

Facial Expression Recognition using Label- and Scale-Distribution-aware Loss-assisted Cross Vision Transformer

Bharathi Manchenahalli Gopalakrishna Rao^{1*}, Divya Savithramma Ramakrishna Reddy²

Assistant Professor, Department of Computer Science, Vijaya College, Basavanagudi, Bangalore, India.

E-mail: ¹bharathimg@gmail.com, ²divyasr@vijayacollege.ac.in

Abstract

In recent years, Facial Expression Recognition (FER) has gained attention as a crucial area of computer vision. However, FER methods depend on single-source facial data that often experience reduced accuracy and instability because of challenges such as facial occlusion and variations in illumination. Moreover, existing FER approaches fail to understand the salient facial regions, resulting in poor interpretability and performance. Hence, this research proposes the Label- and Scale- Distribution-aware Loss-assisted Cross Vision Transformer (LSDL-CrossViT) to solve the issues in FER. The proposed model captures the global facial structure and local expression-specific cues through cross-attention-based feature learning. By incorporating the attention mechanism through discriminative loss design, the pipeline improves class separability and reduces classification errors across multiple facial expression categories. This research considers three benchmarks such as RAF-DB, AffectNet-8, and CK+, to estimate the complete performance of the proposed approach. The experimental results illustrate that the proposed LSDL-CrossViT gains a balanced accuracy of 93.18%, 71.84%, and 99.01% on the RAF-DB, AffectNet-8, and CK+ datasets, respectively, as compared to the reproduced Distract your Attention Network (DAN) and Self Learning Weight Network related to Label Distribution Training (SLW-LDT) models.

Keywords: Attention Mechanism, Cross Vision Transformer, Facial Expression Recognition, Global Facial Structure, Label- and Scale- Distribution-aware Loss.

1. Introduction

Facial expressions are important indicators of emotional states and intentions in human interaction [1]. Face detection and recognition are widely considered research subjects in the area of computer vision and have applications in a wide variety of areas such as emotion recognition, digital entertainment, and security monitoring [2][3]. Facial Expression Recognition (FER) is a crucial task in affective computing, with applications in human-computer interaction, education, and mental health support [4]. An individual image conveys more information, as facial expressions represent human emotional states such as happy, sad, fear, disgust, neutral, surprise, and anger. However, the recognition of facial emotions from images is a complex operation in the area of computer vision [5]. Facial expressions are obtained from coordinated facial muscle movements or subtle facial

distortions that convey both physical and emotional states [6]. However, subtle facial details, particularly around the eyes and lips, may occupy only a small region of a facial image. These fine-grained features can be difficult to capture accurately, making facial expression recognition challenging [7] [8]. To address this challenge, researchers have developed various Deep Learning (DL)-based FER approaches for learning discriminative facial features and recognizing emotional states from facial images [9].

Automatic FER has become a crucial topic in research owing to the development of artificial intelligence (AI) techniques and advancements in computer vision [10] [11]. An important FER framework involves different key phases such as feature extraction, detection, and classification [12]. The face detection phase is attended through a facial arrangement step to ensure that each image is spatially symmetrical [13]. In the FER pipeline, face detection is commonly performed using open-source implementations and recently developed Deep Learning (DL)-based detection approaches [14]. Several studies have focused on global feature extraction from the entire face, such as appearance and contour characteristics [15]. However, existing FER approaches primarily rely on DL and attention mechanisms to extract facial features or classify expressions. These methods may not sufficiently model the semantic relationships between facial regions and their corresponding expressions [16].

1.1 Problem Statement and Objective

FER remains challenging due to occlusions and illumination changes because traditional Convolutional Neural Network (CNN)-assisted approaches depend on local texture patterns and often struggle to model long-range facial dependencies. Hence, this research proposes a Label- and Scale- Distribution-aware Loss-assisted Cross Vision Transformer (LSDL-CrossViT) approach that emphasizes informative facial regions while maintaining global context. The proposed contribution lies in its task-specific integration with Region-Aware Attention (RAA) and Label- and Scale-Distribution-aware Loss (LSDL) for FER. RAA operates on intermediate feature representations to emphasize expression-sensitive facial regions, whereas LSDL operates on the classification objective by jointly considering difficult samples, class distribution, and multi-scale branch contributions. Thus, the proposed framework combines multi-scale feature interaction, region-aware refinement, and distribution-aware optimization within a unified FER pipeline. The main contributions of this research are as follows:

- This research proposes the RAA refinement integrated with CrossViT for facial expression recognition, allowing an approach for understanding global facial structure and fine-grained local expression cues.
- Through the incorporation of cross-attention among multi-scale feature streams, the proposed framework minimizes the confusion between visually similar emotions. This framework improves discriminative representation learning over traditional CNN and standard transformer models.
- The importance of the proposed approach is validated through experiments on three benchmark datasets (RAF-DB, AffectNet-8, and CK+), demonstrating consistent accuracy improvements over seven reproduced existing methods.

This research paper is arranged as follows: Section 2 specifies the related work. Section 3 shows the proposed methodology. Section 4 specifies the results and discussion. The conclusion of this research work is shown in Section 5.

2. Related Works

Nowadays, DL approaches have been widely applied to FER operations and have attained improved results. Abbas Issa Jabbooree et al. [17] introduced an FER approach that integrated CNN and Support Vector Machine (SVM) for the extraction of local and global features, thus improving discriminative power. The introduced FER approach utilized a β -skeleton undirected graph to enhance geometric feature extraction through diverse pre-processing phases. Saining Zhang et al. [18] suggested the Dual Direction Attention Mixed Feature Network (DDAMFN) approach for FER, which enhanced its lightweight characteristics. The DDAMFN pipeline comprises different key constituents such as the Mixed Feature Network (MFN), which serves as the baseline network, and the DDAMFN module, which functions as the classification head. Resilient features were extracted through mixed-size kernels to improve the model's capability in the MFN. Zhengyao Wen et al. [19] presented a Distract your Attention Network (DAN) for FER. The DAN approach consists of three main components: the Attention Fusion Network (AFN), Multi-head Attention Network (MAN), and Feature Clustering Network (FCN).

DanChen et al. [20] implemented the FER approach using an improved ConvNeXt model. Incorporating the Squeeze-and-Excitation Network (SENet) attention mechanism into the ConvNeXt block enhances the extraction of discriminative facial features. Moreover, the integration of the Focal Loss (FL) function optimized the classification performance of the network. Samia Nawaz Yousafzai et al. [21] introduced a graph-assisted FER pipeline, which incorporated multi-scale feature extraction through attention mechanisms for improved emotion detection. Primarily, YOLOv8 detected the faces and allowed for the design of multi-scale graphs to estimate the spatial relationships between the features. Yangbo Chen et al. [22] presented the Self Learning Weight Network related to Label Distribution Training (SLW-LDT) for FER. Based on the ShuffleNet-V2 core network, the SLW-LDT developed a local feature extraction block that emphasized particular expression-based features by compressing the facial image into different local areas.

2.1 Research Gap

Recent FER research increasingly combines convolutional feature extraction with transformer-based representation learning to overcome the limited receptive-field characteristics of conventional CNN architectures. However, transformer-based approaches can still face challenges associated with computational cost, class imbalance, and fine-grained discrimination among visually similar expressions. To solve these limitations, this research developed an improved FER framework that integrates multi-scale cross-attention interaction with an LSDL to enhance fine-grained emotion discrimination. The proposed LSDL-CrossViT introduces a task-specific optimization strategy for facial expression recognition. The development of joint integration of multi-scale cross-attention learning, region-aware feature refinement, and an LSDL, simultaneously addresses class imbalance and sample difficulty.

3. Proposed Methodology

The proposed framework introduces an LSDL-CrossViT architecture, developed to address the restrictions of conventional CNN and transformer-assisted FER models. The proposed approach incorporates cross-attention between global and local facial representations. Initially, this research discusses the datasets considered in the implementation and experiments. Figure 1 outlines a comprehensive procedure of the proposed methodology in FER.

3.1 Dataset Description

To estimate the importance of the proposed CrossViT approach, several standard datasets, namely RAF-DB [23], AffectNet-8 [24], and CK+ [25] are considered in this research. The RAF-DB dataset is selected for its training and stimulating constraints, which is used to estimate the FER approaches using different constituents such as lighting, occlusion, and pose. RAF-DB is a FER benchmark obtained from the Internet, which involves 29,672 facial samples labeled with different common emotion labels and 11 compound emotion labels. The 29,672-image figure describes the full RAF-DB corpus, which includes both the seven basic-expression classes and additional compound-expression classes. This work uses only the standard seven-class basic-expression subset (15,339 images total: 12,271 training + 3,068 testing, detailed below). This RAF-DB dataset involves 12,271 training instances (happy (4772), fear (281), neutral (2524), angry (705), surprise (1290), sad (1982), and disgust (717)), and 3068 testing instances obtainable for FER.

The AffectNet-8 dataset is a widely used dataset for FER that involves eight emotion classes, including contempt, which is represented by 3,750 additional training images. Consequently, the training set contains 287,651 images, while the test set contains 4,000 images distributed across the eight classes. The CK+ dataset is basically utilized in FER setups, which involve 593 sequences collected from 123 subjects; however, 327 emotion-labeled sequences are used and typically utilized for facial expression recognition tasks. Peak-expression frames were extracted from the emotion-labelled sequences to construct the still-image subset used in this study. Specifically, three frames were selected from each sequence: the apex frame and the two frames immediately preceding it. All three frames were assigned the emotion label of their corresponding sequence. To avoid subject leakage during CK+ evaluation, subject-independent partitioning is performed.

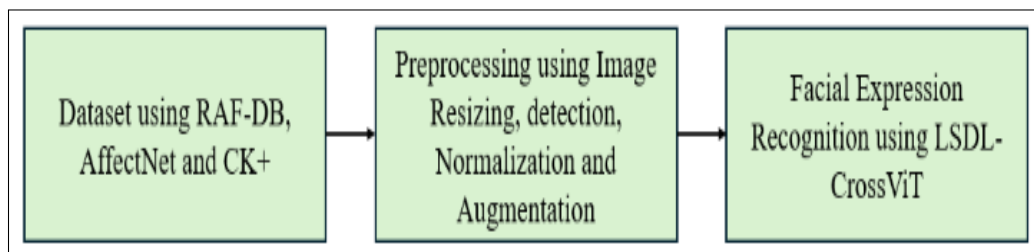


Figure 1. Overall workflow of the proposed LSDL-CrossViT framework for facial expression recognition task

3.2 Data Preprocessing

The data preprocessing phase is important to establish the basic input conditions because each image in the dataset contains customized parameters such as illumination, pose, and resolution. Hence, this research performs diverse preprocessing steps such as image resizing, face detection, normalization, and data augmentation. The details are provided in the following sections.

3.2.1 Image Resizing and Face Detection

The images in the considered datasets have different sizes; therefore, each image is resized to a similar size 224×224 , which improves the performance of the model. In this study, face localization is applied using the Haar Cascade detector. The purpose of this step is to provide a computationally efficient and consistent procedure for detecting and cropping

facial regions, rather than to introduce a new face-detection model. Haar Cascade has a relatively low computational overhead and performs rapid frontal-face localization. The computationally demanding DL detectors such as MTCNN and RetinaFace provide stronger performance for unconstrained pose, scale, and occlusion conditions; however, using these alternative detectors would increase the computational cost and introduce an additional variable into the comparison. Therefore, the Haar Cascade is retained as a lightweight and consistent preprocessing component across the evaluated datasets. However, its performance can be affected by large pose variations, occlusion, scale changes, and non-frontal faces. More advanced detectors such as MTCNN or RetinaFace may provide stronger localization under unconstrained conditions but would introduce additional computational cost and another variable into the comparison.

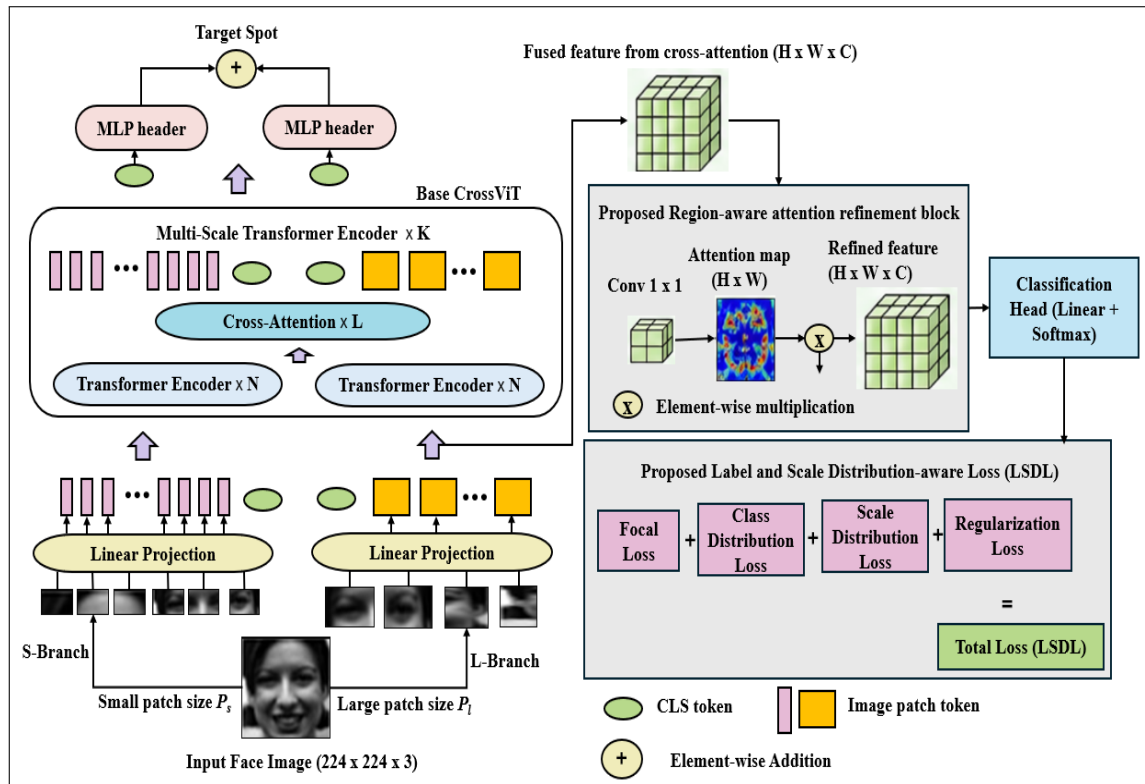


Figure 2. Architecture of CrossViT which comprises feature extraction, fusion, and classification modules for FER

3.2.2 Image Normalization

Normalizing the input data in a uniform manner is an important step for improving model performance. Most studies have employed the 0 to 1 normalization technique prior to providing the input value to the classifier. However, the images in the considered dataset involve irregular light and shade due to shadows, illumination, and external conditions, resulting in poor FER performance. Hence, this research employs the Contrast Limited Adaptive Histogram Equalization (CLAHE) approach for processing the input image rather than plain Histogram Equalization (HE). The images are processed using a clip limit of 2.0 and an 8×8 tile grid. The common concept in this approach is to design a similar distribution from an actual image histogram, which is computed locally within each tile to avoid over-amplifying noise in near-uniform regions.

3.2.3 Data Augmentation

DL approaches depend on important training data to attain improved recognition results. To solve this problem, data augmentation approaches have been developed to enhance FER datasets, thus allowing the recognition approach training without overfitting. By utilizing the image data generator parameters, these approaches produce a number of images from a single input using augmentation approaches such as flipping, sharpening, rotation at various angles, and Gaussian blur. Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to improve contrast consistency across facial images rather than plain HE. Although excessive contrast enhancement may potentially alter fine-grained textural information, the applied normalization parameters are selected conservatively to maintain expression-related facial details. Each augmentation operation is applied independently to a given training image with its own probability: Horizontal flip = 0.5, rotation of $\pm 15^\circ$ (angle sampled uniformly, applied with probability 0.3), brightness of $\pm 20\%$ (factor sampled uniformly, applied with probability 0.3), scaling of 0.9-1.1 (factor sampled uniformly, applied with probability 0.3), gaussian blur of kernel 3×3 , probability 0.2, and random sharpening of probability 0.2. Gaussian blur augmentation is integrated with a low probability to improve performance against image quality variations.

3.3 Facial Expression Recognition using CrossViT

This research adopts the CrossViT approach, which is obtained from the base model of Vision Transformer (ViT). The proposed framework adopts CrossViT as the backbone architecture because of its ability to model multi-scale visual representations. As compared to the existing CNN approach, CrossViT utilizes cross-attention mechanisms for modeling the global dependencies among the various image scales. This ability crucially enhances the model performance in understanding large visual data. To preserve the spatial relationships among facial image patches, positional embeddings are incorporated into the token representations before transformer processing. These positional embeddings provide information regarding the relative location of each patch within the facial image, enabling the transformer to capture spatial dependencies that are essential for facial expression recognition. By combining positional information with patch embeddings, the model learns both appearance-based and spatial characteristics of facial expressions.

Before transformer-based feature extraction, each facial image is categorized as a fixed-size, non-overlapping patch. These patches are flattened and transformed into feature embeddings using a learnable projection layer, producing a sequence of visual tokens. The generated token representations capture local facial characteristics and serve as the input to the CrossViT architecture. This tokenization process enables the transformer to efficiently model facial expression information at multiple spatial scales. Scale distribution-awareness refers to the adaptive weighting of multi-scale patch features together with class-distribution-aware loss weighting, enabling the model to learn discriminative representations across both feature scales and imbalanced emotion categories.

Figure 2 illustrates the complete framework of CrossViT, which comprises three modules for FER. In the input phase, an image is divided into different patch scales, such as small and large, to attain a balance between modeling fine-grained global and local semantic data. In each branch, standard self-attention is primarily used for the respective patch token for modeling intra-scale feature depictions. This strategy utilizes the CLS token from a branch as

a query, whereas patch tokens from various branches provide key and value in cross-attention estimation. An input to the cross-attention function is illustrated in Equation (1) as:

$$x'^l = [f^l(x_{cls}^l) \| x_{patch}^s] \quad (1)$$

Here, x'^l specifies the updated feature representation of the large branch after cross-scale fusion; x_{patch}^s demonstrates the patch token embeddings from the small-scale branch; s and l specifies the small and large-scale branches individually. $f^l(.)$ denotes a projection function utilized for the dimension alignment. The projection operator is employed to align the embedding dimensions prior to cross-scale fusion. Consequently, the fused representation maintains semantic data from both local and global facial regions while maintaining dimensional consistency. x_{cls}^l illustrates a CLS token of the large branch, and after an approach employs the cross-attention estimation among x_{cls}^l and x'^l , as formulated in Equations (2) and (3) as:

$$q = x'^l W_q, \quad k = x'^l W_k, \quad v = x'^l W_v \quad (2)$$

$$A = softmax\left(\frac{qk^T}{\sqrt{C/h}}\right), \quad CA(x'^l) = Av \quad (3)$$

Here, $W_q, W_k, W_v \in R^{C \times (C/h)}$ denote the learnable weight parameters for query, key, and value projections individually; h denotes the count of attention heads, and C denotes the embedding dimension. q, k and v specify the query vector, key matrix, and value matrix individually. $\sqrt{C/h}$ demonstrates the scaling factor for stabilizing the dot-product attention; A means attention weight matrix; $softmax(.)$ illustrates the softmax activation function; and $(.)^T$ specifies the matrix transpose operation. The scaling factor is introduced to stabilize the magnitude of the dot-product operation during attention computation. The small-scale and large-scale branches operate in parallel and exchange information through periodic cross-attention interactions. During each interaction stage, representative tokens from one branch attend to the token representations of the other branch, enabling mutual information transfer while preserving branch-specific feature characteristics. This synchronization mechanism facilitates the integration of local and global facial representations.

In the proposed cross-attention module, the [CLS] token is used as the query vector. Since the query attends to the patch tokens of the other branch, the attention computation scales linearly with the number of patch tokens. Therefore, the computational and memory costs of this operation are lower than those of full self-attention, which has quadratic complexity with respect to the number of tokens. This framework enhances the complete performance of the process. Particularly, the result of the cross-attention output is merged with the normalized and modified CLS token by the residual connection and merged through patch tokens for the following encoding, which is formulated in Equations (4) and (5) as follows:

$$y_{cls}'^l = f^l(x_{cls}^l) + MCA\left(LN\left([f^l(x_{cls}^l) \| x_{patch}^s]\right)\right) \quad (4)$$

$$Z^l = [g^l(x_{cls}^l) \| x_{patch}^l] \quad (5)$$

Where $f^l(.)$ and $g^l(.)$ denote the projection and inverse functions utilized for dimension

arrangement. y_{cls}^l means updated CLS token of large branch after residual cross-attention fusion; x_{patch}^l means the patch tokens from the large-scale branch; $LN(\cdot)$ denotes the layer normalization operation; $MCA(\cdot)$ denotes the Multi-head Cross-Attention function; and Z^l specifies the final large-branch feature representation passed to the next transformer layer. The output of the cross-attention module preserves the dimensionality of the large-scale branch representation. Therefore, the residual connection ensures information preservation while facilitating stable gradient propagation throughout transformer optimization. Subsequently, the CLS tokens from the small and large branches are transmitted to the following transformer encoding layers, enhancing their semantic depiction. For classification, the last CLS tokens are moved to the Multilayer Perceptron (MLP) head. The classified data from both branches are concatenated to identify the class of facial expressions.

3.4 Label- and Scale- Distribution-aware Loss

LSDL is a loss function applied to the model logits during training, which is conceptually and mathematically different from the region-aware spatial attention (RAA) module. It uses a 1×1 convolution to generate a spatial score map, followed by softmax-based spatial weighting to emphasize expression-sensitive facial regions. In contrast, LSDL operates at the output level and guides the model during optimization by modifying the training objective. The two components are complementary but separate: RAA reweights spatial feature locations, while LSDL reweights the classification loss contribution of each training instance and class. This loss function solves the class imbalance issue by treating the sample distribution and leveraging the proportional distribution to employ the classification complexity of each instance, as shown in Equation (6). The RAA module generates a spatial attention map from the intermediate feature representation to emphasize expression-sensitive facial regions, which is formulated in Equation (7) as follows:

$$FL(p_t) = \alpha_t (1 - p_t)^\gamma CE(p, y) \quad (6)$$

$$RAA = Conv_{1 \times 1}(F). A_{h,w} F(h, w, c) \quad (7)$$

Where α_t illustrates the weight coefficient in a range from 0 to 1 for fine-tuning the positive and negative instance weights. $FL(p_t)$ demonstrates the focal loss value for a training instance; p_t means the predicted likelihood of the true class; γ means focusing parameter that minimizes the loss contribution of simple samples; $CE(p, y)$ illustrates a cross-entropy loss function and $A_{h,w} F(h, w, c)$ denotes a refined feature representation. Moreover, the weight coefficient $(1 - p_t)^\gamma$ is used to adjust the contribution of the classification head on weight adjustments. This adjustment allows an approach to prioritize a particular instance class at the time of training, thus rectifying the class imbalance in the positive class, which is formulated in Equation (8) as follows:

$$Loss(p_t) = -\alpha_t (1 - p_t(c_i))^\gamma P_P(c_i) \log(p_t(c_i)) \quad (8)$$

Where, $Loss(p_t)$ demonstrates the weighted focal-based loss value; $p_t(c_i)$ means predicted probability for class c_i ; $P_P(c_i)$ means proportional distribution weight for class c_i ; and $\log(\cdot)$ denotes the natural logarithm function. The weight ω_i is for performing at a large level and is combined into the loss estimation to ascertain the important loss value. By proportionally distributing the instances among the classes, the proposed strategy reduces class

imbalance and improves the representation of minority classes. Subsequently, the proposed LSDL function is formulated in Equations (9) to (11). To compensate for class imbalance, the class-distribution weighting factor is introduced, which is formulated in Equation (12) as follows:

$$Loss' = (1 + \omega_i) Loss \quad (9)$$

$$LSDL - Loss = -\alpha_t \left(2 + \frac{t + \beta}{\log(t)} \right) (1 - p_t(c_i))^\gamma \quad (10)$$

$$LSDL - Loss = \frac{\left(\frac{1}{n_i}\right)^\rho}{\sum_{j=1}^t \left(\frac{1}{n_j}\right)^\rho} \log(p_t(c_i)) \quad (11)$$

$$w_c = \frac{1}{\log(1 + n_c)} \quad (12)$$

Where ω_i specifies the sample-level weighting factor associated with sample i and the parameter β specifies the focal-loss focusing parameter that controls the contribution of difficult samples, and the parameter ρ denotes the distribution-related weighting parameter used to adjust the contribution of samples based on the distribution of samples across the classes. k illustrates the class count; n_i specifies the instances count in i th class; w_c denotes the inverse-frequency class weight assigned to class c ; n_c is the raw number of training samples in class c ; The $w_c = 1/\log(1 + n_c)$ as n_c increases, minority classes (small n_c) receive systematically larger weights than majority classes, satisfying the inverse-frequency weighting requirement for class-imbalance correction. This is an inverse (logarithmically damped) weighting function rather than a simple proportional weight, which avoids assigning excessively large weights to rare classes. To make the scale-distribution-aware component of LSDL, a branch-specific scale weight is introduced: given the small-branch (16×16 patch) prediction $p_t^{s(c_i)}$ and large-branch (32×32 patch) prediction $p_t^{l(c_i)}$ for class c_i , a per-class learnable scale weight μ_s and μ_l (with $\mu_s + \mu_l = 1$) is applied so that the final scale-aware prediction is $p_t^{s(c_i)} = \mu_s \cdot p_t^{s(c_i)} + \mu_l \cdot p_t^{l(c_i)}$ and this scale-combined $p_t^{s(c_i)}$ is the quantity used in Equations (9) and (10) above, rather than a single-branch prediction. μ_s and μ_l are optimized jointly with the network weights via backpropagation that is initialized at 0.5 each. The proposed LSDL is expected to indirectly improve larger class-level feature margins by applying larger optimization weights to minority and difficult samples, since the LSDL weighting mechanism increases the contribution of minority samples during optimization, which improves the learned decision boundaries, thereby improving classification performance. However, LSDL does not explicitly impose a metric-learning constraint different feature distance. At training, harder samples contribute high gradients to parameter updates, which may shift the decision boundary toward improved class separation.

The focal modulation terms are integrated to emphasize difficult samples, resulting in the proposed LSDL formulation. Although focal loss emphasizes hard samples through probability modulation, it does not explicitly consider the global class-distribution characteristics of the training dataset. In contrast, the proposed LSDL introduces an additional distribution-aware weighting factor derived from class frequency statistics. Consequently, focal loss performs instance-level reweighting, whereas the LSDL combines instance-level and class-level optimization. This dual weighting strategy improves classification performance on highly

imbalanced FER datasets such as RAF-DB and AffectNet-8.

In the proposed LSDL formulation, the small-scale and large-scale branch contributions are controlled by learnable scale weights, denoted by α_s and α_l , respectively. The weights are initialized equally as $\alpha_s = 0.5$ and $\alpha_l = 0.5$, which are jointly optimized with the network parameters through backpropagation. The scale weights are constrained to maintain a valid relative contribution of the two branches with $\alpha_s + \alpha_l = 1$. The class-distribution weight is calculated from the training-fold class frequencies using the inverse-frequency logarithmically damped formulation defined in Equation (11). The target-domain class distribution is not used when performing cross-dataset evaluation.

Algorithm 1: Pseudocode of the proposed LSDL-CrossViT method for better reproducibility

Input: Training dataset D , Learning rate η , Batch size B , and Number of epochs E

Output: Trained LSDL-CrossViT model

- 1: Initialize CrossViT parameters θ and scale weights $\mu_s = \mu_l = 0.5$
 - 2: Compute inverse-frequency class weights $w_c = 1/\log(1 + n_c)$ for each class c using training-fold sample counts n_c (Equation 11)
 - 3: for epoch = 1 to E do
 - 4: for each fold $\rightarrow$ initialize model $\rightarrow$ for epoch=1... $E \rightarrow$ train/validate $\rightarrow$ early stop $\rightarrow$ save best checkpoint
 - 5: Sample mini-batch B from the training-fold data D_{train}
 - 6: Apply preprocessing (CLAHE, resizing) and, for training data only, augmentation
 - 7: Generate patch embeddings at two branch-specific patch sizes (16×16 and 32×32)
 - 8: Extract multi-scale features using CrossViT self-attention within each branch
 - 9: Apply bidirectional CLS-token cross-attention fusion between branches
 - 10: Apply RAA refinement
 - 11: Compute scale-combined prediction $p_t^{s(c_i)} = \mu_s \cdot p_t^{s(c_i)} + \mu_l \cdot p_t^{l(c_i)}$
 - 12: Compute LSDL loss using $p_t^{(c_i)}$ focal terms (α, γ), and class weights w_c
 - 13: if training-fold batch then
 - 14: Backpropagate LSDL loss and update θ and μ_s, μ_l using Adam optimizer (lr = 0.0001, weight decay = 0.0005)
 - 15: else
 - 16: Accumulate validation-fold loss and accuracy (no parameter update)
 - 17: end if
 - 18: end for
 - 19: Evaluate mean validation loss for this epoch
 - 20: if validation loss has not improved for 10 consecutive epochs then
 - 21: Apply early stopping and break
 - 22: end if
 - 23: end for
 - 24: Select the model checkpoint with the lowest validation loss
 - 25: Evaluate the selected model once on the untouched official test set
 - 26: Return trained LSDL-CrossViT model and test-set predictions
-

4. Experimental Results

The internal experimental parameter settings and implementation details of this research for FER are described as follows. All experiments are conducted on a workstation equipped

with an Intel Core i9-13900K CPU, 64 GB RAM, an NVIDIA RTX 3090 GPU (24 GB VRAM), Ubuntu 22.04, CUDA 12.1, and PyTorch 2.1. For FER operations, the CrossViT approach is selected as the backbone model due to its improved performance. Facial feature map extraction is employed through the convolutional layers. In this research, identical hyperparameter settings are employed for both the proposed approach and existing methods, as reported in the performance analysis. The Adam optimization approach is used for optimizing the model weights. Moreover, the learning rate weight decay, batch size, and epoch counts are fixed to 0.0001, 0.0005, 32, and 100, respectively, with Adam's exponential decay rates set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To ensure reproducibility, five independent training runs are conducted using the random seeds 42, 43, 44, 45, and 46. The same seed was consistently applied to parameter initialization, mini-batch generation, and optimization within each run. A LSDL function is applied to handle class imbalance and improve discrimination between similar facial expressions. This loss function incorporates both class distribution and scale-aware weighting, enabling the model to focus on difficult and minority classes. The experimental evaluation follows standard and widely accepted benchmarking protocols to ensure consistency and reliability.

4.1 Performance Analysis

The complete experimental findings acquired through the LSDL-CrossViT on different broadly utilized standard datasets such as RAF-DB, AffectNet-8, and CK+ are illustrated. Baseline models including CNN, ResNet50, Swin transformer (ST), Vision Transformer (ViT), Progressive Attention Vision Transformer (PAVCT), Distract Your Attention Network (DAN), and Dual-Direction Attention Mixed Feature Network (DDAMFN) are considered. For a fair comparison, the reproduced baseline models are evaluated using the same dataset partitions, input resolution, preprocessing pipeline, augmentation policy, evaluation metrics, batch size, and optimizer. The official test partitions were kept unchanged across all compared methods.

Table 1 illustrates the performance analysis of the proposed method with baseline approaches. These test partitions are excluded from model training, hyperparameter tuning, and K-fold model selection. Although the proposed model achieves an accuracy of 71.84% on AffectNet-8, the recall is comparatively lower at 68.37%. This difference is mainly associated with the strong class imbalance of AffectNet-8, where majority classes such as happiness and neutral contain substantially more samples than minority classes such as contempt, fear, and disgust.

Table 2 specifies the K-fold analysis of the proposed approach through different values. The different K values such as 2, 3, 5, 7 and 8 are considered to validate the model performance. For each fold, k-1 folds are used for training and the remaining fold is used only for validation. This process is repeated until each fold has served as the validation fold. The official test partition remains completely untouched during K-fold validation and it is used only once after the final model configuration is selected.

Table 1. Performance comparison of the proposed LSDL-CrossViT framework with baseline approaches across benchmark FER datasets

Dataset	Methods	Balanced Accuracy (Mean $\pm$ SD)	Precision (Mean $\pm$ SD)	Recall (Mean $\pm$ SD)	Macro F1-score (Mean $\pm$ SD)
RAF-DB	CNN	85.42 $\pm$ 0.81	84.31 $\pm$ 0.76	83.76 $\pm$ 0.79	84.03 $\pm$ 0.77
	ResNet50	88.63 $\pm$ 0.68	87.92 $\pm$ 0.64	87.10 $\pm$ 0.67	87.51 $\pm$ 0.65
	ST	90.47 $\pm$ 0.57	89.85 $\pm$ 0.53	89.32 $\pm$ 0.55	89.58 $\pm$ 0.54
	ViT	91.62 $\pm$ 0.49	90.94 $\pm$ 0.46	90.11 $\pm$ 0.48	90.52 $\pm$ 0.47
	PAVCT	91.94 $\pm$ 0.45	91.22 $\pm$ 0.43	90.61 $\pm$ 0.44	90.91 $\pm$ 0.43
	DAN	92.41 $\pm$ 0.41	91.73 $\pm$ 0.39	91.08 $\pm$ 0.40	91.40 $\pm$ 0.39
	DDAMFN	92.83 $\pm$ 0.38	92.12 $\pm$ 0.36	91.34 $\pm$ 0.37	91.73 $\pm$ 0.36
	Proposed LSDL-CrossViT	93.18 $\pm$ 0.42	92.39 $\pm$ 0.39	91.46 $\pm$ 0.41	91.92 $\pm$ 0.40
AffectNet-8	CNN	61.84 $\pm$ 1.12	60.93 $\pm$ 1.08	60.11 $\pm$ 1.05	60.52 $\pm$ 1.06
	ResNet50	64.72 $\pm$ 0.96	63.81 $\pm$ 0.92	63.04 $\pm$ 0.90	63.42 $\pm$ 0.91
	ST	66.95 $\pm$ 0.84	66.11 $\pm$ 0.80	65.38 $\pm$ 0.78	65.74 $\pm$ 0.79
	ViT	67.88 $\pm$ 0.79	67.02 $\pm$ 0.75	66.27 $\pm$ 0.73	66.64 $\pm$ 0.74
	PAVCT	68.74 $\pm$ 0.73	67.93 $\pm$ 0.70	67.18 $\pm$ 0.68	67.55 $\pm$ 0.69
	DAN	69.48 $\pm$ 0.68	68.72 $\pm$ 0.64	67.94 $\pm$ 0.63	68.33 $\pm$ 0.64
	DDAMFN	70.63 $\pm$ 0.62	69.85 $\pm$ 0.59	68.91 $\pm$ 0.58	69.38 $\pm$ 0.58
	Proposed LSDL-CrossViT	71.84 $\pm$ 0.58	69.91 $\pm$ 0.56	68.37 $\pm$ 0.54	69.13 $\pm$ 0.55
CK+	CNN	95.24 $\pm$ 0.52	94.18 $\pm$ 0.48	94.72 $\pm$ 0.49	94.45 $\pm$ 0.48
	ResNet50	97.36 $\pm$ 0.39	96.42 $\pm$ 0.37	96.98 $\pm$ 0.38	96.70 $\pm$ 0.37
	ST	98.12 $\pm$ 0.31	97.26 $\pm$ 0.29	97.81 $\pm$ 0.30	97.53 $\pm$ 0.29
	ViT	98.63 $\pm$ 0.27	98.21 $\pm$ 0.25	97.94 $\pm$ 0.26	98.07 $\pm$ 0.25
	PAVCT	98.71 $\pm$ 0.24	98.33 $\pm$ 0.23	98.01 $\pm$ 0.23	98.17 $\pm$ 0.23
	DAN	98.84 $\pm$ 0.22	98.47 $\pm$ 0.21	98.13 $\pm$ 0.21	98.30 $\pm$ 0.21
	DDAMFN	98.92 $\pm$ 0.20	98.58 $\pm$ 0.19	98.26 $\pm$ 0.19	98.42 $\pm$ 0.19
	Proposed LSDL-CrossViT	99.01 $\pm$ 0.19	98.72 $\pm$ 0.18	98.41 $\pm$ 0.18	98.56 $\pm$ 0.18

Table 2. K-fold validation analysis of the proposed LSDL-CrossViT framework using varying fold configurations across datasets

Dataset	K values	Accuracy (Mean $\pm$ SD)	Precision (Mean $\pm$ SD)	Recall (Mean $\pm$ SD)	F1-score (Mean $\pm$ SD)
RAF-DB	K = 2	92.21 $\pm$ 0.63	91.42 $\pm$ 0.58	90.38 $\pm$ 0.61	90.90 $\pm$ 0.59
	K = 3	92.74 $\pm$ 0.51	91.96 $\pm$ 0.48	90.92 $\pm$ 0.49	91.43 $\pm$ 0.48
	K = 5	93.16 $\pm$ 0.39	92.22 $\pm$ 0.36	91.40 $\pm$ 0.39	91.89 $\pm$ 0.36
	K = 7	93.05 $\pm$ 0.45	92.18 $\pm$ 0.42	91.21 $\pm$ 0.43	91.69 $\pm$ 0.42
	K = 8	92.87 $\pm$ 0.47	92.01 $\pm$ 0.44	90.98 $\pm$ 0.45	91.49 $\pm$ 0.44
AffectNet-8	K = 2	68.24 $\pm$ 0.92	67.18 $\pm$ 0.88	66.12 $\pm$ 0.86	66.65 $\pm$ 0.87
	K = 3	68.91 $\pm$ 0.79	67.92 $\pm$ 0.75	66.88 $\pm$ 0.73	67.40 $\pm$ 0.74
	K = 5	71.81 $\pm$ 0.51	69.88 $\pm$ 0.52	68.34 $\pm$ 0.52	69.09 $\pm$ 0.53
	K = 7	69.12 $\pm$ 0.71	68.01 $\pm$ 0.68	66.94 $\pm$ 0.66	67.47 $\pm$ 0.67
	K = 8	68.73 $\pm$ 0.76	67.82 $\pm$ 0.73	66.71 $\pm$ 0.71	67.26 $\pm$ 0.72

CK+	K = 2	98.24 $\pm$ 0.34	96.32 $\pm$ 0.31	97.18 $\pm$ 0.32	96.75 $\pm$ 0.31
	K = 3	98.71 $\pm$ 0.27	96.84 $\pm$ 0.25	97.72 $\pm$ 0.26	97.28 $\pm$ 0.25
	K = 5	99.87 $\pm$ 0.15	98.69 $\pm$ 0.14	98.38 $\pm$ 0.15	98.48 $\pm$ 0.14
	K = 7	98.92 $\pm$ 0.22	97.10 $\pm$ 0.21	98.02 $\pm$ 0.21	97.56 $\pm$ 0.21
	K = 8	98.63 $\pm$ 0.25	96.88 $\pm$ 0.23	97.71 $\pm$ 0.24	97.29 $\pm$ 0.23

Table 3 presents the cross-dataset evaluation of the proposed LSDL-CrossViT model and the baseline methods using three predefined transfer directions: RAF-DB $\rightarrow$ AffectNet-8, AffectNet-8 $\rightarrow$ RAF-DB, and CK+ $\rightarrow$ RAF-DB. For each direction, the model is trained on the source dataset and evaluated on the target dataset using only the emotion classes shared by both datasets. For the CK+ $\rightarrow$ RAF-DB cross-dataset evaluation, the label space was restricted to the seven emotion classes shared by the two datasets: angry, disgust, fear, happy, neutral, sad, and surprise. The contempt category was excluded because it is not part of the selected seven-class RAF-DB label space. The class mapping was applied consistently during source-domain training and target-domain evaluation. The neutral category was not treated as a CK+ class unless neutral-labelled samples were explicitly available in the dataset used for the experiments. The confusion matrix was configured to display only the classes included in the corresponding evaluation.

Table 4 illustrates the computational complexity comparison of the proposed method with different models for FER using different datasets. The evaluation corresponds to the average inference durations measured under identical hardware and software conditions using batch-wise evaluation on the same GPU environment. The computational complexity values reported in this analysis are expressed in MegaFLOPs (MFLOPs), where 1 MFLOP represents (10^6) floating-point operations. The computational feasibility of the proposed framework was evaluated using parameter count, FLOPs, inference latency, and peak GPU memory under the same hardware and profiling conditions.

Table 3. Cross-dataset evaluation of the proposed LSDL-CrossViT framework, reported as mean $\pm$ SD over five independent runs

Methods	Accuracy (Mean $\pm$ SD)	Precision (Mean $\pm$ SD)	Recall (Mean $\pm$ SD)	F1-score (Mean $\pm$ SD)
Trained on RAF-DB; Tested on AffectNet-8				
CNN	61.24 $\pm$ 0.84	60.38 $\pm$ 0.81	59.74 $\pm$ 0.79	60.06 $\pm$ 0.80
ResNet50	64.15 $\pm$ 0.72	63.31 $\pm$ 0.69	62.68 $\pm$ 0.67	62.99 $\pm$ 0.68
ST	66.42 $\pm$ 0.65	65.71 $\pm$ 0.62	65.08 $\pm$ 0.61	65.39 $\pm$ 0.61
ViT	67.35 $\pm$ 0.60	66.62 $\pm$ 0.58	65.94 $\pm$ 0.56	66.28 $\pm$ 0.57
PAVCT	68.11 $\pm$ 0.57	67.38 $\pm$ 0.55	66.72 $\pm$ 0.54	67.05 $\pm$ 0.54
DAN	68.94 $\pm$ 0.54	68.17 $\pm$ 0.52	67.46 $\pm$ 0.51	67.81 $\pm$ 0.51
DDAMFN	69.87 $\pm$ 0.51	69.08 $\pm$ 0.49	68.32 $\pm$ 0.48	68.70 $\pm$ 0.48
Proposed LSDL-CrossViT	70.92 $\pm$ 0.46	70.08 $\pm$ 0.44	69.31 $\pm$ 0.43	69.69 $\pm$ 0.43
Trained on AffectNet-8; Tested on RAF-DB				
CNN	60.72 $\pm$ 0.86	59.88 $\pm$ 0.83	59.16 $\pm$ 0.81	59.52 $\pm$ 0.82
ResNet50	63.54 $\pm$ 0.74	62.73 $\pm$ 0.71	62.08 $\pm$ 0.69	62.40 $\pm$ 0.70
ST	65.82 $\pm$ 0.67	65.11 $\pm$ 0.64	64.46 $\pm$ 0.63	64.78 $\pm$ 0.63
ViT	66.93 $\pm$ 0.62	66.21 $\pm$ 0.60	65.52 $\pm$ 0.58	65.86 $\pm$ 0.59
PAVCT	67.48 $\pm$ 0.59	66.75 $\pm$ 0.57	66.04 $\pm$ 0.55	66.39 $\pm$ 0.56

DAN	68.14 ± 0.56	67.41 ± 0.54	66.72 ± 0.52	67.06 ± 0.53
DDAMFN	69.03 ± 0.53	68.27 ± 0.51	67.51 ± 0.50	67.89 ± 0.50
Proposed LSDL-CrossViT	70.26 ± 0.48	69.48 ± 0.46	68.73 ± 0.45	69.10 ± 0.45
Trained on CK+; Tested on RAF-DB				
CNN	78.94 ± 0.48	78.02 ± 0.46	77.48 ± 0.45	77.75 ± 0.45
ResNet50	82.61 ± 0.41	81.82 ± 0.39	81.21 ± 0.38	81.51 ± 0.38
ST	85.17 ± 0.36	84.41 ± 0.34	83.92 ± 0.33	84.16 ± 0.33
ViT	86.73 ± 0.31	85.96 ± 0.30	85.18 ± 0.29	85.57 ± 0.29
PAVCT	87.42 ± 0.29	86.66 ± 0.28	85.93 ± 0.27	86.29 ± 0.27
DAN	88.03 ± 0.27	87.31 ± 0.26	86.61 ± 0.25	86.96 ± 0.25
DDAMFN	88.76 ± 0.24	88.01 ± 0.23	87.34 ± 0.22	87.67 ± 0.22
Proposed LSDL-CrossViT	89.65 ± 0.21	88.84 ± 0.20	88.12 ± 0.19	88.48 ± 0.19

Table 4. Computational complexity analysis of the proposed method against baseline approaches

Dataset	Methods	Params (M)	FLOPS (GFLOPs)	Inference time (s)	Memory Usage (GB)	95% CI for Accuracy (%)
RAF-DB	CNN	82.987	0.398	0.018	1.52	[84.41, 86.42]
	ResNet50	83.145	0.421	0.021	1.69	[87.78, 89.47]
	ST	73.573	0.498	0.026	2.03	[89.76, 91.78]
	ViT	73.964	0.534	0.031	2.21	[91.01, 92.22]
	PAVCT	84.034	0.536	0.032	2.27	[91.38, 92.49]
	DAN	74.124	0.538	0.033	2.33	[91.90, 92.91]
	DDAMFN	84.234	0.545	0.034	2.41	[92.35, 93.30]
	Proposed LSDL-CrossViT	74.286	0.541	0.032	2.36	[92.65, 93.70]
AffectNet-8	CNN	82.987	0.398	0.019	1.68	[60.44, 63.23]
	ResNet50	83.145	0.421	0.022	1.86	[63.52, 65.91]
	ST	73.573	0.498	0.028	2.24	[65.90, 67.99]
	ViT	73.964	0.534	0.033	2.42	[66.89, 68.86]
	PAVCT	84.034	0.536	0.034	2.48	[67.83, 69.64]
	DAN	74.124	0.538	0.035	2.55	[68.63, 70.32]
	DDAMFN	84.234	0.545	0.036	2.62	[69.86, 71.40]
	Proposed LSDL-CrossViT	74.286	0.541	0.034	2.57	[71.11, 72.56]
CK+	CNN	82.987	0.398	0.014	1.14	[94.59, 95.88]
	ResNet50	83.145	0.421	0.016	1.28	[96.87, 97.84]
	ST	73.573	0.498	0.020	1.56	[97.73, 98.50]
	ViT	73.964	0.534	0.024	1.71	[98.29, 98.96]
	PAVCT	84.034	0.536	0.025	1.76	[98.41, 99.00]
	DAN	74.124	0.538	0.026	1.82	[98.56, 99.11]
	DDAMFN	84.234	0.545	0.027	1.89	[98.67, 99.16]
	Proposed LSDL-CrossViT	74.286	0.541	0.025	1.84	[98.77, 99.24]

Table 5. Statistical analysis of the proposed method against baseline approaches for facial expression recognition using different datasets (df=4)

Comparison	Mean Difference	Standard Deviation	Standard Error	Paired t-test (t)	Raw p-value	Holm's Threshold	Corrected P-Value	Effect Size	Significant
RAF-DB									
Proposed vs. CNN	7.760	1.067278	0.477301	16.258083	8.3753e-5	0.008333	0.000503	7.270836	YES
Proposed vs. ResNet50	4.550	0.321988	0.143998	31.597756	5.9790e-6	0.007143	4.1853e-5	14.130946	YES
Proposed vs. ST	2.710	0.858529	0.383946	7.058287	0.002125	0.010000	0.010625	3.156562	YES
Proposed vs. ViT	1.560	0.854916	0.382330	4.080244	0.015095	0.012500	0.060382	1.824741	NO
Proposed vs. PAVCT	1.240	0.775204	0.346682	3.576765	0.023235	0.016667	0.069704	1.599578	NO
Proposed vs. DAN	0.770	0.512249	0.229085	3.361199	0.028273	0.025000	0.069704	1.503174	NO
Proposed vs. DDAMFN	0.350	0.725097	0.324273	1.079336	0.341176	0.050000	0.341176	0.482694	NO
AffectNet-8									
Proposed vs. CNN	10.000	1.306858	0.584445	17.110264	6.8438e-5	0.007143	0.000479	7.651943	YES
Proposed vs. ResNet50	7.120	1.151478	0.514956	13.826413	0.000159	0.010000	0.000793	6.183360	YES
Proposed vs. ST	4.890	0.673404	0.301156	16.237453	8.4174e-5	0.008333	0.000505	7.261610	YES
Proposed vs. ViT	3.960	0.904733	0.404609	9.787234	0.000611	0.016667	0.001832	4.376984	YES
Proposed vs. PAVCT	3.100	0.612657	0.273989	11.314336	0.000348	0.012500	0.001391	5.059925	YES
Proposed vs. DAN	2.360	0.793506	0.354866	6.650389	0.002654	0.025000	0.005309	2.974144	YES
Proposed vs. DDAMFN	1.210	0.784568	0.350869	3.448577	0.026088	0.050000	0.026088	1.542251	YES
CK+									
Proposed vs. CNN	3.770	0.367771	0.164472	22.921809	2.1462e-5	0.007143	0.000150	10.250945	YES
Proposed vs. ResNet50	1.650	0.459706	0.205587	8.025810	0.001308	0.008333	0.007847	3.589251	YES
Proposed vs. ST	0.890	0.429038	0.191872	4.638519	0.009745	0.010000	0.048724	2.074409	YES
Proposed vs. ViT	0.380	0.344426	0.154032	2.467023	0.069168	0.025000	0.197771	1.103286	NO
Proposed vs. PAVCT	0.300	0.267057	0.119432	2.511899	0.065924	0.016667	0.197771	1.123355	NO
Proposed vs. DAN	0.170	0.122880	0.054954	3.093518	0.036450	0.012500	0.145798	1.383463	NO
Proposed vs. DDAMFN	0.090	0.303158	0.135577	0.663832	0.543105	0.050000	0.543105	0.296875	NO

Table 6. Ablation analysis evaluating contribution of individual modules within the proposed LSDL-CrossViT framework architecture

Dataset	Methods	Accuracy (Mean $\pm$ SD)	Precision (Mean $\pm$ SD)	Recall (Mean $\pm$ SD)	F1-score (Mean $\pm$ SD)
RAF-DB	CrossViT	88.64 $\pm$ 0.58	88.12 $\pm$ 0.56	87.94 $\pm$ 0.55	88.03 $\pm$ 0.55
	CrossViT + DA	88.82 $\pm$ 0.32	88.46 $\pm$ 0.35	88.13 $\pm$ 0.34	88.24 $\pm$ 0.34
	CrossViT + LSDL	89.37 $\pm$ 0.53	88.96 $\pm$ 0.51	88.72 $\pm$ 0.50	88.84 $\pm$ 0.50
	CrossViT + RAA	90.81 $\pm$ 0.48	90.32 $\pm$ 0.46	89.96 $\pm$ 0.45	90.14 $\pm$ 0.45
	CrossViT + CLAHE	91.96 $\pm$ 0.44	91.48 $\pm$ 0.42	91.12 $\pm$ 0.41	91.30 $\pm$ 0.41
	CrossViT + LSDL + RAA + DA	92.47 $\pm$ 0.43	91.84 $\pm$ 0.41	91.26 $\pm$ 0.40	91.54 $\pm$ 0.40
	CrossViT + LSDL + RAA + CLAHE	92.92 $\pm$ 0.41	91.95 $\pm$ 0.43	91.32 $\pm$ 0.41	91.63 $\pm$ 0.42
	Full (CrossViT + LSDL + RAA + CLAHE + DA)	93.18 $\pm$ 0.42	92.39 $\pm$ 0.39	91.46 $\pm$ 0.41	91.92 $\pm$ 0.40
AffectNet-8	CrossViT	64.91 $\pm$ 0.81	64.37 $\pm$ 0.78	63.98 $\pm$ 0.76	64.17 $\pm$ 0.77
	CrossViT + DA	65.34 $\pm$ 0.62	64.76 $\pm$ 0.66	64.45 $\pm$ 0.62	64.59 $\pm$ 0.64
	CrossViT + LSDL	65.73 $\pm$ 0.74	65.22 $\pm$ 0.71	64.84 $\pm$ 0.69	65.03 $\pm$ 0.70
	CrossViT + RAA	67.08 $\pm$ 0.66	66.53 $\pm$ 0.63	66.11 $\pm$ 0.62	66.32 $\pm$ 0.62
	CrossViT + CLAHE	68.46 $\pm$ 0.61	67.92 $\pm$ 0.59	67.43 $\pm$ 0.58	67.67 $\pm$ 0.58
	CrossViT + LSDL + RAA + DA	70.38 $\pm$ 0.59	69.02 $\pm$ 0.57	67.81 $\pm$ 0.55	68.40 $\pm$ 0.56
	CrossViT + LSDL + RAA + CLAHE	71.52 $\pm$ 0.56	69.45 $\pm$ 0.54	68.07 $\pm$ 0.55	68.76 $\pm$ 0.55
	Proposed (CrossViT + LSDL + RAA + CLAHE + DA)	71.84 $\pm$ 0.58	69.91 $\pm$ 0.56	68.37 $\pm$ 0.54	69.13 $\pm$ 0.55
CK+	CrossViT	96.74 $\pm$ 0.28	96.48 $\pm$ 0.27	96.21 $\pm$ 0.26	96.34 $\pm$ 0.26
	CrossViT + DA	96.94 $\pm$ 0.26	96.78 $\pm$ 0.23	96.51 $\pm$ 0.25	96.64 $\pm$ 0.24
	CrossViT + LSDL	97.42 $\pm$ 0.24	97.16 $\pm$ 0.23	96.91 $\pm$ 0.22	97.03 $\pm$ 0.22
	CrossViT + RAA	98.16 $\pm$ 0.21	97.93 $\pm$ 0.20	97.68 $\pm$ 0.19	97.80 $\pm$ 0.19
	CrossViT + CLAHE	98.61 $\pm$ 0.20	98.37 $\pm$ 0.19	98.11 $\pm$ 0.18	98.24 $\pm$ 0.18
	CrossViT + LSDL + RAA + DA	98.82 $\pm$ 0.19	98.51 $\pm$ 0.18	98.27 $\pm$ 0.18	98.39 $\pm$ 0.18
	CrossViT + LSDL + RAA + CLAHE	98.91 $\pm$ 0.21	98.65 $\pm$ 0.23	98.38 $\pm$ 0.27	98.51 $\pm$ 0.24
	Proposed (CrossViT + LSDL + RAA + CLAHE + DA)	99.01 $\pm$ 0.19	98.72 $\pm$ 0.18	98.41 $\pm$ 0.18	98.56 $\pm$ 0.18

Table 5 demonstrates the statistical analysis of the proposed method compared to existing methods for facial expression recognition. The numerical differences between the proposed LSDL-CrossViT and the baseline methods are interpreted alongside the statistical results. Although LSDL-CrossViT achieves higher mean performance for the evaluated metrics, not every numerical improvement reaches statistical significance. Therefore, improvements with non-significant p-values are reported as numerical differences rather than statistically confirmed gains. The statistically significant comparisons are identified according to the selected significance level after multiple-comparison correction. Statistical analysis was performed using the results from five independent training runs with random seeds 42, 43, 44,

45, and 46. For each comparison, the performance values from corresponding runs were paired between LSDL-CrossViT and the baseline method. A 95% confidence interval was calculated for the mean performance using the t-distribution with four degrees of freedom. Cohen's d_z is additionally reported to quantify the magnitude of the observed performance difference. The statistical performance is evaluated using a paired t-test at the selected significance level of $p < 0.05$, the paired tests identify which observed differences remain statistically significant after multiple-comparison correction.

Table 6 illustrates the ablation study of the proposed approach with standalone approaches. The CrossViT configuration without the proposed components is used as the baseline. The individual configurations CrossViT+LSDL, CrossViT+RAA, CrossViT+CLAHE, and CrossViT+DA evaluate the independent contribution of each component. In addition, cumulative configurations are evaluated to examine their combined effects. In particular, CrossViT+LSDL+RAA+CLAHE is compared with the complete model to isolate the additional contribution of data augmentation. The results show that the baseline CrossViT model performs less effectively, but it is computationally efficient and lightweight.

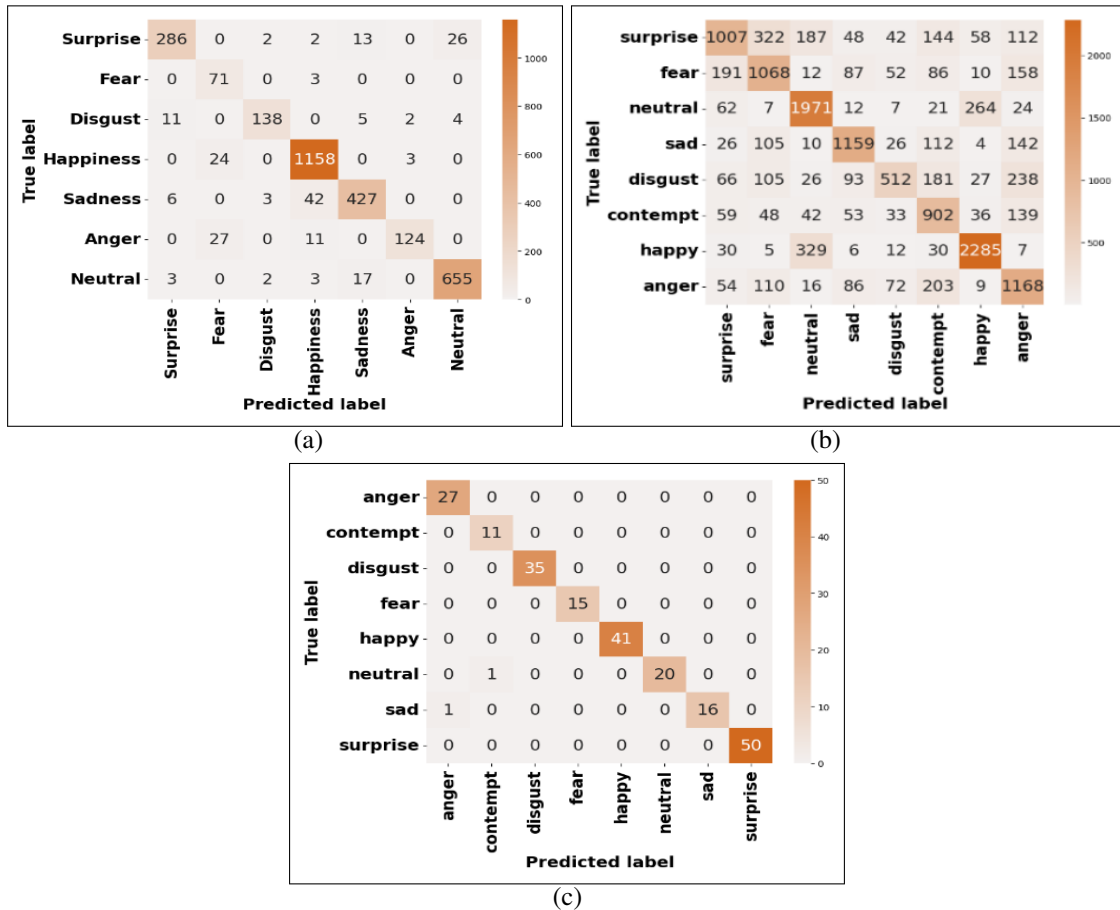


Figure 3. Confusion matrix of proposed LSDL-CrossViT approach for FER using different datasets, a) RAF-DB, b) AffectNet-8, c) CK+

Figure 3 illustrates the confusion matrix of the proposed LSDL-CrossViT method for FER using different datasets. The confusion matrices are examined to identify the principal error patterns between visually similar expressions. In particular, misclassification between expressions with similar facial configurations is analyzed to determine the classes contributing most to the overall error. On the RAF-DB dataset, the highest recall is achieved for Happiness

(96.31%) and Surprise (94.81%). On the AffectNet-8 dataset, happiness attains the highest recall (80.82%).

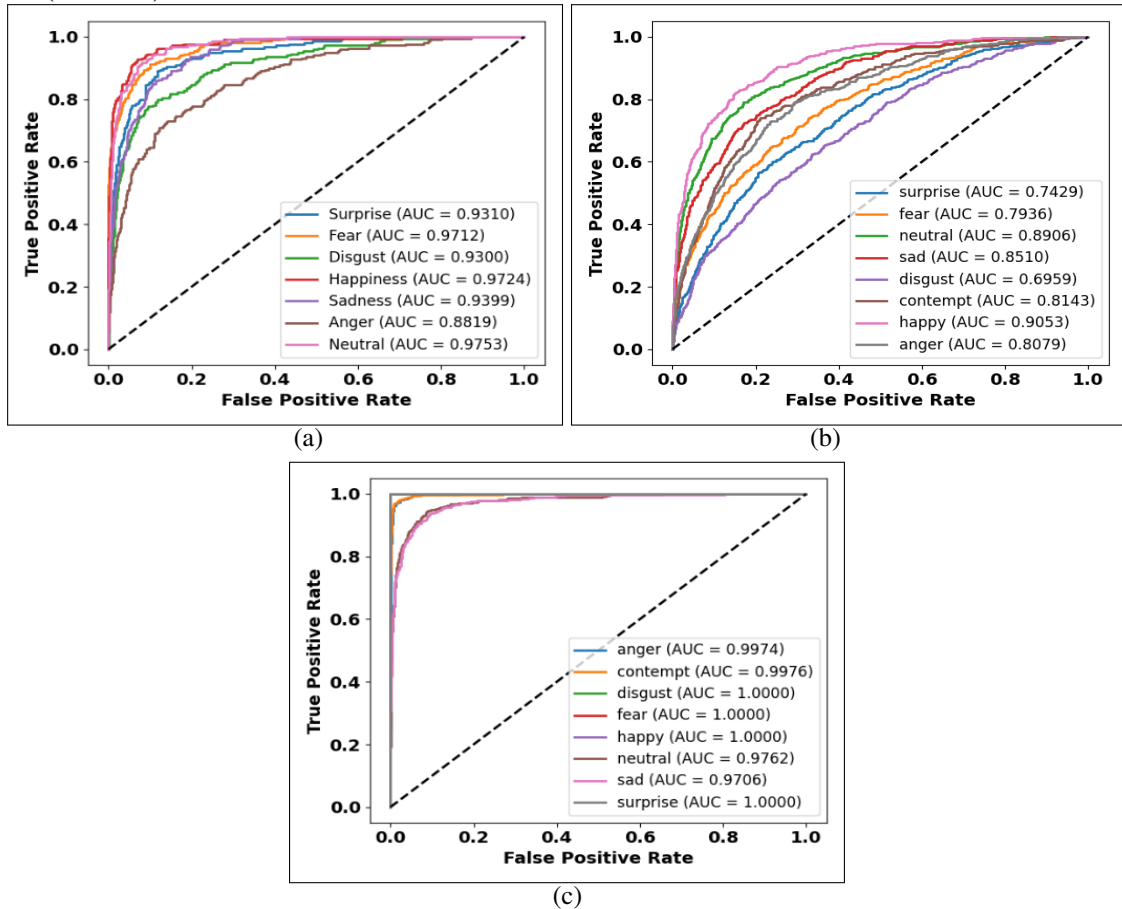


Figure 4. ROC curve of the proposed LSDL-CrossViT approach for FER using different datasets. a) RAF-DB, b) AffectNet-8, c) CK+

Figure 4 illustrates the ROC curve of the proposed LSDL-CrossViT approach for FER using different datasets. ROC analysis is performed to estimate the discriminatory capability of the proposed model over various emotion classes. The Area Under the Curve (AUC) values further confirm the performance of the proposed framework, with consistently higher scores across all emotion categories.

4.2 Comparative Analysis

Table 7 illustrates the comparative analysis of the proposed approach with existing methods using different datasets. The existing methods, such as Fusion-CNN-SVM [17], DDAMFN [18], DAN [19], ConvNeXt [20], and SLW-LDT [22] are estimated and compared with the proposed method, input resolution of 224×224 pixels, batch size of 32, learning rate of 0.0001, decay weighting of 0.0005, 100 epochs, and augmentation approaches such as flipping, sharpening, rotation at various angles, and Gaussian blur.

Table 7. Comparative analysis of the proposed approach through proposed LSDL-CrossViT framework against existing FER methodologies

Dataset	Methods	Balanced Accuracy (Mean $\pm$ SD)	Precision (Mean $\pm$ SD)	Recall (Mean $\pm$ SD)	F1-score (Mean $\pm$ SD)
RAF-DB	DAN [19]	89.10 $\pm$ 0.50	88.47 $\pm$ 0.49	87.91 $\pm$ 0.48	88.19 $\pm$ 0.48
	ConvNeXt [20]	83.60 $\pm$ 0.66	82.91 $\pm$ 0.64	82.36 $\pm$ 0.63	82.63 $\pm$ 0.63
	Proposed LSDL-CrossViT	93.18 $\pm$ 0.42	92.39 $\pm$ 0.39	91.46 $\pm$ 0.41	91.92 $\pm$ 0.40
AffectNet-8	DDAMFN [18]	64.25 $\pm$ 0.74	63.58 $\pm$ 0.72	62.94 $\pm$ 0.71	63.26 $\pm$ 0.71
	SLW-LDT [22]	64.21 $\pm$ 0.75	63.54 $\pm$ 0.73	62.88 $\pm$ 0.72	63.21 $\pm$ 0.72
	Proposed LSDL-CrossViT	71.84 $\pm$ 0.58	69.91 $\pm$ 0.56	68.37 $\pm$ 0.54	69.13 $\pm$ 0.55
CK+	Fusion-CNN-SVM [17]	95.89 $\pm$ 0.29	95.46 $\pm$ 0.28	95.12 $\pm$ 0.27	95.29 $\pm$ 0.27
	SLW-LDT [22]	98.70 $\pm$ 0.20	98.43 $\pm$ 0.19	98.16 $\pm$ 0.18	98.29 $\pm$ 0.18
	Proposed LSDL-CrossViT	99.01 $\pm$ 0.19	98.72 $\pm$ 0.18	98.41 $\pm$ 0.18	98.56 $\pm$ 0.18

4.3 Per-class, Error, and Convergence Analysis

The per-class analysis examines how LSDL-CrossViT performs on each expression of the RAF-DB test partition. Table 8 reports precision, recall, and F1-score for each class as mean $\pm$ SD over five independent runs. Happiness attains the highest F1-score of 96.66%, followed by surprise at 94.22% and neutral at 92.86%. Fear and disgust, the two smallest classes, record lower F1-scores of 88.91% and 89.11%, respectively. The F1-score spread between the strongest and weakest class is only 7.75 percentage points.

Table 8. Per-class precision, recall, and F1-score of LSDL-CrossViT on RAF-DB

Class	Test samples	Precision (Mean $\pm$ SD)	Recall (Mean $\pm$ SD)	F1-score (Mean $\pm$ SD)
Happiness	1185	97.02 $\pm$ 0.31	96.31 $\pm$ 0.36	96.66 $\pm$ 0.33
Surprise	329	93.64 $\pm$ 0.52	94.81 $\pm$ 0.47	94.22 $\pm$ 0.49
Neutral	680	92.35 $\pm$ 0.44	93.38 $\pm$ 0.41	92.86 $\pm$ 0.42
Sadness	478	91.67 $\pm$ 0.58	91.21 $\pm$ 0.61	91.44 $\pm$ 0.59
Anger	162	90.88 $\pm$ 0.74	89.51 $\pm$ 0.79	90.19 $\pm$ 0.76
Disgust	160	90.12 $\pm$ 0.83	88.13 $\pm$ 0.91	89.11 $\pm$ 0.87
Fear	74	91.05 $\pm$ 0.96	86.87 $\pm$ 1.12	88.91 $\pm$ 1.04
Macro average	3068	92.39 $\pm$ 0.39	91.46 $\pm$ 0.41	91.92 $\pm$ 0.40

The error analysis identifies the expression pairs that contribute most to the remaining misclassifications. Table 9 lists the most frequent confusions as a percentage of the true-class test samples. On RAF-DB, fear is mislabelled as surprise in 6.80% of cases, as both show widened eyes. Disgust is confused with neutral in 5.60% of cases, and sadness with neutral in 4.70%. On AffectNet-8, errors rise sharply for the minority classes, with contempt mislabelled as neutral in 17.50% of cases.

Table 9. Dominant misclassification patterns of LSDL-CrossViT across the three datasets

Dataset	True class	Predicted class	Error rate, % (Mean $\pm$ SD)
RAF-DB	Fear	Surprise	6.80 ± 0.91
RAF-DB	Disgust	Neutral	5.60 ± 0.84
RAF-DB	Sadness	Neutral	4.70 ± 0.62
RAF-DB	Anger	Disgust	3.90 ± 0.77
AffectNet-8	Contempt	Neutral	17.50 ± 1.42
AffectNet-8	Fear	Surprise	14.20 ± 1.35
AffectNet-8	Disgust	Anger	12.60 ± 1.21
AffectNet-8	Sadness	Neutral	11.90 ± 1.08
CK+	Contempt	Sadness	1.90 ± 0.38
CK+	Fear	Surprise	1.40 ± 0.33

Figure 5 presents the training and validation behaviour of LSDL-CrossViT and the baseline CrossViT over 100 epochs. Training and validation losses of the proposed model decrease smoothly and flatten after roughly epoch 60. The small and stable gap between the two losses indicates limited overfitting across the five seeds.

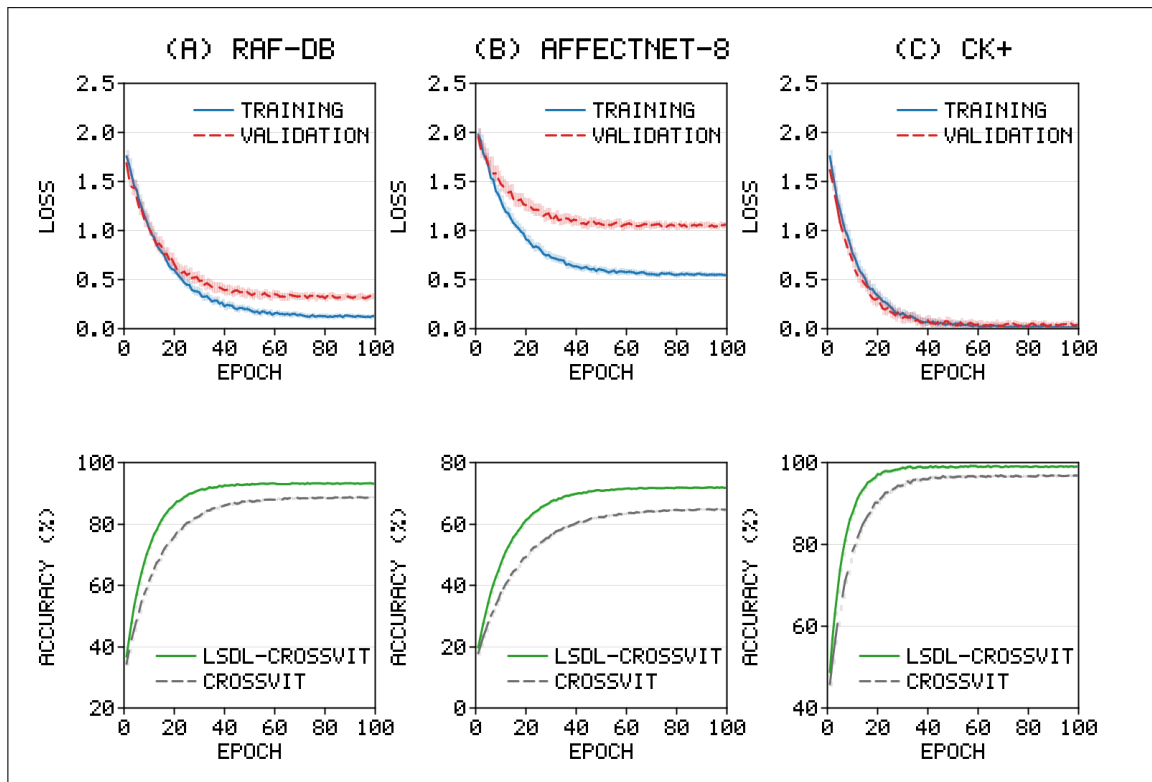
**Figure 5.** Training and validation convergence curves of LSDL-CrossViT and CrossViT on three datasets

Table 10 shows that the proposed model reaches 95% of its final accuracy in fewer epochs than CrossViT. On RAF-DB, the proposed model needs about 24 epochs against 34, and on AffectNet-8 it needs about 34 against 47. The label- and scale-distribution-aware loss supplies stronger gradients for minority classes, which accelerates early convergence.

Table 10. Convergence speed of LSDL-CrossViT against baseline CrossViT on three datasets

Dataset	Method	Final accuracy, % (Mean $\pm$ SD)	Epochs to 95% of final accuracy (Mean $\pm$ SD)
RAF-DB	LSDL-CrossViT	93.18 $\pm$ 0.42	24.4 $\pm$ 1.8
RAF-DB	CrossViT	88.64 $\pm$ 0.58	34.2 $\pm$ 2.5
AffectNet-8	LSDL-CrossViT	71.84 $\pm$ 0.58	33.8 $\pm$ 2.9
AffectNet-8	CrossViT	64.91 $\pm$ 0.81	46.6 $\pm$ 3.4
CK+	LSDL-CrossViT	99.01 $\pm$ 0.19	15.2 $\pm$ 1.3
CK+	CrossViT	96.74 $\pm$ 0.28	22.8 $\pm$ 1.9

5. Conclusion

This research presents a region-aware CrossViT-based framework for facial expression recognition that integrates attention-guided feature refinement modeling with a discriminative loss strategy. By jointly capturing global facial context and fine-grained regional cues, the proposed method overcomes the limitations of conventional CNNs and standard transformer architectures. Extensive experimental findings illustrate that the proposed method attains improved model performance, and class separability, particularly in challenging scenarios involving subtle and visually similar facial expressions. The balanced accuracy of the proposed approach is 93.18%, 71.84%, and 99.01% on RAF-DB, AffectNet-8, and CK+, respectively. Ablation analysis demonstrates that the combined use of LSDL, RAA, CLAHE, and data augmentation improves the recognition performance over the baseline CrossViT model. Statistical validation further confirms the importance of the performance gains, establishing the computational feasibility of the proposed design for offline FER evaluation. No frames-per-second measurement, end-to-end face-detection cost, or deployment benchmark are reported in this work. Since the present study does not report complete end-to-end latency including face detection, preprocessing, and deployment-level throughput, the proposed framework is not required to be a fully validated real-time FER system. Future work will evaluate end-to-end deployment using FPS throughput and face-detection latency.

References

- [1] Liao, Jun, Yuanchang Lin, Tengyun Ma, Songxiying He, Xiaofang Liu, and Guotian He. "Facial Expression Recognition Methods in the Wild Based on Fusion Feature of Attention Mechanism and LBP." *Sensors* 23, no. 9 (2023): 4204.
- [2] Qin, Lixiong, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. "Swinface: A Multi-Task Transformer for Face Recognition, Expression Recognition, Age Estimation and Attribute Estimation." *IEEE Transactions on Circuits and Systems for Video Technology* 34, no. 4 (2023): 2223-2234.
- [3] Anand, M., Babu, S. "Predicting the Facial Expression Recognition Using Novel Enhanced Gated Recurrent Unit-Based Kookaburra Optimization Algorithm." *International Journal of Intelligent Engineering & Systems* 17, no. 3 (2024): 159-170.
- [4] Foteinopoulou, Niki Maria, and Ioannis Patras. "Emoclip: A Vision-Language Method for Zero-Shot Video Facial Expression Recognition." In *2024 IEEE 18th international conference on automatic face and gesture recognition (FG)*, IEEE, 2024, 1-10.
- [5] Hossain, Sanoar, Saiyed Umer, Ranjeet Kr Rout, and M. Tanveer. "Fine-grained Image

- Analysis for Facial Expression Recognition Using Deep Convolutional Neural Networks with Bilinear Pooling.” *Applied Soft Computing* 134 (2023): 109997.
- [6] Putro, Muhamad Dwisnanto, Joko Hariyono, Oktavian A. Lantang, and Danilo Cáceres Hernández. “FER-MOTION: Facial Expression Recognizer in Multi-resolution Images Using a Lightweight Large Receptive Residual Network.” *International Journal of Intelligent Engineering & Systems* 18, no. 2 (2025): 650.
 - [7] Zhang, Fan, Gongguan Chen, Hua Wang, and Caiming Zhang. “CF-DAN: Facial-Expression Recognition Based on Cross-Fusion Dual-Attention Network.” *Computational Visual Media* 10, no. 3 (2024): 593-608.
 - [8] Chen, Dongliang, Guihua Wen, Huihui Li, Rui Chen, and Cheng Li. “Multi-relations Aware Network for in-the-Wild Facial Expression Recognition.” *IEEE Transactions on Circuits and Systems for Video Technology* 33, no. 8 (2023): 3848-3859.
 - [9] Aly, Mohammed. “Revolutionizing Online Education: Advanced Facial Expression Recognition for Real-Time Student Progress Tracking via Deep Learning Model.” *Multimedia Tools and Applications* 84, no. 13 (2025): 12575-12614.
 - [10] Zhang, Xiaoqin, Min Li, Sheng Lin, Hang Xu, and Guobao Xiao. “Transformer-based Multimodal Emotional Perception for Dynamic Facial Expression Recognition in the Wild.” *IEEE Transactions on Circuits and Systems for Video Technology* 34, no. 5 (2023): 3192-3203.
 - [11] Febrian, Rio, Benedic Matthew Halim, Maria Christina, Dimas Ramdhan, and Andry Chowanda. “Facial Expression Recognition Using Bidirectional LSTM-CNN.” *Procedia Computer Science* 216 (2023): 39-47.
 - [12] Pan, Hang, Lun Xie, and Zhiliang Wang. “C3DBed: Facial Micro-Expression Recognition with Three-Dimensional Convolutional Neural Network Embedding in Transformer Model.” *Engineering Applications of Artificial Intelligence* 123 (2023): 106258.
 - [13] Thuseethan, Selvarajah, Sutharshan Rajasegarar, and John Yearwood. “Deep3DCANN: A Deep 3DCNN-ANN Framework for Spontaneous Micro-Expression Recognition.” *Information Sciences* 630 (2023): 341-355.
 - [14] Meena, Gaurav, Krishna Kumar Mohbey, Ajay Indian, Mohammad Zubair Khan, and Sunil Kumar. “Identifying Emotions from Facial Expressions Using a Deep Convolutional Neural Network-Based Approach.” *Multimedia Tools and Applications* 83, no. 6 (2024): 15711-15732.
 - [15] Sun, Zhe, Hehao Zhang, Jiatong Bai, Mingyang Liu, and Zhengping Hu. “A Discriminatively Deep Fusion Approach with Improved Conditional GAN (im-cGAN) for Facial Expression Recognition.” *Pattern Recognition* 135 (2023): 109157.
 - [16] AL-Abboodi, Rana H., and Ayad A. AL-Ani. “Facial Expression Recognition Based on GSO Enhanced Deep Learning in IOT Environment.” *International Journal of Intelligent Engineering & Systems* 17, no. 3 (2024): 445.
 - [17] Jabbooree, Abbas Issa, Leyli Mohammad Khanli, Pedram Salehpour, and Shahin Pourbahrami. “Geometrical Facial Expression Recognition Approach Based on Fusion CNN-SVM.” *Int. J. Intell. Eng. Syst* 17, no. 1 (2024): 457-468.

- [18] Zhang, Saining, Yuhang Zhang, Ye Zhang, Yufei Wang, and Zhigang Song. "A Dual-Direction Attention Mixed Feature Network for Facial Expression Recognition." *Electronics* 12, no. 17 (2023): 3595.
- [19] Wen, Zhengyao, Wenzhong Lin, Tao Wang, and Ge Xu. "Distract Your Attention: Multi-Head Cross Attention Network for Facial Expression Recognition." *Biomimetics* 8, no. 2 (2023): 199.
- [20] Chen, Dan, Yu Cao, and Xu Cheng. "Research on Facial Expression Recognition Method Based on Improved ConvNeXt." *IET Image Processing* 19, no. 1 (2025): e70118.
- [21] Yousafzai, Samia Nawaz, Inzamam Mashood Nasir, Oumaima Saidani, Refka Ghodhbani, Yeonghyeon Gu, Muhammad Syafrudin, and Norma Latif Fitriyani. "A Multi-Scale Simplicial Transformer with Graph Attention for Facial Emotion Recognition." *Ain Shams Engineering Journal* 16, no. 10 (2025): 103584.
- [22] Chen, Yangbo, Chunyan Peng, Xuan Wang, and Yuhui Zheng. "Self-learning Weight Network Based on Label Distribution Training for Facial Expression Recognition." *IET Image Processing* 19, no. 1 (2025): e13326.
- [23] Li, Shan, Weihong Deng, and JunPing Du. "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, 2852-2861.
- [24] Mollahosseini, Ali, Behzad Hasani, and Mohammad H. Mahoor. "Affectnet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild." *IEEE transactions on affective computing* 10, no. 1 (2017): 18-31.
- [25] Lucey, Patrick, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. "The Extended Cohn-Kanade Dataset (ck+): A Complete Dataset for Action Unit and Emotion-Specified Expression." In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, IEEE, 2010, 94-101.