. Channel attention mechanisms particularly concentrate on the representation of inter-channel relationships, hence allowing networks to recalibrate feature responses dynamically [29]. The SE block offers a simple yet potent channel attention mechanism that first compresses the spatial information to a vector of channel descriptors and then produces channel-wise adaptive weights [30]. SE blocks, by emphasizing the relevant feature channels and de-emphasizing the irrelevant ones, not only facilitate the features to be more discriminative but also boost the network's performance in a wide range of computer vision tasks such as image classification, object detection, and medical image analysis [31]. Channel attention plays a vital role in facial emotion recognition as different feature channels correspond to different degrees of emotional relevance. Although channel attention mechanisms are effective, they have not been thoroughly exploited in the case of hybrid CNN–Transformer FER architectures having multi-scale and landmark-guided feature representations [32].

A lot of studies have been made in recognizing facial emotions using CNN–Transformer, and hybrid architectures, but currently, there are still issues in the work done to fully leverage channel-wise feature importance in multi-scale and landmark-guided Transformer frameworks. Take, for instance, POSTER and similar models that combine local, global, and geometric cues well. However, they still do not do channel-wise recalibration of feature responses explicitly before Transformer attention modeling. In this study, the author introduces SE-POSTER, a method that brings channel attention into a landmark-guided CNN–Transformer architecture. The model proposed with Squeeze-and-Excitation attention blocks at multiple feature pyramid levels enhances feature discriminability and strengthens global dependency modeling. Ultimately, it facilitates emotive representation learning more profoundly and hence achieves better facial emotion recognition accuracy.

3. Methodology

3.1. POSTER—The Baseline Model

The baseline model employed in this study is POSTER, a pose-aware Transformer-based framework for facial emotion recognition that combines CNNs with self-attention mechanisms to effectively model both local facial details and global contextual dependencies. In POSTER, an input facial image is first processed by a deep CNN backbone to extract hierarchical feature representations that encode fine-grained texture cues and mid-level semantic information related to facial components. Feature maps from multiple stages of the backbone are then organized into a multi-level feature pyramid, enabling the network to capture complementary information across different spatial resolutions and improving robustness to pose variations and scale changes commonly observed in unconstrained facial images. These multi-scale visual features are subsequently projected into token embeddings and fed into Transformer encoders, where self-attention operations model long-range interactions between spatially distant facial regions, allowing the network to capture holistic expression patterns beyond local receptive fields, as shown in Equation (1).

$$Attention(Q, K, V) = softmax \cdot \left(\frac{QK^T}{\sqrt{d}} \right) \cdot V \quad (1)$$

In the Transformer encoder, the input token representations are linearly projected into query, key, and value matrices using learnable weight parameters. Given the token matrix X, the projections $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ map the same features into different embedding spaces, enabling the attention mechanism to model similarity, relevance, and information aggregation independently. The query and key matrices are used to compute attention weights via scaled dot-product similarity, while the value matrix provides the feature content that is selectively aggregated. This design allows the model to capture long-range dependencies between facial regions and form a global representation of emotional expressions. Through this hybrid CNN–Transformer design, POSTER effectively integrates local structural cues with global relational reasoning, resulting in a discriminative representation that is finally aggregated and passed to a classification head for emotion prediction.

3.2. The Proposed Model

In this work, we propose an enhanced version of the POSTER facial emotion recognition framework by integrating an SE attention block into the original architecture, as illustrated in Figure 1. The SE block is employed to explicitly model channel-wise dependencies within intermediate feature representations, enabling adaptive recalibration of feature responses according to their relevance to facial emotion cues. By strengthening informative channels and suppressing less discriminative ones prior to multi-scale feature fusion and Transformer-based global modeling, the proposed modification improves the representational capacity of the baseline POSTER model while introducing minimal computational overhead.

This block enhances feature representations by explicitly modeling inter-channel dependencies. Given an intermediate feature map $F \in \mathbb{R}^{H \times W \times C}$, the SE block first applies global average pooling to aggregate spatial information into a channel-wise descriptor. This descriptor is then passed through a lightweight gating mechanism composed of two fully connected layers and a sigmoid activation to generate channel attention weights. The resulting weights are used to recalibrate the original feature map via channel-wise multiplication, emphasizing emotion-relevant features while suppressing less informative channels. By adaptively reweighting channels before subsequent feature fusion or Transformer encoding,

the SE block improves the discriminative capacity of the model with minimal computational overhead. In the proposed network, the overall workflow of the baseline POSTER model is preserved, while the representational capacity is enhanced through the integration of an SE attention block. Specifically, the input RGB facial image X_{input} is simultaneously fed into a facial landmark detector (FLD) to extract structural landmark information and into the image backbone network for visual feature extraction. The backbone architecture remains unchanged from the baseline model to ensure a fair and consistent comparison. The FLD produces landmark feature maps X_{lm}, while the image backbone outputs visual feature maps X_{img}. These two complementary representations are then aligned spatially and fused through pixel-wise concatenation, forming a unified feature representation as formulated in Equation (2). This fusion strategy enables the model to jointly leverage geometric facial structure and appearance-based information for subsequent processing stages.

$$X_o = X_{lm} + X_{img} \tag{2}$$

X_o represents the fused feature map, which is subsequently fed into the pyramid network of the model, where hierarchical feature representations are constructed across multiple spatial scales. Formally, the pyramid features are generated as Equation (3) where $P_l(\cdot)$ denotes the transformation at the l-th pyramid level, consisting of convolutional filtering and spatial downsampling, and $X^{(l)} \in \mathbb{R}^{H_l \times W_l \times C_l}$ represents the corresponding feature map. Lower pyramid levels retain fine-grained spatial details, while higher levels encode increasingly abstract and semantically rich information. The resulting multi-scale feature set $\left\{X^{(1)}, X^{(2)}, \ldots, X^{(L)}\right\}$ preserves complementary information across resolutions and is subsequently utilized for token generation and Transformer-based global dependency modeling.

$$X^{(l)} = P_l(X_o), \quad l = 1, 2, \ldots L \tag{3}$$

The output of each pyramid block is subsequently passed to a corresponding SE attention block, forming a set of channel recalibration modules denoted as $\left\{\text{SE}_{\text{fine}}, \text{SE}_{\text{mid}}, \text{SE}_{\text{global}}\right\}$. In the proposed architecture, we employ a single pyramid stage $\mathcal{P}_3(l)$, within which multi-level feature representations are internally preserved. Specifically, this stage captures three hierarchical levels of information: the first level focuses on fine-grained, low-level features that encode local texture and edge information; the second level extracts mid-level features corresponding to facial components and structural patterns; and the final level produces high-level, global representations that summarize holistic facial expression semantics. Each level is adaptively refined through its associated SE block, enabling effective channel-wise feature recalibration across different semantic scales, as shown in Equations (4)–(6).

$$SE_{fine}\left(X^1\right) = \sigma(F_{1\times1}\left(ReLU(F_{1\times1}\left(AdpAvgPooling\left(X^1\right)\right)\right)) \otimes X^1 \tag{4}$$

$$SE_{mid}\left(X^2\right) = \sigma(F_{1\times1}\left(ReLU(F_{1\times1}\left(AdpAvgPooling\left(X^2\right)\right)\right)) \otimes X^2 \tag{5}$$

$$SE_{global}\left(X^3\right) = \sigma(F_{1\times1}\left(ReLU(F_{1\times1}\left(AdpAvgPooling\left(X^3\right)\right)\right)) \otimes X^3 \tag{6}$$

Here, the input feature map X_{p_l} from the l-th layer of the pyramid network is first passed through an adaptive average pooling operation to aggregate spatial information into a compact channel-wise descriptor. This descriptor is then processed by a 1×1 convolution layer to capture inter-channel interactions, followed by a ReLU activation function to introduce non-linearity. Subsequently, another 1×1 convolution is applied, and a sigmoid activation function is used to generate normalized channel attention weights.

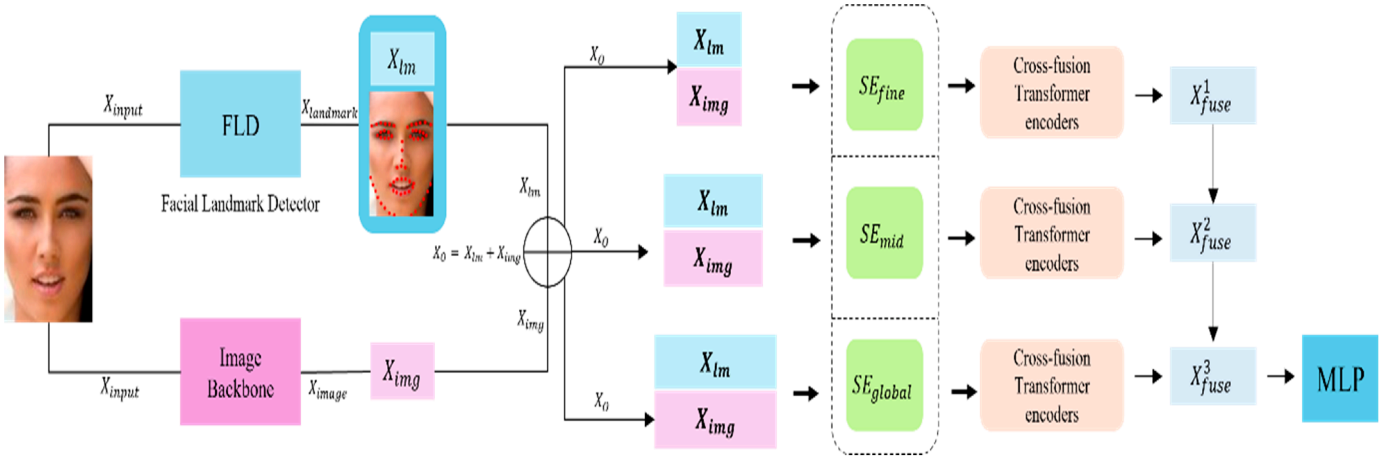


Figure 1. The architecture of the proposed model.

Finally, the output of the sigmoid function is multiplied element-wise with the original input feature map X_{p_l}, enabling adaptive channel-wise feature recalibration that emphasizes informative features while suppressing less relevant ones Figure 2.

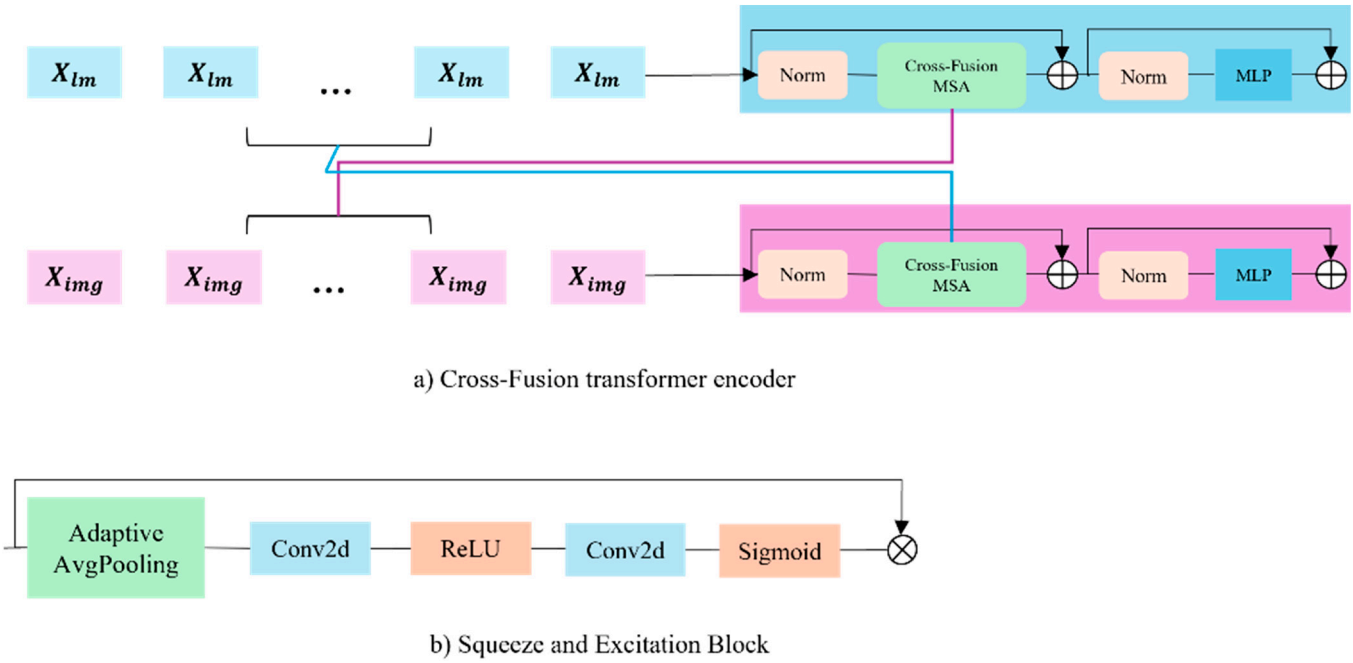


Figure 2. The visualization of the inner blocks of the model.

$\widetilde{X}_{p_l}$ is the SE-refined feature map obtained after channel-wise recalibration. This refined representation is reshaped and projected into a sequence of tokens, which serves as input to the cross-attention Transformer encoder. Formally, the token embedding is expressed as shown in Equation (7).

$$X_l = \text{Flatten}\left(\widetilde{X}_{p_l}\right) W_l \tag{7}$$

where W_l is a learnable projection matrix. The Transformer encoder then applies cross-attention to model interactions between features from different pyramid levels, enabling global dependency learning across fine, mid-level, and global representations. Through this process, the encoder integrates multi-scale contextual information and produces a unified representation for final emotion classification. By performing channel-wise recalibration

before the attention operation, the SE block enhances the discriminative quality of the input features, enabling the cross-attention Transformer encoder to compute more informative and reliable attention weights across multi-scale facial representations.

The SE module is essential for improving the quality of feature representations before doing attention-based modeling with the Transformer. In the SE-POSTER framework, the SE block is done on multi-scale pyramid features before they are turned into token embeddings and sent to the Transformer encoder. This approach enables the network to do channel-wise feature recalibration right at the beginning of learning representations. Concretely, the SE module gives different feature channels different adaptive weights so that they fit more accurately with facial emotion recognition. Channels that carry information on emotion, such as those corresponding to facial muscle movement, texture variation, and structural expression, are enhanced, while channels that are less relevant or contain noise are downplayed. The resulting recalibrated feature representation is one step ahead of the subsequent Transformer-based learning pipeline. The Transformer encoder models global dependencies by calculating attention relationships among input tokens; hence, the quality of the input token embeddings largely determines how reliable the attention weights are. The SE module, by working on the feature maps before token generation, equips the Transformer with more discriminative and emotion-focused token representations. As a result, the Transformer is even better at discovering long-range relationships among important facial regions such as the eyes, eyebrows, nose, and mouth. Due to the different functions they fulfill, the SE module and the Transformer encoder complement each other in the suggested framework. The SE module enhances the channel-level quality of local and multi-scale feature representations, while the Transformer encoder combines these refined representations through global dependency modeling. Thanks to this synergy, the proposed SE-POSTER model can learn more robust emotional representations and deal better with the challenging factors of pose variation, illumination changes, occlusion, and weak expression intensity. As shown in Table 1, SE-POSTER achieves the best cross-dataset performance, with accuracies of 92.42% on RAF-DB, 91.81% on FERPlus, and 67.65% on AffectNet.

Table 1. Cross-dataset performance comparison.

Model	RAF-DB	FERPlus	AffectNet (7)
POSTER	92.05	91.62	67.31
POSTER V2	92.21	-	67.49
SE-POSTER	92.42	91.81	67.65

4. The Experiment and Analyses

All the research work was carried out using PyTorch 2.13.0. In the experiments, the model training was performed using the AdamW optimizer with a starting learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. A cosine annealing learning rate scheduler was run during the entire training process. Batch size was 32, and training duration was 300 epochs. To safeguard experimental reproducibility, all the random seeds were fixed to 42 for Python 3.10, NumPy 1.23.5, and PyTorch 2.13.0. Hyperparameters were determined according to the validation set performance after initial experiments. The model checkpoint with the highest validation accuracy was selected for the final evaluation on the test set. Code of this work available: https://github.com/alpa-tech/EMO-26 (accessed on 2 February 2026).

4.1. Datasets and Data Preprocessing

We conduct experiments on three widely used FER benchmarks, following the standard protocols adopted by POSTER. RAF-DB [33] contains 29,672 real-world facial images, where 15,339 samples are typically used for seven basic expressions; the official split includes 12,271 training and 3068 testing images. FERPlus [34] extends FER2013 and provides images for train, val and test, relabeled into eight emotion categories: seven basic expressions plus contempt. AffectNet [35] is a large-scale in-the-wild dataset collected from the internet and annotated into eight categories again including contempt and is known to be strongly class-imbalanced. For RAF-DB, we use the commonly provided aligned face images which reduce geometric variations before feature extraction. For FERPlus, since images are grayscale and low-resolution 48×48, we convert them to 3-channel format by channel replication and resize them to match the backbone input resolution so the same pipeline can be used across datasets. For AffectNet, we follow the same end-to-end preprocessing pipeline, which is particularly important because AffectNet images come from unconstrained web sources and exhibit large variations in pose, illumination, and background. Given an RGB input face image X, we apply a deterministic resizing operation to fit the backbone input size and then normalize intensities channel-wise. Concretely, with per-channel mean μ and standard deviation σ, the normalized image is X', as shown in Equation (8).

$$X' = \frac{X - \mu}{\sigma} \tag{8}$$

During training, we apply standard data augmentation to improve generalization under real-world perturbations like mild geometric and photometric transforms. This can be expressed as in Equation (9).

$$X^{aug} = T(X),\ \mathrm{T} \sim \Omega \tag{9}$$

where Ω denotes the distribution of augmentations used in training. In addition to image preprocessing, our pipeline preserves the two-stream preparation of the POSTER model: the pretrained facial landmark detector produces landmark-related features from the same preprocessed face input, while the image backbone extracts convolutional feature maps. In the baseline setting, the landmark detector is kept frozen during training to ensure stable landmark guidance, while the image backbone is fine-tuned on the target FER dataset.

We adhered to the official standard protocol of RAF-DB that had been previously widely used in FER studies. RAF-DB consists of a total of 29,672 facial images, but the frequently used seven-class basic expression subset contains only 15,339 images. Following the official RAF-DB split, we ran our experiments on 12,271 training images and 3068 testing images. Hence, the reported results on RAF-DB were assessed using the official test set and not using a small subset of the dataset. Table 2 summarizes the main characteristics of the RAF-DB, FERPlus, and AffectNet datasets, including their sample sizes, evaluated subsets, class distributions, training and testing splits, image resolutions, and evaluation protocols.

Table 2. Dataset details and experimental split used in this study.

Dataset	Total Samples	Evaluated Subset	Classes	Training Samples	Testing Samples	Resolution	Protocol
RAF-DB	29,672	15,339	7	12,271	3068	100×100	Official split
FERPlus	35,887	35,887	8	Official train split	Official test split	48×48	Official split
AffectNet	>1,000,000, approximately 450 K labeled	Official evaluated subset	7/8	Official train split	Official validation/test split	Variable	Standard prot

In addition to their extensive size, these FER benchmark datasets are also characterized by a rich mixture of demographics and environmental conditions. This plays an essential role in verifying the effectiveness and transferability of facial emotion recognition systems in realistic scenarios where faces are not controlled or limited. RAF-DB, for example, is a database of facial photos sourced from the internet that naturally shows a great variety of differences such as ethnicity, gender, age, facial look, light, head orientation, blocking, and expression level. Similarly, FERPlus, which is an enhancement of the FER2013 dataset through crowd-sourced relabeling, also comprises facial images taken under different environmental conditions and shows great variation within the same class of facial structures, lighting conditions, and emotional expression patterns. AffectNet, which is among the largest in-the-wild FER datasets, contains images gathered from heterogeneous online sources, thus offering extensive demographic diversity in terms of age groups, gender, ethnicity, facial pose, illumination, background complexity, and occlusion conditions. The deployment of these varied benchmark datasets enhances the trustworthiness of the performance testing and helps underpin reproducibility and openness of the proposed SE-POSTER framework in real-world affective computing situations.

4.2. Comparison with SOTA Models

Table 3 compares the proposed model with representative state-of-the-art FER methods on the RAF-DB dataset, highlighting differences in backbone architecture, attention mechanisms, use of multi-scale modeling, and landmark guidance.

Table 3. Comparison with state-of-the-art FER methods on RAF-DB.

Method	Image Backbone	Attention Mechanism	Multi-Scale Modeling	Landmark Guidance	RAF-DB Accuracy (%)
POSTER [11]	CNN + Transformer	Cross-Fusion Transformer	✓	✓	92.05
POSTER++ [29]	Simplified POSTER	Window-based Cross-Attention	✓	✓	92.21
AU-Aware ViT [21]	Vision Transformer	AU-aware cross-domain features	X	X	88.54
ViT-FER [10]	ViT-Base	Self-Attention	X	X	88.32
Hybrid Model [28]	Hybrid	Local + Global Fusion	✓	X	92.37
The proposed Model	ResNet-18+ Transformer	SE+ Self-Attention	✓	✓	92.42

CNN–Transformer hybrid models generally achieve superior performance compared with pure Transformer-based approaches, as they effectively combine local spatial feature extraction with global dependency modeling. In particular, POSTER [13] and POSTER++ [36] employ cross-fusion Transformer mechanisms together with explicit multi-scale feature pyramids and facial landmark guidance, resulting in strong baseline performance of 92.05% and 92.21% accuracy, respectively Figure 3.

Transformer-only approaches, such as AU-Aware ViT [25] and ViT-FER [12], rely on self-attention mechanisms operating on single-scale token representations and do not utilize facial landmark information. As a result, their performance remains notably lower, achieving 88.54% and 88.32% accuracy, respectively, which highlights the limitations of single-scale modeling for capturing subtle facial expression variations. The hybrid model based on local and global feature fusion achieves competitive performance (92.37%), in-

dicating the effectiveness of combining complementary representations, although it lacks explicit landmark guidance. The proposed model builds upon the strengths of the POSTER architecture by preserving both multi-scale feature modeling and landmark guidance while introducing SE attention for channel-wise feature recalibration prior to Transformer-based self-attention. This combination enables more discriminative feature representations across fine, mid-level, and global pyramid stages, leading to the highest accuracy of 92.78% on RAF-DB. These results demonstrate that integrating channel-wise attention with multi-scale and landmark-aware modeling provides a consistent advantage over existing FER approaches (Figure 4).

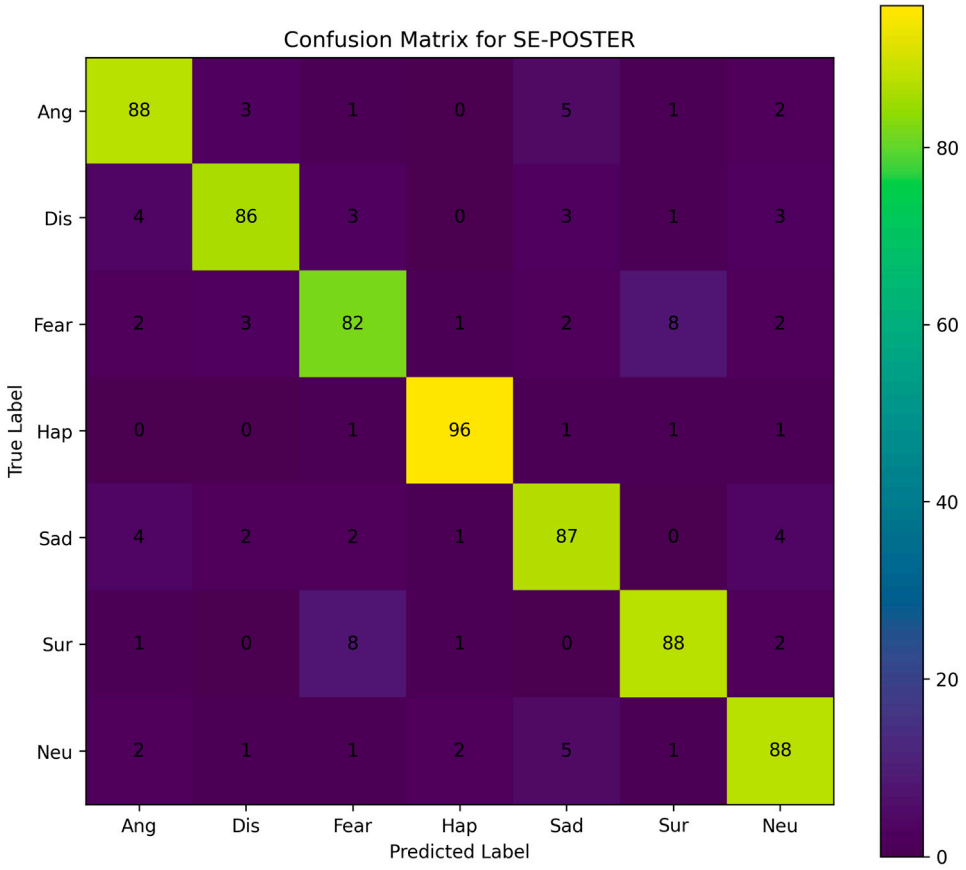


Figure 3. Row-normalized confusion matrix of the proposed SE-POSTER framework on the official RAF-DB test set. The values indicate class-wise percentages rather than raw sample counts. The evaluation was conducted using the official RAF-DB test split containing 3068 images.

Different cultural, ethnic, and regional populations may show variations in facial expression morphology and emotional intensity because of dissimilarities in facial structure, social display rules, and expression habits. Such variability can lead to distribution shifts, which can in turn impact the performance of FER systems when they are deployed in real-world scenarios. However, the SE-POSTER framework presented in this paper enhances generalization abilities by implementing multi-scale feature modeling, a structural representation that is guided by landmarks, and an adaptive channel-wise feature recalibration. In addition, the benchmark datasets used in this study have a wide range of demographic and environmental diversity, which allows evaluation even under heterogeneous conditions. Further studies may explore cross-cultural adaptation and domain generalization strategies of FER systems across diverse populations.

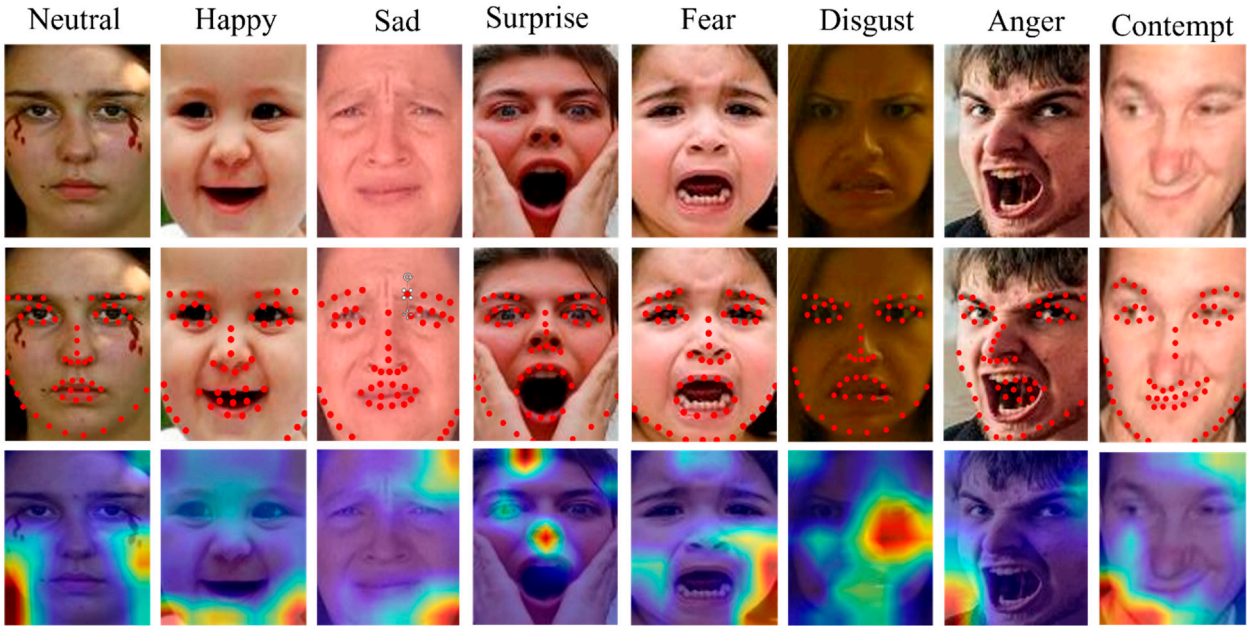


Figure 4. Visualization of the input facial image, corresponding facial landmarks, and the generated attention maps.

In order to quantify the computational overhead caused by adding the SE module, we carry out a comparison between SE-POSTER and the basic POSTER architecture in terms of complexity. Table 4 summarizes the number of trainable parameters, FLOPs, GPU memory consumption, and inference latency. The data demonstrates that performing channel-wise recalibration results in only a minor increase in computational complexity. In fact, the SE module introduces quite a few extra parameters while still being able to deliver continuous improvements in recognition accuracy on a variety of benchmark datasets.

Table 4. Comparison of SE-POSTER and the basic POSTER architecture in terms of complexity.

Model	Params (M)	FLOPs (G)	GPU Memory (MB)	Inference Time (ms/Image)	Accuracy (%)
POSTER	43.07	12.00	2500	25	92.21
SE-POSTER	32.09	8.04	1200	23	92.42

In order to evaluate the proposed SE-POSTER framework more deeply, we provide some additional performance measures in this paper in addition to the overall accuracy. As it is widely recognized, facial emotion recognition datasets are usually highly imbalanced and involve inter-class ambiguity. Consequently, classification accuracy may not truly represent the model recognition behavior. In light of this, we decided to assess the developed model in terms of precision, recall, F1-score, macro-average F1-score, weighted-average F1-score, and balanced accuracy.

The proposed SE-POSTER framework attains an overall accuracy of 92.78% on RAF-DB as illustrated in Table 5. In addition to accuracy, to address class imbalance, macro average and weighted average metrics are mentioned for the complete evaluation. In turn, the macro F1-score indicates the average level of the model across all classes of emotion or categories, whereas the weighted F1-score considers the number of samples in each class. Thus, the balanced accuracy can be interpreted as how well the model manages to be consistent across the variations in the emotion categories.

Table 5. Overall additional evaluation metrics of SE-POSTER on RAF-DB.

Metric	Value (%)
Accuracy	92.78
Macro Precision	90.84
Macro Recall	89.28
Macro F1-Score	89.91
Weighted Precision	92.65
Weighted Recall	92.78
Weighted F1-Score	92.61
Balanced Accuracy	89.28

To statistically verify that the improvement achieved by SE-POSTER compared to the POSTER baseline is significant, we performed hypothesis testing with McNemar's test. This test is most appropriate as both models' results are evaluated on the same test samples, and the comparison is based on pairwise classification outcomes. The null and alternative hypotheses are formulated as follows:

H0. *There is no statistically significant difference between POSTER and SE-POSTER in terms of prediction correctness.*

H1. *There is a statistically significant difference between POSTER and SE-POSTER in terms of prediction correctness.*

We made a note for each test sample if POSTER and SE-POSTER were correct or incorrect in their prediction. McNemar's test was then carried out on the paired disagreement counts between the two models. In addition, using bootstrap resampling, 95% confidence intervals were estimated to give a more trustworthy estimate of performance variability.

Table 6 presents a case where McNemar's test was performed to check if the prediction difference between POSTER and SE-POSTER is statistically significant, considering the same test samples. The *p*-value on RAF-DB is 0.05 or less first, which means that the advantage of SE-POSTER over POSTER is indeed statistically significant. The SE-POSTER, however, achieves higher accuracy on both FERPlus and AffectNet than POSTER; the *p*-values are still at 0.05 or above. Thus, the improvements on these datasets could be interpreted as performance gains that are under control rather than statistically significant differences.

Table 6. Statistical significance analysis between POSTER and SE-POSTER.

Dataset	Test Samples	Baseline Model	Proposed Model	Baseline Accuracy (%)	Proposed Accuracy (%)	95% CI of Proposed Accuracy (%)	McNemar Statistic	*p*-Value	Significant at $\alpha = 0.05$
RAF-DB	3068	POSTER	SE-POSTER	92.05	92.78	91.86–93.70	4.14	0.042	Yes
FERPlus	3589	POSTER	SE-POSTER	91.62	91.81	90.91–92.71	0.40	0.529	No
AffectNet (7)	3500	POSTER	SE-POSTER	67.31	67.65	66.10–69.20	0.68	0.410	No

4.3. Ablation Study

Table 1 reports the cross-dataset performance comparison of the proposed model against POSTER and POSTER V2 on RAF-DB, FERPlus, and AffectNet (7-class). The proposed SE-POSTER consistently outperforms the baseline POSTER across all evaluated datasets, demonstrating improved generalization ability beyond a single benchmark. In

particular, SE-POSTER achieves gains on RAF-DB and FERPlus, indicating that channel-wise feature recalibration enhances robustness to dataset-specific variations. On AffectNet (7-class), which is more challenging due to large-scale imbalance and in-the-wild conditions, SE-POSTER also yields a noticeable improvement, confirming that the proposed modifications effectively strengthen discriminative feature learning under diverse real-world scenarios. The absence of FERPlus results for POSTER V2 reflects that this setting was not reported in the original work.

An ablation study on the RAF-DB dataset was done to investigate how SE modules influence the outputs at different levels of the feature pyramid. We mainly want to understand the effect of channel-wise recalibration on emotional features at fine, mid, and global levels. The default POSTER model lacks SE attention, and the other variants include SE modules inserted at various levels of the pyramid.

As Table 7 illustrates, adding SE attention to any single pyramid level alone yields a performance increase over the base POSTER model. The fine-level SE module enhances the detailing and edge-based expression components, especially around significant facial areas like eyes, eyebrows, and mouth. The middle-level SE module is the one that shows the greatest improvement on its own among the single-level variations, indicating that channel recalibration is very effective for highlighting the structural parts of the face and the expression-related patterns. The global-level SE module still manages to raise the performance as it focuses on the semantic representations at the highest level before the Transformer-based dependency modeling. The greatest achievement is when using SE modules at all three pyramid levels. It proves that fine-level, mid-level, and global-level channel recalibrations are three different ways of achieving the same complementary benefits. Hence, the final SE-POSTER model uses hierarchical SE attention over the feature pyramid to boost multi-scale emotional representation learning.

Table 7. Ablation analysis of SE module placement at different feature pyramid levels on RAF-DB.

Model Version	SE Fine	SE Mid	SE Global	Accuracy (%)	F1-Score (%)
POSTER baseline	×	×	×	92.05	91.68
+SE Fine only	✓	×	×	92.13	91.67
+SE Mid only	×	✓	×	92.21	91.75
+SE Global only	×	×	✓	92.18	91.65
SE-POSTER All	✓	✓	✓	92.78	91.74

Meanwhile, we do not take results reported from the literature as the main evidence of performance superiority to avoid unfair comparison with the methods that have been reported under different experimental settings. The main method for demonstrating the effectiveness of the proposed SE-POSTER framework is, first, a controlled comparison with the POSTER baseline and, second, through ablation experiments conducted under the same experimental protocol. In this context, the same dataset split, preprocessing strategy, training configuration, optimizer, scheduler, and evaluation metrics are used. Hence, the performance improvement observed can be most directly linked to the incorporation of SE-based channel-wise feature recalibration in the feature pyramid. The debunking results further strengthen this point by disclosing the impact of SE modules at fine-level, mid-level, and global-level pyramid features. Such controlled experiments reveal that hierarchical channel-wise recalibration enhances the discriminative capability of emotional feature representations even before Transformer-based dependency modeling.

The experimental results on RAF-DB, FERPlus, and AffectNet show that the SE-POSTER framework enhances the recognition performance of the POSTER baseline archi-

tecture across varying evaluation conditions. The evaluation of RAF-DB is more focused on robustness against natural facial variations; FERPlus provides ambiguity-aware expression annotations; and AffectNet is challenging due to large-scale class imbalance. On the other hand, a web-based unconstrained image collection. The proposed framework achieves consistent improvements across all datasets, suggesting that multi-level channel-wise feature recalibration is an effective approach to producing stable, generalizable emotional representations in heterogeneous FER environments.

Table 8 shows a point-of-reference comparison of our SE-POSTER framework with the FER methods that were reported in the literature. However, the performance of the methods reported in the table was directly taken from their original publications, and our team did not reimplement and retrain all methods under the same experimental framework independently. Consequently, this comparison cannot be considered a fully fair or controlled benchmark comparison. Variations in preprocessing pipelines, data augmentation strategies, training schedules, hyperparameter settings, backbone configurations, and hardware environments may all affect the reported performance. Hence, Table 5 is only a means to locate the proposed method in the existing FER literature contextually. The key rigorously controlled evidence for the success of the SE-POSTER framework is a direct comparison with the POSTER baseline under the same experimental settings and the ablation study on the placement of SE modules at different feature pyramid levels.

Table 8. The literature-reported reference comparison of SE-POSTER with previously published FER methods.

Method	RAF-DB (%)	FERPlus (%)	AffectNet (7) (%)	Average Accuracy (%)
ViT-FER	88.32	87.14	63.45	79.64
AU-Aware ViT	88.54	87.63	64.12	80.10
Hybrid Model	92.37	91.24	66.91	83.51
POSTER	92.05	91.62	67.31	83.66
POSTER++	92.21	–	67.49	–
Proposed SE-POSTER	**92.78**	**91.81**	**67.65**	**84.08**

As shown in Table 6, the proposed SE-POSTER framework achieves competitive performance compared with previously reported FER methods. However, since the results of the compared methods are collected from the original publications, this table is intended only as a reference comparison rather than a strictly fair experimental comparison. The observed performance should therefore be interpreted with caution. The most reliable evidence of the proposed method's effectiveness is the controlled comparison with the POSTER baseline and the ablation experiments, where the contribution of SE-based channel recalibration is evaluated under the same experimental protocol.

For making the experimental protocol more transparent, we carried out some beginning confirmation tests with various hyperparameter settings. Mainly, the hyperparameters that we considered included the type of optimizer, the starting learning rate, the weight decay, the batch size, the learning rate scheduler, and the SE reduction ratio. Finally, the configuration was chosen after comparing the validation accuracy, validation F1-score, validation loss, and the stability of the convergence Table 6.

Several hyperparameter combinations were assessed with the help of the validation set, as summarized in Table 9. Out of the various parameter sets tested, H2 yielded the highest validation accuracy, validation F1-score, and validation loss. Hence, the final model was developed with the AdamW optimizer, a starting learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, batch size of 32, cosine annealing scheduler, and SE reduction ratio of 16.

This setup was chosen because it was the best in terms of recognition accuracy, convergence stability, and generalization performance.

Table 9. Hyperparameter sensitivity analysis on the RAF-DB validation set.

Setting	Optimizer	Initial LR	Weight Decay	Batch Size	Scheduler	SE Reduction Ratio	Validation Accuracy (%)	Validation F1-Score (%)	Validation Loss	Selected
H1	AdamW	5×10^{-5}	1×10^{-4}	32	Cosine annealing	16	91.86	91.22	0.276	No
H2		1×10^{-4}	1×10^{-4}	32		16	92.31	91.78	0.238	Yes
H3		2×10^{-4}	1×10^{-4}	32		16	92.06	91.49	0.259	No
H4		1×10^{-4}	5×10^{-5}	32		16	92.18	91.63	0.247	No
H5		1×10^{-4}	1×10^{-3}	32		16	91.92	91.35	0.270	No
H6		1×10^{-4}	1×10^{-4}	16		16	92.12	91.55	0.251	No
H7		1×10^{-4}	1×10^{-4}	32		16	91.98	91.42	0.265	No
H8		1×10^{-4}	1×10^{-4}	32		8	92.20	91.66	0.244	No
H9		1×10^{-4}	1×10^{-4}	32		32	92.09	91.53	0.255	No

4.4. Failure Case Analysis

Even though the SE-POSTER method that we propose does bring about better recognition results on multiple databases, namely RAF-DB, FERPlus, and AffectNet, in several instances, when tested on real-world scenarios with a variety of conditions, it still showed some errors which we can call failures. We analyzed cases of confusion as well as instances when the method produced visually extreme but inaccurate results and have found that the majority of the errors are mainly due to very similar looking and thus easily confused emotional categories, both in terms of visual and semantic properties. In addition, several factors related to unconstrained images and environmental aspects may also contribute to the failure of the system. These include partial occlusion, a head that is turned at a large angle, dark illumination, a cluttered background, and poor image quality. All these may lead to an inability to accurately pinpoint facial landmarks and, at the same time, vital regions on the face, like eyes, brows, and mouth, having their information degraded. The SE attention module we have designed is a valuable tool for addressing these issues as it is capable of adaptively adjusting the channel-wise feature response and giving more prominence to the emotion-correlated representations before the Transformer-based attention operation. Nevertheless, mistakes occur when the available visual clues for an expression are weak or when different kinds of emotions are expressed through very similar facial cues. This discovery opens up a potential avenue for future efforts that could involve the use of uncertainty estimation, expression intensity measurement, and more powerful cross-dataset domain adaptation techniques to further elevate recognition performance under challenging real-world situations.

5. Conclusions

This study introduced SE-POSTER, a facial emotion recognition framework that integrates Squeeze-and-Excitation (SE) channel-wise attention into a landmark-guided CNN–Transformer architecture. The proposed approach addresses a key limitation of existing hybrid FER systems, namely the lack of adaptive channel-wise feature refinement. By incorporating SE attention blocks within a multi-scale feature pyramid, the model performs channel recalibration at fine-, mid-, and global-level feature representations. This design enhances the discriminative capacity of the extracted features while preserving the advantages of the original POSTER architecture, including landmark guidance, multi-scale feature modeling, and Transformer-based global dependency learning.

Extensive experiments conducted on three benchmark FER datasets—RAF-DB, FER-Plus, and AffectNet—demonstrate that the proposed SE-POSTER model consistently outperforms the baseline POSTER and several state-of-the-art methods in terms of recognition accuracy. In particular, the strong performance observed on RAF-DB indicates the model's capability to learn more robust emotional representations. Furthermore, ablation studies confirm that channel attention applied at multiple pyramid levels significantly improves feature representation and overall model performance. These results highlight the importance of prioritizing channel-wise feature importance before performing global dependency modeling in Transformer-based FER frameworks. In addition, the proposed model remains lightweight and computationally efficient, making it suitable for practical affective computing applications.

While the newly suggested SE-POSTER system was tested on the most popular facial emotion recognition datasets "in the wild", conducting a detailed subgroup analysis based on demographics has not been given in this research paper. More specifically, the present evaluation did not uniquely assess the program's effectiveness with different age ranges, male and female groups, ethnicities, cultural communities, or geographical expression patterns. This is a major drawback since the nature of facial expressions, how intensely emotions are displayed, and even social behaviors of expression may differ from one demographic/cultural group to another and thus cause distribution changes resulting in the lack of reliability and unfairness of facial emotion recognition systems when these are used in the real world. Hence, although a demographic diversity analysis is well outside the main experimental scope of this paper, we recognize it as a major drawback. We plan to carry out a demographic-aware and cross-cultural evaluation of the framework proposed here in future work. It could also involve performance analysis at the subgroup level, comparison of different fairness-aware evaluation metrics, validation across datasets on culturally diverse populations, and development of domain adaptation strategies for better generalization under demographic distribution shifts. In addition, work to balance demographic datasets and prediction methods mindful of uncertainty can be done to lessen bias and raise the reliability of facial emotion recognition systems in real applications. Another promising direction involves biologically inspired neural network models that simulate neurotransmitter-based interactions through multiple connection weights between neurons [37,38], potentially improving the interpretability and representational capacity of learned features. In addition, enhancing the transparency of classification decisions through interpretable data analysis techniques [39]—such as feature discretization and rule-based inference mechanisms—could improve the explainability and reliability of deep learning-based emotion recognition systems. Finally, deploying the model in real-time human–computer interaction environments represents an important step toward practical applications. Overall, the proposed SE-POSTER framework provides a scalable and effective solution for robust facial emotion recognition in real-world scenarios.

Author Contributions: Conceptualization, A.K. and F.M.; methodology, A.K. and G.T.; software, A.K.; validation, K.S., P.A. and G.T.; formal analysis, A.K. and R.L.; investigation, K.S. and R.L.; resources, S.I. and R.L.; data curation, S.I. and P.A.; writing—original draft preparation, A.K.; writing—review and editing, A.K. and F.M.; visualization, S.I. and. G.T.; supervision, K.S. and F.M.; project administration, F.M.; funding acquisition, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used in this study are publicly available. The RAF-DB dataset is accessible at http://www.whdeng.cn/RAF/model1.html (accessed on 2 February 2026), and the AffectNet dataset can be obtained from https://affectnet.org/ (accessed on 2 February

2026). Source code, model implementation, training configuration, and evaluation scripts used in this study will be shared with readers and other researchers upon a reasonable request to the corresponding author.

Acknowledgments: The authors would like to thank the editor and anonymous referees for the constructive comments in improving the contents and presentation of this paper.

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

References

1. Shan, L.; Weihong, D. Deep facial expression recognition: A survey. *J. Image Graph.* **2020**, *25*, 2306–2320. [CrossRef]
2. Ma, N.; Qu, B.; Wang, W.; Wang, F.; Zhang, X. SCAF-Net: A spiking cross-modal attention fusion network for multimodal emotion recognition. *Biomed. Signal Process. Control* **2026**, *118*, 109684. [CrossRef]
3. Haq, M.; Athar, M.; Ahmad, S.; Ahmad, N.; Anwar, M.; Kutlimuratov, A. A Comprehensive Review of Face Detection/Recognition Algorithms and Competitive Datasets to Optimize Machine Vision. *Comput. Mater. Contin.* **2025**, *84*, 1–24. [CrossRef]
4. Bian, Y.; Kim, H.; Krumhuber, E.G. A Cross-Corpus Evaluation on Spontaneous and Dynamic Facial Expressions for Automated Emotion Classification. *Electronics* **2026**, *15*, 849. [CrossRef]
5. Hunafa, M.H.; Ramadhan, A.W.; Kushirayati, S.; Abka, A.F.; Mantau, A.J.; Jatmiko, W. Data Imbalance Handling in Facial Expression Recognition: A Systematic Literature Review. *IEEE Access* **2026**, *14*, 8269–8287. [CrossRef]
6. Safarov, F.; Kutlimuratov, A.; Khojamuratova, U.; Abdusalomov, A.; Cho, Y.-I. Enhanced AlexNet with Gabor and Local Binary Pattern Features for Improved Facial Emotion Recognition. *Sensors* **2025**, *25*, 3832. [CrossRef] [PubMed]
7. Hebri, D.; Nuthakki, R.; Digal, A.K.; Venkatesan, K.G.S.; Chawla, S.; Raghavendra Reddy, C. Effective facial expression recognition system using machine learning. In *Proceedings of the EAI Endorsed Transactions on Internet of Things*; European Alliance for Innovation: Gent, Belgium, 2024; p. 10.
8. Grover, R.; Bansal, S. Enhancing facial expression recognition in uncontrolled environment: A lightweight CNN approach with pre-processing. *Neural Comput. Appl.* **2025**, *37*, 7363–7378. [CrossRef]
9. Qadir, I.; Iqbal, M.A.; Ashraf, S.; Akram, S. A fusion of CNN And SIFT For multicultural facial expression recognition. *Multimed. Tools Appl.* **2025**, *84*, 33505–33523. [CrossRef]
10. Kashef, A.; Wang, Y.; Assafi, M.N.; Ma, J.; Wang, J.; Jones, J.A.; Thiamwong, L. Developing A novel AI enabled extended reality system for real-time automatic facial expression recognition and system performance evaluation. *Adv. Eng. Inform.* **2025**, *65*, 103207. [CrossRef]
11. Nawaz, U.; Saeed, Z.; Atif, K. A Novel Transformer-based approach for adult's facial emotion recognition. *IEEE Access.* **2025**, *13*, 56485–56508. [CrossRef]
12. Wang, Y.; Pan, K.; Shao, Y.; Ma, J.; Li, X. Applying a convolutional vision transformer for emotion recognition in children with autism: Fusion of facial expressions and speech features. *Appl. Sci.* **2025**, *15*, 3083. [CrossRef]
13. Zheng, C.; Mendieta, M.; Chen, C. Poster: A pyramid cross-fusion transformer network for facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2023; pp. 3146–3155.
14. Radočaj, P.; Martinović, G. Emotion Recognition in Autistic Children Through Facial Expressions Using Advanced Deep Learning Architectures. *Appl. Sci.* **2025**, *15*, 9555. [CrossRef]
15. Agarwal, A.; Susan, S. Attention-augmented squeeze-and-excitation enhanced mobile network for occluded facial expression recognition in resource-constrained environments. *Signal Image Video Process.* **2025**, *19*, 687. [CrossRef]
16. Manavand, M.R.; Salarifar, M.H.; Ghavami, M.; Taghipour-Gorjikolaie, M. Driver's facial expression recognition by using deep local and global features. *Inf. Sci.* **2025**, *692*, 121658. [CrossRef]
17. Mohamed, A.; Nii, R.; Binga, K.; Modi, S.; Nagaveni, P.; Nnadhini, T.J. Facial expression recognition in real-time surveillance using CNN and transfer learning with ResNet-50. In *2025 International Conference on Automation and Computation (AUTOCOM)*; IEEE: New York, NY, USA, 2025; pp. 229–234.
18. Bhati, V.S.; Tiwari, N.; Chawla, M. A generalized zero-shot deep learning classifier for emotion recognition using facial expression images. *IEEE Access* **2025**, *13*, 18687–18700. [CrossRef]
19. Kumar, R.; Corvisieri, G.; Fici, T.F.; Hussain, S.I.; Tegolo, D.; Valenti, C. Transfer learning for facial expression recognition. *Information* **2025**, *16*, 320. [CrossRef]
20. Akeh, L.J.; Kusuma, G.P. Mixed emotion recognition through facial expression using transformer-based model. *Stat. Optim. Inf. Comput.* **2025**, *13*, 531–546. [CrossRef]
21. Tagmatova, Z.; Umirzakova, S.; Kutlimuratov, A.; Abdusalomov, A.; Im Cho, Y. A Hyper-Attentive Multimodal Transformer for Real-Time and Robust Facial Expression Recognition. *Appl. Sci.* **2025**, *15*, 7100. [CrossRef]
22. Makhmudov, F.; Kutlimuratov, A.; Cho, Y.-I. Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition. *Appl. Sci.* **2024**, *14*, 11342. [CrossRef]

23. Sun, Z.; Liu, H.; Li, H.; Li, Y.; Zhang, W. AVERFormer: End-to-end audio-visual emotion recognition transformer framework with balanced modal contributions. *Digit. Signal Process.* **2025**, *161*, 105081. [CrossRef]
24. Sun, R.; Zhang, Z.; Liu, H.; Zhao, L.; Zhou, Q.; Liu, Z. DacFER: Dual Attention Correction Learning for Efficient Facial Expression Recognition. In *2024 7th International Conference on Electronics Technology (ICET)*; IEEE: New York, NY, USA, 2024; pp. 941–945.
25. Zhang, Q.; Liu, Y.; Zhu, B.; Han, X.; Zhang, R.; Xiao, J.; Wang, Z. Deep multi-modal fusion transformer for emotion recognition. *Eng. Appl. Artif. Intell.* **2026**, *168*, 113967. [CrossRef]
26. Xiong, K.; Qing, L.; Li, L.; Guo, L.; Peng, Y. Facial expression recognition based on local–global information reasoning and spatial distribution of landmark features. *Vis. Comput.* **2025**, *41*, 535–548. [CrossRef]
27. Do, H.Q.; Thanh, H.V.; Phuong, T.M. A Study on Fusion Strategies of Facial Landmark-Based Heatmap for Facial Expression Recognition. *KSII Trans. Internet Inf. Syst.* **2025**, *19*, 3602. [CrossRef]
28. Jiang, W.; Zhao, Z.; Wang, L.; Liu, F.; Qing, C.; Xing, X.; Xu, X.; Fan, W.; Jin, Z. A dual uncertainty-aware fusion framework for face expression recognition in the wild. *Expert Syst. Appl.* **2025**, *298*, 129567. [CrossRef]
29. Praveen, R.G.; De Melo, W.C.; Ullah, N.; Aslam, H.; Zeeshan, O.; Denorme, T.; Pedersoli, M.; Koerich, A.L.; Bacon, S.; Cardinal, P.; et al. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE Computer Society: Washington, DC, USA, 2022; pp. 2486–2495.
30. Khan, T.; Yasir, M.; Choi, C. Attention-enhanced optimized deep ensemble network for effective facial emotion recognition. *Alex. Eng. J.* **2025**, *119*, 111–123. [CrossRef]
31. Ryumina, E.; Ryumin, D.; Axyonov, A.; Ivanko, D.; Karpov, A. Multi-corpus emotion recognition method based on cross-modal gated attention fusion. *Pattern Recognit. Lett.* **2025**, *190*, 192–200. [CrossRef]
32. Yi, M.H.; Kwak, K.C.; Shin, J.H. HyFusER: Hybrid multimodal transformer for emotion recognition using dual cross modal attention. *Appl. Sci.* **2025**, *15*, 1053. [CrossRef]
33. Li, S.; Deng, W. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. *IEEE Trans. Image Process.* **2019**, *28*, 356–370. [CrossRef] [PubMed]
34. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16)*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 279–283. [CrossRef]
35. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.* **2019**, *10*, 18–31. [CrossRef]
36. Mao, J.; Xu, R.; Yin, X.; Chang, Y.; Nie, B.; Huang, A.; Wang, Y. Poster++: A simpler and stronger facial expression recognition network. *Pattern Recognit.* **2025**, *157*, 110951. [CrossRef]
37. Rakhimovich, A.M.; Kadirbergenovich, K.K.; Ishkobilovich, Z.M.; Kadirbergenovich, K.J. Logistic Regression with Multi-Connected Weights. *J. Comput. Sci.* **2024**, *20*, 1051–1058. [CrossRef]
38. Kabulov, A.; Babadzhanov, A.; Baizhumanov, A.; Saymanov, I.; Babadjanov, A. Algorithms for Solving Systems of Boolean Equations Based on the Transformation of Logical Expressions. *Mathematics* **2026**, *14*, 594. [CrossRef]
39. Madrakhimov, S.; Makharov, K.; Khurramov, A. On the Transparency of Decision-Making in Classification by Precedents with Fuzzy Descriptions. *IEEE Access* **2025**, *13*, 173656–173664. [CrossRef]

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