Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.086137

ARTICLE

# Dual-Stream Facial Emotion Recognition with Self-Supervised Pre-Training and Evidential Uncertainty

**Rashid Jahangir**[1,*], **Nazik Alturki**[2] **and Mohammed Alreshoodi**[3]

[1]Department of Computer Science, COMSATS University Islamabad, Vehari Campus, Vehari, Pakistan
[2]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[3]Unit of Scientific Research, Applied College, Qassim University, Buraydah, Saudi Arabia

*Corresponding Author: Rashid Jahangir. Email: rashidjahangir@cuivehari.edu.pk

Received: 25 May 2026; Accepted: 07 September 2026; Published: 28 September 2026

**ABSTRACT:** Facial emotion recognition (FER) remains difficult in real-world settings. Inter-subject variability, lighting changes, occlusion, and class imbalance all limit performance. Most FER systems rely on one convolutional or transformer backbone. This narrows the features available for classification. This paper presents Dual-Stream FERNet. It is a carefully evaluated integration of an EfficientNetV2-S backbone with a Swin Transformer Tiny backbone, joined by a learnable sigmoid-gated fusion module. Before fine-tuning, both branches undergo SimCLR-style self-supervised pre-training on two augmented views. This gives a stronger initialization without extra labels. An Evidential Deep Learning head then produces class probabilities and Dirichlet-parameterized uncertainty together. The model is tested on two benchmarks, KDEF and CK+. Under subject-disjoint 5-fold cross-validation, the model reached 93.84 ± 1.73% accuracy on KDEF and 93.07 ± 1.22% on CK+. The model is compared against fair, SSL-matched baselines on KDEF, ResNet50, EfficientNet-B0, and Swin-Small. The model's real advantage is calibrated uncertainty, not higher accuracy. On KDEF the model runs at 459.67 FPS (NVIDIA RTX 4070, FP32, batch size 16, batch-1 median latency 18.96 ms). Grad-CAM shows the model attending to facial regions tied to FACS action units on both datasets. Overall, the model matches strong single-stream baselines in accuracy and adds calibrated uncertainty on top.

**KEYWORDS:** Facial emotion recognition; dual-stream CNN-transformer; EfficientNetV2-S; swin transformer; sigmoid-gated feature fusion; self-supervised learning

## 1 Introduction

Automatic facial emotion recognition (FER) is a core problem in computer vision and affective computing. Reading emotional states from facial images supports applications such as online gaming, patient monitoring, mental health assessment, and driver fatigue detection [1]. Yet robust FER under realistic, unconstrained conditions is still an open problem [2]. The same emotion can look different across individuals due to differences in facial muscle structure, cultural display rules, and expressiveness. Real-world images add further difficulties. Lighting changes, occlusion from accessories, and non-frontal poses all distort the spatial layout of facial action units [3,4]. Class imbalance compounds these problems, since neutral and happy expressions appear far more often than fear or disgust in most datasets, biasing classifiers trained with standard cross-entropy (CE) loss toward the majority classes [5].

Early FER methods relied on hand-crafted descriptors—Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), and Gabor filter banks—paired with Support Vector Machines or Random

Forest classifiers. These methods are interpretable but require careful feature engineering and generalize poorly across imaging conditions. Deep Convolutional Neural Networks (CNNs) advanced the field by learning hierarchical representations directly from pixels [6]. Still, CNNs trained with plain cross-entropy loss face three recurring problems in FER. They overfit easily when labeled data is scarce, they produce overconfident softmax scores that poorly reflect true uncertainty, and they rely on a single backbone that captures only one type of feature representation. Standard convolutional architectures extract local convolutional features but lack any self-attention mechanism, while transformer architectures capture global self-attention relationships but lack an inductive bias toward local convolutional structure. Neither captures both simultaneously [7,8].

Transformer-based models have recently shown strong FER performance by capturing long-range dependencies between distant facial action units [9,10]. But pure transformers need large labeled datasets and heavy compute. Lightweight CNNs, such as EfficientNet variants, run faster but often miss global relationships between distant facial regions [11]. Combining both paradigms in one framework remains underexplored for FER [4,8]. Self-supervised learning (SSL) offers one way to ease data scarcity by pre-training backbones on unlabeled data through pretext tasks. SimCLR-style contrastive learning [12] trains a network to pull together two augmented views of the same image in embedding space, producing representations that transfer well to downstream classification. For FER specifically, SSL pre-training on the target dataset's unlabeled images adapts the backbone more closely to facial image statistics than generic ImageNet pre-training does [13,14].

Standard FER classifiers also cannot quantify uncertainty. A model that assigns 95% confidence to "happy" for a subtly ambiguous micro-expression looks identical, on paper, to one that assigns 95% confidence to a clearly happy face. Evidential Deep Learning (EDL) addresses this by modeling class probabilities as a Dirichlet distribution parameterized by evidence, where low total evidence signals high uncertainty [15]. This uncertainty can be used during training to down-weight mislabeled or ambiguous samples, and during deployment to flag cases for human review [16]. This paper proposes Dual-Stream FERNet, a unified architecture that addresses all four challenges above. The contributions of this study are as follows.

- A dual-stream backbone pairing EfficientNetV2-S (speed, local CNN features) with Swin-T (global self-attention features), fused via a learnable sigmoid-gated feature fusion module that computes a per-dimension gating vector and produces a dynamically weighted combination of both branch outputs—a gated weighted sum rather than a query-key-value attention operation.
- A dual-branch SimCLR-style SSL pre-training phase applied jointly to both backbones as lightweight domain adaptation of ImageNet pre-trained representations toward the FER distribution, prior to supervised fine-tuning.
- An Evidential Deep Learning (EDL) classification head replacing softmax with Dirichlet-parameterized evidence outputs, providing calibrated uncertainty estimates used both as an inference signal and as a per-sample loss-weighting mechanism during training.

The remainder of this paper is organized as follows. Section 2 reviews related work on deep learning-based FER. Section 3 describes the proposed Dual-Stream FERNet architecture, SSL pre-training strategy, and training procedure. Section 4 presents experimental results and discussion. Section 5 concludes with directions for future work.

## 2 Literature Review

The evolution of FER systems spans three broad phases, hand-crafted feature methods, shallow neural networks, and deep learning-based approaches. This review focuses on deep learning-based FER methods,

organized around three technical dimensions central to current research, backbone architecture, training strategy, and uncertainty modelling.

### *2.1 Single-Backbone CNN and Hybrid Architectures*

Early deep learning FER systems established that CNNs trained end-to-end on facial image data substantially outperform hand-crafted feature approaches. A hierarchical deep neural network structure for FER was proposed [17] and achieved 96.5% accuracy on CK+ and 91.3% on JAFFE. The proposed method improved upon prior baselines by 1.3% and 1.5%, respectively. The approach reclassified top erroneous predictions using a hierarchical structure but operated on a single backbone without attention mechanisms. A hybrid CNN-RNN architecture was introduced [18] to jointly learn spatial features via CNN and temporal sequence patterns via RNN on the MMI and JAFFE datasets. This architecture achieved 94.91% accuracy with six hidden layers. An extension of this work replaced the temporal dataset with CK+ and incorporated deep residual blocks with a fully connected network. The extended research [19] achieved 93.24% and 95.23% accuracy on CK+ and JAFFE, respectively. These results demonstrate the benefit of combining spatial feature extraction with sequential modelling. However, the study did not address the limitations of a single convolutional stream.

A CNN combined with image edge detection was evaluated [20] on a mixed Fer-2013/LFW dataset. The model achieved 88.56% average accuracy with training speed 1.5 times faster than competing methods. An attentional convolutional network (Deep-Emotion) incorporating a spatial attention mechanism was applied to FER-2013, CK+, JAFFE, and FERG datasets. This demonstrated that explicit attention mechanisms substantially improved over standard CNN feature aggregation. Two CNN models for FER in the wild were proposed [21], a lightweight external network for industrial deployment and a unified network combining deep feature extraction with LBP learning in parallel.

### *2.2 Transfer Learning and Data Augmentation*

Transfer learning from large-scale image classification datasets remains the dominant way to address data scarcity in FER. Models such as VGG16, ResNet50, and InceptionV3, pre-trained on ImageNet, are commonly used as backbone initializations and then fine-tuned on FER datasets. However, the domain gap between natural image classification and facial expression recognition limits how well ImageNet-derived features transfer, particularly for subtle emotion categories such as fear and disgust.

Data augmentation has also been studied as a complementary strategy. A detailed comparison of augmentation techniques and deep learning features showed that augmentation meaningfully improves FER accuracy on the FER-2013, extended FER-2013, and AffectNet datasets [22]. Separately, fusing handcrafted and automatically extracted features on the same datasets yielded roughly 1% higher accuracy than standard approaches [23]. Both studies, however, used fixed, predefined augmentation pipelines and did not adapt augmentation strength to the target domain. A hierarchical weighted random forest (WRF) classifier for driver FER, evaluated on the CK+, KMU-FED, and MMI databases, reached 92.6% accuracy on CK+ [24].

Knowledge distillation offers another route to real-time FER deployment. A real-time emotion recognition system based on adaptive knowledge distillation [25] trains a lightweight student network to match the soft-probability outputs of a larger, pre-trained teacher model, achieving competitive accuracy at much lower inference cost. This differs fundamentally from architectural parallelism or model compression as a strategy for meeting real-time requirements.

### *2.3 Transformer-Based and Multi-Branch Architectures*

The introduction of Vision Transformers (ViT) and hierarchical Swin Transformers has established a new direction for facial expression recognition (FER) research. In contrast to convolutional architectures, transformers employ self-attention mechanisms that capture long-range dependencies between distant facial regions [26]. This capability enables more effective modeling of the interactions among multiple action units, including brow raising, lip corner pulling, and eye widening, which collectively define facial expressions.

Temporal transformer architectures have been utilized in video-based facial expression recognition (FER). STT-Net [27] introduces a simplified temporal transformer that reduces the computational complexity associated with full self-attention across video frame sequences, while maintaining competitive accuracy on sequence-level FER benchmarks. These temporal models are particularly effective for video-sequence FER, where the dynamics of expressions provide discriminative temporal information. In contrast, the current study addresses static single-image FER, a setting in which temporal information is absent and the quality of spatial features primarily determines classification performance.

EfficientNet variants have demonstrated state-of-the-art accuracy-efficiency trade-offs on image classification benchmarks and have been successfully applied to FER. However, pure EfficientNet architectures do not explicitly model global context. To date, no published studies have investigated dual-stream architectures that combine a convolutional backbone with a transformer backbone for FER using a learnable gated fusion mechanism. Additionally, the application of self-supervised contrastive pre-training (SimCLR) to FER has been restricted to single-branch settings [28]. This study addresses both gaps by introducing a gated EfficientNetV2-S and Swin-T fusion architecture with dual-branch self-supervised learning (SSL) pre-training.

### *2.4 Uncertainty Modelling in FER*

Standard FER classifiers produce softmax probability vectors that are systematically overconfident. They assign high probability to a single class even for genuinely ambiguous expressions. This overconfidence is problematic in safety-critical applications such as driver monitoring or clinical mental health screening, where awareness of model uncertainty is essential for appropriate system behavior. Evidential Deep Learning (EDL) provides a principled alternative by modelling the class probability vector as a Dirichlet distribution [29]. The concentration parameters of the Dirichlet are output by the network as evidence values, and the total evidence determines the precision of the distribution. Low total evidence corresponds to high uncertainty, providing a signal for genuinely ambiguous expressions. Out-of-distribution detection is not evaluated in this study and accordingly do not claim this uncertainty signal detects out-of-distribution inputs. EDL has been applied in image classification and medical imaging tasks but has not previously been applied to FER in a dual-stream gated fusion setting. No prior published study has combined Evidential Deep Learning with dual-backbone CNN-Transformer sigmoid-gated feature fusion and SimCLR-style dual-branch SSL pre-training within a single FER framework.

## 3 Materials and Methods

The proposed FER system takes a facial image as input and classifies it into one of seven or eight emotion categories, depending on the dataset. The framework operates in four stages, (a) data preparation and augmentation, (b) dual-branch self-supervised pre-training, (c) supervised fine-tuning with evidential uncertainty, and (d) evaluation and visualization. The overall pipeline is illustrated in Fig. 1.

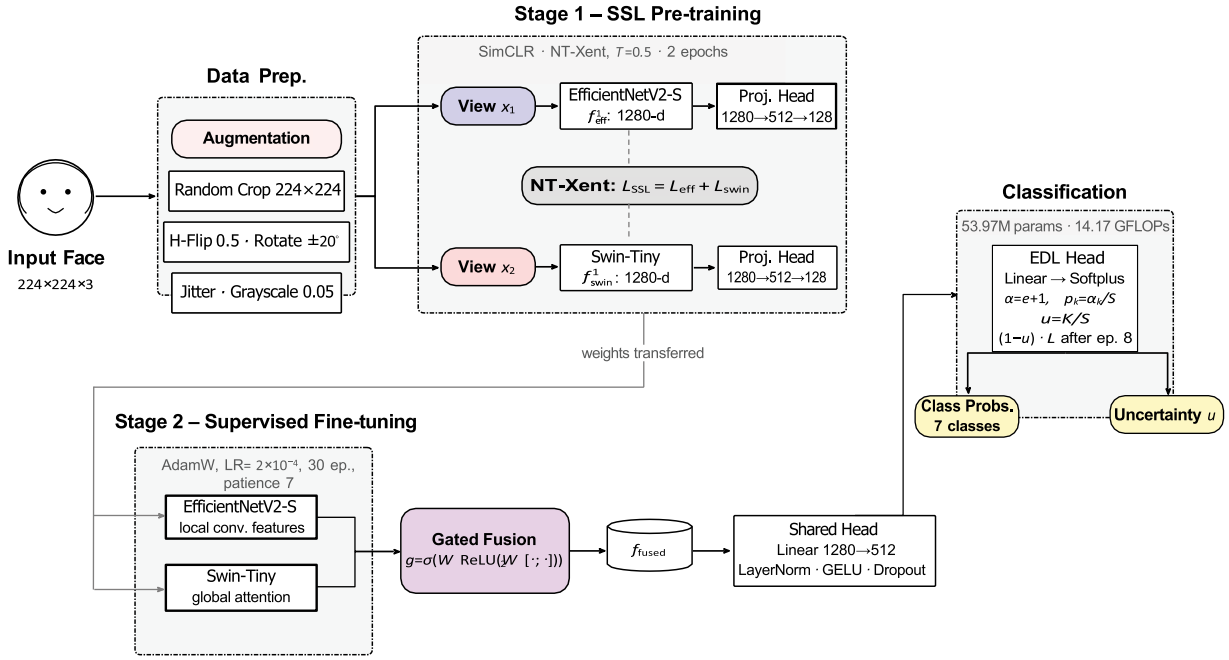


**Figure 1:** Proposed Dual-Stream FERNet pipeline.

### *3.1 Datasets*

Two benchmark datasets are used for evaluation. The KDEF (Karolinska Directed Emotional Faces) [30] contains images of 70 subjects (35 male, 35 female) posed in seven emotion categories, anger, disgust, fear, happy, neutral, sad, and surprise. 2938 images across the seven classes are used in this study, drawn from a publicly available pre-sorted mirror of the dataset. Images are captured under controlled laboratory conditions with consistent illumination and frontal or near-frontal poses. The second CK+ dataset [31] is drawn from a publicly available mirror providing 981 images across seven emotion categories, anger, contempt, disgust, fear, happy, sadness, and surprise.

Model performance is evaluated using subject-disjoint stratified 5-fold cross-validation. The StratifiedGroupKFold ensured subject identity so that no subject's images appear in both the training and test partition of any fold. For CK+, all images belonging to a given subject are likewise constrained to a single fold. Each fold trains on approximately 80% and tests on the remaining approximately 20% of subjects' images, with the exact per-fold counts varying slightly because grouping by subject does not yield perfectly equal partitions.

### *3.2 Data Augmentation*

Data augmentation artificially increases the effective size of the training set by applying random transformations to existing images while preserving their semantic label [19]. For training images, the following augmentation pipeline is applied sequentially.

- Random resized crop to 224 × 224 with scale range (0.75, 1.0). This introduced spatial variation while preserving face structure.
- Random horizontal flip with probability 0.5 to provide mirror-image invariance.
- Random rotation within ±20° to generate rotation-invariant feature maps.
- Color jitter with brightness ±0.30, contrast ±0.30, saturation ±0.20, hue ±0.06. This method simulated diverse illumination and camera conditions.

- Random grayscale conversion with probability 0.05 to encourage color-invariant features.
- Normalization using ImageNet statistics (mean [0.485, 0.456, 0.406], std [0.229, 0.224, 0.225]).

Evaluation images are only resized to 224 × 224 and normalized. The SSL augmentation pipeline additionally applied Gaussian blur (kernel size 3), more aggressive color jitter (brightness/contrast 0.4, saturation 0.3, hue 0.08), and random grayscale with probability 0.15 to generate two sufficiently different yet semantically consistent views of each image.

### *3.3 Proposed Dual-Stream FERNet Architecture*

The proposed Dual-Stream FERNet model adopted a dual-backbone design with a sigmoid-gated attention fusion module, a shared projection head, and an EDL classification head. The architecture summary is given in Table 1. The EfficientNetV2-S backbone served as the speed-optimised stream. The original classification head was removed, and the backbone generated a 1280-dimensional feature vector $\mathbf{f}_{\text{eff}}$ per image. EfficientNetV2-S used Fused-MBConv blocks in earlier stages and standard MBConv blocks in later stages. This mechanism achieved superior throughput relative to accuracy. The Swin-T backbone captured global spatial relationships through shifted-window self-attention. It produced a 768-dimensional vector $\mathbf{f}_{\text{swin}}$, which was projected to 1280 dimensions via a learnable linear layer to match the EfficientNetV2-S feature space. The two branch features were fused via a sigmoid-gated mechanism as follows.

$$\mathbf{g} = \sigma\left(\mathbf{W}_2\,\mathrm{ReLU}\left(\mathbf{W}_1\left[\mathbf{f}_{\text{eff}};\,\mathbf{f}'_{\text{swin}}\right]\right)\right) \tag{1}$$

$$\mathbf{f}_{\text{fused}} = \mathbf{g} \odot \mathbf{f}_{\text{eff}} + (1 - \mathbf{g}) \odot \mathbf{f}'_{\text{swin}} \tag{2}$$

This formulation allowed the model to dynamically weight each branch's contribution on a per-sample basis. Gate values average 0.675. EfficientNetV2-S gets favored over Swin-T about two to one. This held up on real test data across two separate experimental runs. Layer normalization was applied to $\mathbf{f}_{\text{fused}}$ before the shared projection head. A linear layer reduced $\mathbf{f}_{\text{fused}}$ from 1280 to 512 dimensions, followed by Layer Normalisation, GELU activation, and 50% dropout. This 512-dimensional representation served as the final shared feature for the EDL head. A linear layer mapped the 512-dimensional feature to $K$ class dimensions ($K = 7$ for both KDEF and CK+, checked directly on the trained model). This was corrected from an earlier version that listed 8 outputs for CK+, which was a writing mistake, followed by a Softplus activation. Class probabilities $\mathbf{p}$ and per-sample uncertainty $u$ were computed as follows.

$$\boldsymbol{\alpha} = \mathbf{e} + 1, \quad S = \sum_k \alpha_k, \quad p_k = \frac{\alpha_k}{S}, \quad u = \frac{K}{S} \tag{3}$$

High $u$ indicates that the model accumulates little evidence for any class, a signal for genuinely ambiguous inputs. The EDL loss combined a mean-squared error term between predicted probabilities and one-hot targets with a KL-divergence regularizer annealed over the first ten epochs. After epoch eight, the supervised loss was weighted by $(1 - u)$ per sample to reduce the influence of uncertain examples during fine-tuning.

**Table 1:** Dual-Stream FERNet architecture summary.

| Component | Configuration | Output Dim. |
|---|---|---|
| EfficientNetV2-S | ImageNet-1K pre-trained, classifier removed | 1280 |
| Swin-Tiny | ImageNet-1K pre-trained, head removed + Linear projection | 1280 |
| Sigmoid-Gate (Fusion) | Linear(2560 → 1280) → ReLU → Linear(1280 → 1280) → Sigmoid | 1280 |
| Fusion + LayerNorm | Gated weighted sum + normalization | 1280 |
| Shared Head | Linear(1280 → 512) + LayerNorm + GELU + Dropout(0.5) | 512 |
| EDL Head (KDEF) | Linear(512 → 7) + Softplus | 7 |
| EDL Head (CK+) | Linear(512 → 7) + Softplus | 7 |
| Total Parameters | 54,261,969[a] (KDEF)/54,261,969[a] (CK+)—all trainable | — |
| Model Size | 208.15 MB (KDEF)/208.16 MB (CK+) | — |

Note: [a]The parameter count refers exclusively to the supervised fine-tuning stage. During the SSL pre-training phase, approximately 1,704,320 additional parameters are added. These parameters are discarded after SSL pre-training and do not form part of the supervised model. No backbone, fusion, or classification-head parameters are frozen at any stage.

### *3.4 Self-Supervised Pre-Training (Independent Per-Branch Adaptation)*

Before supervised fine-tuning, both backbone branches were jointly pre-trained for two epochs using a SimCLR-style contrastive objective. Each branch trains on its own during this step. There is no shared or joint objective between the two branches at this stage. For each mini-batch, two independently augmented views $(x_1, x_2)$ were generated from the same set of images. Both views passed through the respective backbone branches to produce four feature tensors, $(\mathbf{f}^1_{\text{eff}}, \mathbf{f}^1_{\text{swin}})$ and $(\mathbf{f}^2_{\text{eff}}, \mathbf{f}^2_{\text{swin}})$. Separate projection heads (1280 → 512 → 128 dimensions) were applied to each branch for the SSL phase. The NT-Xent contrastive loss was computed independently for each branch at temperature 0.5, as follows.

$$\mathcal{L}_{\text{SSL}} = \mathcal{L}_{\text{NCE}}(\mathbf{z}^1_{\text{eff}}, \mathbf{z}^2_{\text{eff}}) + \mathcal{L}_{\text{NCE}}(\mathbf{z}^1_{\text{swin}}, \mathbf{z}^2_{\text{swin}}) \tag{4}$$

AdamW with learning rate $1 \times 10^{-4}$ and weight decay $1 \times 10^{-5}$ were used as SSL optimizers. For KDEF, the SSL loss decreased from 3.5611 (Epoch 1) to 2.8344 (Epoch 2). For CK+, it decreased from 3.9205 (Epoch 1) to 3.0583 (Epoch 2). This trend confirms the convergence of contrastive alignment in both settings. The original SimCLR framework trained from random initialization on ImageNet-scale data and therefore required hundreds of epochs to develop meaningful representations. In this study, both backbone branches enter the SSL phase already initialized from ImageNet-1K pre-trained weights. The objective is therefore lightweight domain adaptation rather than learning representations from scratch. Because this SSL stage uses only two epochs and a mini-batch size of 16, it is described as brief adaptation rather than full self-supervised pre-training. An epoch sweep was run to check this. More SSL training made linear-probe accuracy worse, not better. Two epochs scored highest and twenty five epochs scored lowest. To prevent transductive leakage, SSL pre-training is run independently inside each outer training fold and restricted strictly to that fold's training-partition images. Validation and test subjects are excluded from every SSL update, and this restriction is logged at run time for auditability. The experimental run confirmed both CKA and cosine similarity came out close to zero, around 0.01 and near zero, respectively. This means the two branches are learning different things, not the same representation. The isolated contribution of this SSL phase, disentangled from the Evidential Deep Learning head introduced concurrently, is not tested by the current ablation.

### *3.5 Supervised Fine-Tuning and Training Configuration*

Following SSL pre-training, the full model (both backbones, gated fusion, shared head, and EDL head) was trained in supervised mode. Training hyperparameters were held constant across both datasets to enable fair comparison. The supervised optimizer was AdamW with learning rate $2 \times 10^{-4}$ and weight decay $1 \times 10^{-4}$. A cosine annealing learning rate scheduler was applied over up to 30 epochs with a mini-batch size of 16. Early stopping with patience of 7 epochs prevented overfitting. Gradient norms were clipped to 5.0 for training stability. The model checkpoint with the highest validation accuracy was retained for evaluation. This study used a nested protocol, as an outer 80/20 fold does not by itself provide an independent validation set. Within each outer training fold, a further subject-disjoint inner validation split is carved out and used exclusively for early stopping and checkpoint selection. The outer test fold is exactly used for final evaluation, and never influences training, validation, or model selection. The zero subject overlap between the inner training and validation partitions is verified programmatically.

The complete evidential loss is specified as follows. Given evidence $e \geq 0$ (softplus output), $\alpha = e + 1$, $S = \sum_k \alpha_k$, the data-fit term is $L_{err}(i) = \sum_k (y_{ik} - \alpha_{ik}/S_i)^2$ and the variance term is $L_{var}(i) = \sum_k \alpha_{ik}(S_i - \alpha_{ik})/(S_i^2(S_i + 1))$. The KL regularizer uses the target-adjusted Dirichlet $\tilde{\alpha}_i = y_i + (1 - y_i) \odot \alpha_i$ (evidence for the true class removed) and is annealed as $\lambda_t = \min(1, t/10)$ over the first 10 epochs, as follows. $L_{KL}(i) = \mathrm{KL}(\mathrm{Dir}(\tilde{\alpha}_i) \,\|\, \mathrm{Dir}(\mathbf{1}))$. The total per-sample loss is $L_{EDL}(i) = L_{err}(i) + L_{var}(i) + \lambda_t L_{KL}(i)$. After epoch 8, this is weighted by $(1 - u_i)$ where $u_i = K/S_i$. $u_i$ is detached from the computational graph before this multiplication, so the model cannot reduce its loss by increasing predicted uncertainty rather than improving accuracy.

A natural concern with a 54-million-parameter dual-backbone model is over-parameterization relative to the size of the training partitions (approximately 2300–2410 images per fold for KDEF and 780–795 images per fold for CK+). This concern is substantially mitigated by several design choices. First, both EfficientNetV2-S and Swin-T backbones are initialized from ImageNet-1K pre-trained weights rather than random initialization. Second, four regularization mechanisms act simultaneously. (a) 50% Dropout in the shared projection head. (b) AdamW weight decay $\lambda = 1 \times 10^{-4}$. (c) Gradient norm clipping at 5.0. (d) Early stopping with patience of seven epochs, retaining only the best validation checkpoint. Third, the uncertainty-weighted EDL loss down-weights ambiguous samples.

## 4 Results and Discussion

This section presents and analyses the experimental results obtained from both datasets. Sections 4.1 and 4.2 report training convergence and final performance on KDEF and CK+, respectively. Section 4.5 compares these results against baseline models. Section 4.6 provides a comparison with state-of-the-art methods. Sections 4.7 and 4.8 cover Grad-CAM visualizations and computational efficiency.

### *4.1 KDEF Dataset Results*

The KDEF experiment used images across seven emotion classes, with subject-disjoint stratified 5-fold cross-validation. Fig. 2 shows loss and accuracy for all 5 folds. Thin lines are individual folds. The bold line is the mean. The fold-to-fold variability is visible directly. Mean test accuracy across the 5 folds is 93.84 ± 1.73% (macro-F1 93.82%, MCC 0.928). This is 3.67 points lower than the 97.51% obtained under an image-level, non-subject-disjoint protocol. This drop in accuracy is due to the subject leakage. KDEF has multiple images per subject, and an image-level split lets the same face appear in both train and test. That inflates accuracy. Per-fold accuracy ranged from 90.99% to 95.21% as given in Table 2. Fold 3 ran all 30 epochs and scored 93.52%, lower than Fold 5, which stopped at epoch 28 and scored 95.07%. Epoch count is not the main driver of fold-to-fold variance. Subject composition of each test fold is the more likely factor.

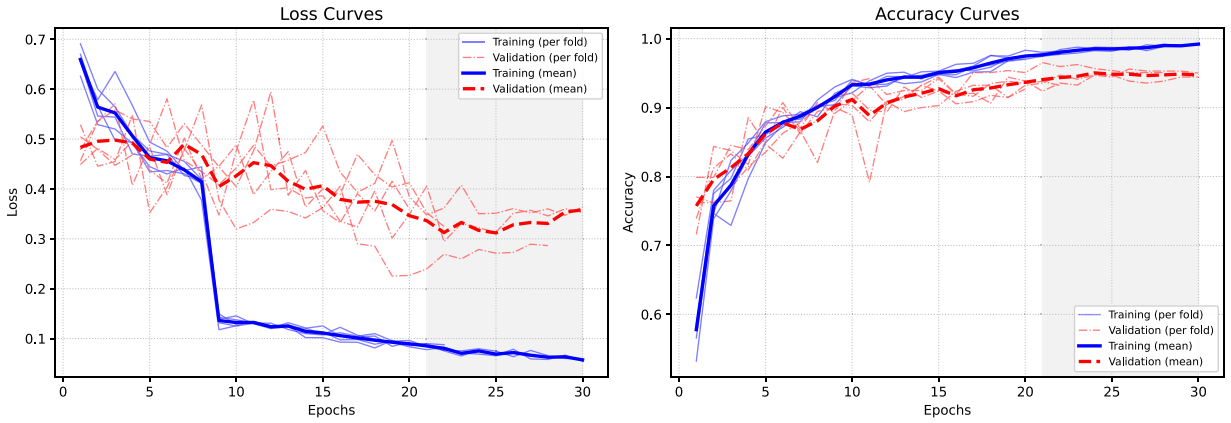


**Figure 2:** KDEF training and validation curves with mean across folds.

**Table 2:** Per-fold breakdown of proposed model performance on the KDEF test set (subject-disjoint 5-fold protocol).

| Fold | Acc. (%) | Prec. (%) | F1 (%) | Macro-F1 (%) | Bal. Acc. (%) | MCC | Mean $u$ |
|---|---|---|---|---|---|---|---|
| 1 | 95.21 | 95.20 | 95.19 | 95.19 | 95.20 | 0.944 | 0.312 |
| 2 | 90.99 | 91.23 | 90.97 | 90.98 | 90.99 | 0.895 | 0.378 |
| 3 | 93.52 | 93.85 | 93.46 | 93.45 | 93.51 | 0.925 | 0.297 |
| 4 | 94.42 | 94.42 | 94.38 | 94.38 | 94.41 | 0.935 | 0.399 |
| 5 | 95.07 | 95.29 | 95.08 | 95.09 | 95.06 | 0.943 | 0.316 |
| **Mean** | **93.84** | 94.00 | 93.82 | 93.82 | 93.83 | 0.928 | 0.340 |

The confusion matrix on the KDEF test set, shown in Fig. 3, reveals that all seven emotion classes are classified with high per-class accuracy. Happy and neutral achieved the highest per-class accuracy, consistent with prior literature attributing strong performance on these categories to their distinctive and less ambiguous facial configurations. Fear and disgust show slightly lower individual class scores. This reflects the established difficulty to distinguish these two categories due to their overlapping action units.

The proposed model performance and mean evidential uncertainty of 0.340 given in Table 3 indicates that the model accumulated strong evidence for its predictions on the majority of test samples. The uncertainty distribution across KDEF test images is shown in Fig. 4a. Samples with high uncertainty predominantly correspond to the boundary cases between fear and disgust, and between sad and neutral, which is consistent with the psychological literatures on emotion ambiguity. The t-SNE projection of the 512-dimensional shared feature embeddings shown in Fig. 4b uses only out-of-fold, subject-disjoint embeddings, with seed 42, perplexity 30, 1000 iterations, and PCA initialization. Cluster separability is quantified directly rather than assessed visually.

### *4.2 CK+ Dataset Results*

The CK+ experiment used 981 images across seven emotion classes, including contempt. The same subject-disjoint 5-fold protocol is used as KDEF. Fig. 5 shows a CK+ training run. The pattern is a general property of CK+'s small folds, seen across runs. Mean test accuracy across the 5 folds is 93.07 ± 1.22% (macro-F1 90.53%, MCC 0.917) as given in Table 4. This variance is smaller than KDEF's (±1.73%). So a smaller dataset does not always mean higher fold-to-fold instability. One fold shows why weighted metrics alone fall short. In Fold 2, weighted F1 is 93.0% but macro-F1 is only 86.5%, a 6.5-point gap. This points to at least one minority class doing much worse than the larger classes in that subject split.


**Figure 3:** Confusion matrix of proposed model using KDEF dataset (mean across 5 subject-disjoint folds).

**Table 3:** Final model performance on the KDEF test set (5-fold mean, subject-disjoint protocol).

| Metric | Value |
|---|---|
| Accuracy | 93.84% |
| Precision (weighted) | 94.00% |
| Recall (weighted) | 93.84% |
| F1-Score (weighted) | 93.82% |
| Macro-F1 | 93.82% |
| Balanced Accuracy | 93.83% |
| MCC | 0.928 |
| Mean Evidential Uncertainty | 0.340 |

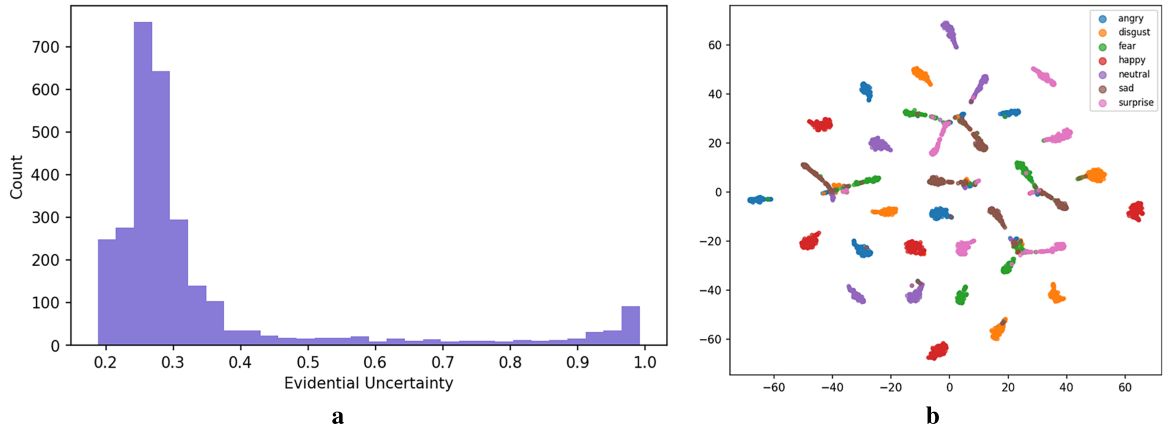


**Figure 4:** Qualitative analysis on the KDEF dataset. (**a**) Evidential uncertainty distribution across test images, and (**b**) t-SNE visualisation of shared feature embeddings.

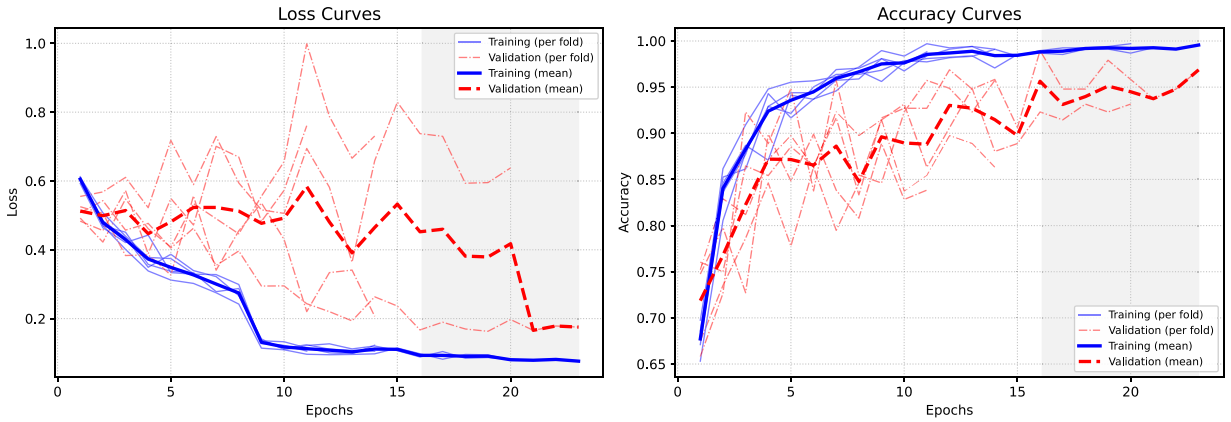


**Figure 5:** CK+ training and validation curves, all 5 subject-disjoint folds with mean across folds.

**Table 4:** Per-fold breakdown of proposed model performance on the CK+ test set (subject-disjoint 5-fold protocol).

| Fold | Acc. (%) | Prec. (%) | F1 (%) | Macro-F1 (%) | Bal. Acc. (%) | MCC | Mean $u$ |
|---|---|---|---|---|---|---|---|
| 1 | 93.01 | 93.73 | 92.80 | 92.14 | 92.04 | 0.919 | 0.442 |
| 2 | 93.23 | 93.96 | 93.02 | 86.49 | 87.99 | 0.917 | 0.343 |
| 3 | 95.02 | 95.79 | 95.12 | 93.98 | 96.19 | 0.941 | 0.401 |
| 4 | 92.04 | 92.79 | 92.01 | 89.34 | 90.34 | 0.905 | 0.512 |
| 5 | 92.04 | 92.98 | 91.52 | 90.69 | 88.95 | 0.904 | 0.396 |
| **Mean** | **93.07** | 93.85 | 92.89 | 90.53 | 91.10 | 0.917 | 0.419 |

Mean evidential uncertainty on CK+ (0.419) given in Table 5 is higher than on KDEF. The values are consistent with the harder classification problem, the smaller training set, and the inclusion of the contempt class. The confusion matrix (Fig. 6) shows contempt as the weakest class by a clear margin (78% recall), with 13% of true contempt samples misclassified as sadness, the single largest off-diagonal error in the matrix. Sadness and anger are confusable in both directions, consistent with the overlapping brow- and mouth-related action units these two expressions share. All other classes exceed 89% recall, with happy (99%) and surprise (97%) the most reliably classified.

**Table 5:** Final model metrics on the CK+ test set (5-fold mean, subject-disjoint protocol).

| Metric | Value |
|---|---|
| Accuracy | 93.07% |
| Precision (weighted) | 93.85% |
| Recall (weighted) | 93.07% |
| F1-Score (weighted) | 92.89% |
| Macro-F1 | 90.53% |
| Balanced Accuracy | 91.10% |
| MCC | 0.917 |
| Mean Evidential Uncertainty | 0.419 |

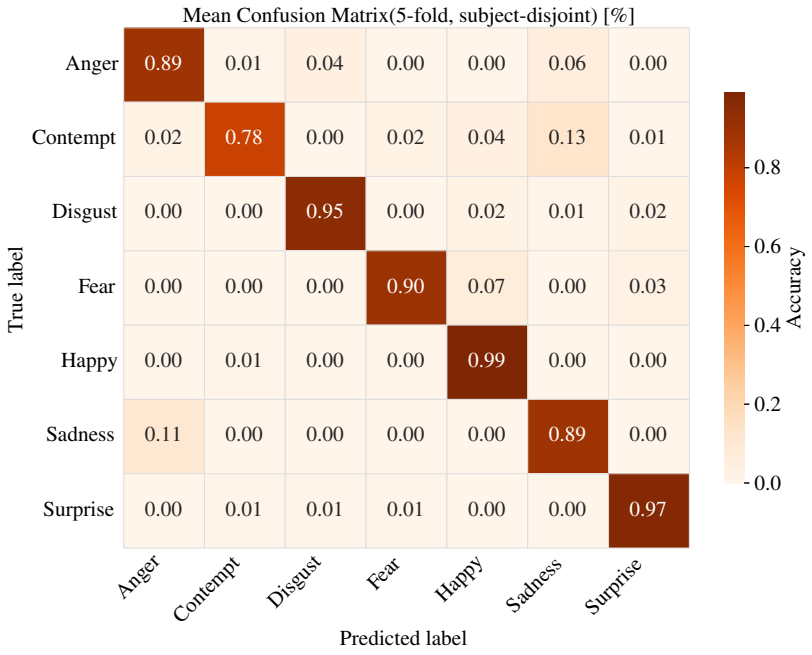


**Figure 6:** Confusion matrix of proposed model using CK+ dataset (mean across 5 subject-disjoint folds).

Fig. 7 presents a qualitative analysis of the proposed Dual-Stream FERNet on the CK+ dataset, using out-of-fold, subject-disjoint predictions and embeddings only. The uncertainty distribution is right-skewed, with most predictions concentrated below $u = 0.45$ and a long tail extending toward $u = 1.0$ consistent with the contempt-driven errors identified in the confusion matrix above. The t-SNE projection shows several well-separated clusters, notably disgust and surprise, alongside a denser, more overlapping region near the origin where anger, contempt, fear, and sadness embeddings mix. This visualization confirms the confusion matrix's off-diagonal errors concentrated among exactly these classes. The t-SNE visualization uses only out-of-fold, subject-disjoint embeddings (each image embedded by a model that never trained on it or its subject), with seed 42, perplexity 30, 1000 iterations, and PCA initialization. Cluster separability is quantified directly rather than assessed visually. The silhouette score in the 512-dimensional shared feature space is 0.115 for CK+.

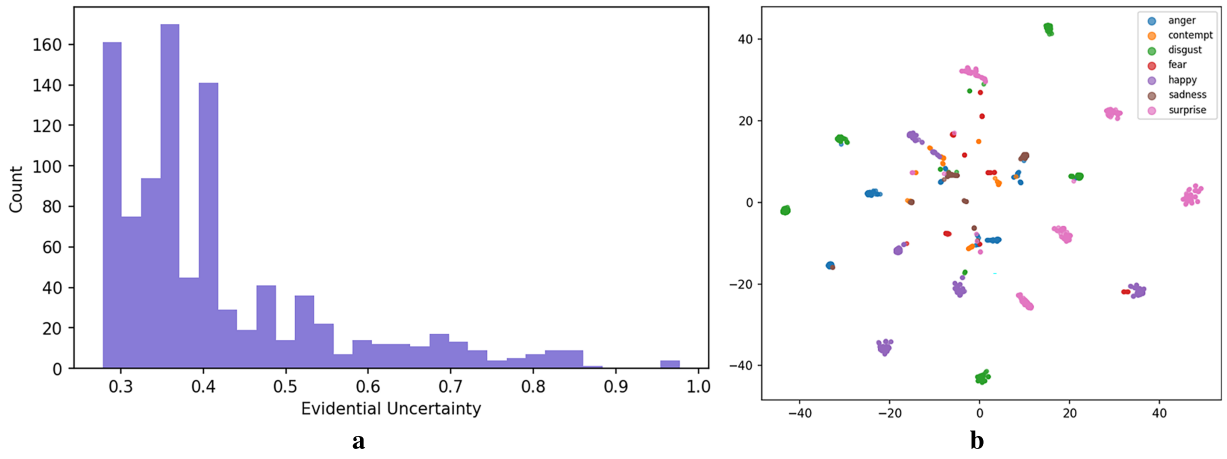


**Figure 7:** Qualitative analysis on the CK+ dataset. (**a**) Evidential uncertainty distribution across test images, and (**b**) t-SNE visualisation of shared feature embeddings.

### *4.3 Calibration Analysis*

A well-calibrated classifier produces confidence scores that reflect the true probability of a correct prediction. This matters particularly in safety-critical FER applications, where the system must not only classify emotions but also reliably signal its own uncertainty. For the proposed model, this study reports Expected Calibration Error (ECE), Adaptive ECE, Negative Log-Likelihood (NLL), and the multiclass Brier score. These metrics are computed on out-of-fold predictions for both KDEF and CK+, rather than a single dataset and a single 15-bin ECE as in the original submission. ECE measures the average gap between predicted confidence and empirical accuracy across confidence bins, with lower values indicating better calibration. NLL measures the average log-loss of the predicted probability assignments. The multiclass Brier score computes the mean squared error between predicted probability vectors and one-hot encoded labels, where lower values indicate both better calibration and higher accuracy.

On KDEF out-of-fold predictions, the model reaches ECE = 0.2258 (95% bootstrap CI [0.2148, 0.2362]), Adaptive ECE = 0.2258, NLL = 0.4923, and Brier = 0.1765 (Table 6). On CK+ out-of-fold predictions, ECE = 0.2859 (95% CI [0.2712, 0.3005]), Adaptive ECE = 0.2855, NLL = 0.5969, Brier = 0.2312. Calibration is worse on CK+ than KDEF on every metric. This fits CK+'s smaller training partitions and higher fold-to-fold variance. Both ECE values sit in the 0.23–0.29 range, which is poor in absolute terms. The model cannot be called "well-calibrated" without a baseline to compare against. Fig. 8 shows reliability diagrams for both datasets. This is underconfidence, not the overconfidence typical of softmax classifiers. It follows directly from the Dirichlet EDL head's evidence-based confidence, and it is a different failure mode from the overconfident miscalibration usually reported in FER calibration work. The model is not compared against softmax cross-entropy, temperature-scaled softmax, or MC-dropout baselines in this study.

**Table 6:** Calibration metrics for the proposed Dual-Stream FERNet.

| Dataset | ECE | Adaptive ECE | NLL | Brier | 95% CI on ECE |
|---|---|---|---|---|---|
| KDEF | 0.2258 | 0.2258 | 0.4923 | 0.1765 | [0.2148, 0.2362] |
| CK+ | 0.2859 | 0.2855 | 0.5969 | 0.2312 | [0.2712, 0.3005] |

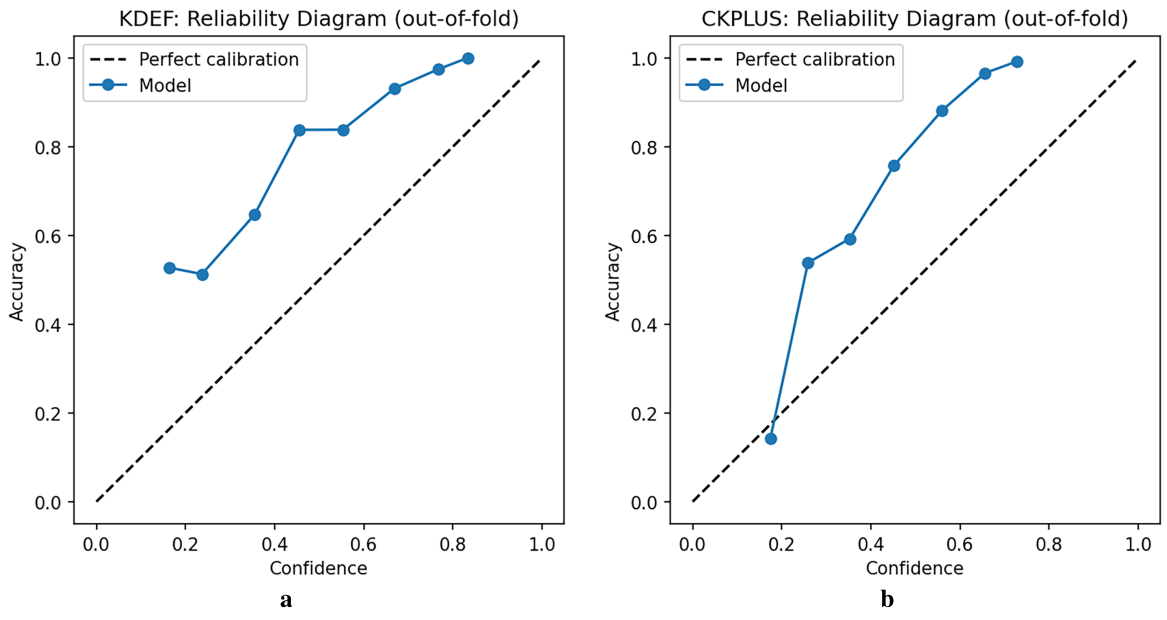


**Figure 8:** Reliability diagrams, out-of-fold predictions. (**a**) KDEF. (**b**) CK+.

### 4.4 Ablation Study

To quantify the individual contribution of each proposed component, a four-variant ablation study is conducted on KDEF, with every variant trained on identical subject-disjoint folds and an identical seed (5 folds × 1 seed = 5 runs per variant). Four conditions are evaluated. Variant A is the Swin-T backbone only, classification head, cross-entropy loss. Variant B is dual-stream concatenation (EfficientNetV2-S + Swin-T), no SSL pre-training, cross-entropy loss. Variant C is dual-stream sigmoid-gated fusion, no SSL pre-training, cross-entropy loss. Variant D is the full proposed model (gated fusion, SSL pre-training, EDL head, and uncertainty weighting). Given the computational cost of retraining under the corrected subject-disjoint protocol, this ablation uses a single seed rather than multiple seeds. Table 7 reports the results. Variant A (Swin-T alone) reaches 90.67 ± 2.10%, the lowest of the four. A single backbone captures less discriminative information than a dual-stream. Variant B (concatenation, no SSL/EDL) reaches 92.50 ± 1.60%, a gain of 1.83 pp over the single-backbone control. Variant C (gated fusion, no SSL/EDL) reaches 90.40 ± 2.09%. This is 2.10 pp *lower* than simple concatenation (Variant B). Three other ways to combine the two branches were evaluated. Average fusion scored 92.96%, scalar-weighted fusion scored 93.35%, and cross-attention fusion scored 93.55%, all beating the gate's 90.40%. Each of these differences was checked with a confidence interval, and none include zero. Variant D, the full proposed model, reaches 93.84 ± 1.73%. This beats every simpler variant, including concatenation (Variant B), by 1.34 pp. The complete system (gating, SSL, and EDL together) beats every other variant evaluated.

**Table 7:** Ablation study results on KDEF, 5-fold subject-disjoint, single seed.

| Var. | Description | KDEF Acc. (%) |
|---|---|---|
| A | Swin-T only, CE | 90.67 ± 2.10 |
| B | Dual concat, no SSL/EDL, CE | 92.50 ± 1.60 |
| C | Dual gated, no SSL/EDL, CE | 90.40 ± 2.09 |
| D | Proposed (gated + SSL + EDL) | 93.84 ± 1.73 |

### 4.5 Comparison with Baseline Models

Table 8 compares the proposed Dual-Stream FERNet against baseline models. All models use the same protocol, augmentation, ImageNet initialization, target-domain SSL pre-training, optimizer budget (AdamW, LR = $2 \times 10^{-4}$, weight decay $10^{-4}$, cosine annealing), and model selection. This study tests significance with a paired Wilcoxon signed-rank test, plus a paired $t$-test as a check. The subject-disjoint fold is the experimental unit. Cohen's $d$, paired bootstrap 95% confidence intervals, and Holm-Bonferroni-corrected $p$-values are reported across all pairwise tests (5 folds each). On KDEF, the proposed model reaches 93.24 ± 1.54% mean accuracy in this experiment. This is a separate training run from the one in Section 4.1, used elsewhere in the paper as the result, 93.84 ± 1.73%. Both numbers come from the same model, protocol, and 5-fold procedure. For comparison, ResNet50 reaches 92.38 ± 2.17%, EfficientNet-B0 reaches 93.75 ± 1.51%, and Swin-Small reaches 93.17 ± 2.92%. None of the pairwise differences are significant after correction (all Holm-adjusted $p = 1.00$). The gain over ResNet50 (+0.86 pp, Cohen's $d = 0.20$, 95% CI $[-3.57, 4.96]$ pp, paired-$t$ $p = 0.76$) and Swin-Small (+0.06 pp, $d = 0.02$, CI $[-4.42, 3.24]$ pp, $p = 0.98$) is small. EfficientNet-B0 numerically beats the proposed model by 0.51 pp ($d = -0.22$, CI $[-3.23, 0.85]$ pp, $p = 0.75$), also not significant. This comparison does not establish an accuracy advantage for the proposed model.

**Table 8:** Fair (SSL-matched) baseline comparison on KDEF, 5-fold subject-disjoint.

| Baseline | Mean Acc. | Wilcoxon $p$ | Paired-$t$ $p$ | Cohen's $d$ | Mean Diff. | 95% CI (pp) | Holm $p$ |
|---|---|---|---|---|---|---|---|
| Proposed FERNet | 93.24% | – | – | – | – | – | – |
| ResNet50 | 92.38% | 0.75 | 0.76 | 0.20 | +0.86 pp | [−3.57, 4.96] | 1.00 |
| EfficientNet-B0 | 93.75% | 1.00 | 0.75 | −0.22 | −0.51 pp | [−3.23, 0.85] | 1.00 |
| Swin-Small | 93.17% | 1.00 | 0.98 | 0.02 | +0.06 pp | [−4.42, 3.24] | 1.00 |

### *4.6 Comparison with State-of-the-Art*

Table 9 presents a comparative analysis of the proposed Dual-Stream FERNet against recent FER methods. No method listed in Table 9 was evaluated under an identical subject-disjoint split, class configuration, and evaluation protocol to the present study. The table is therefore presented as contextual only, and no claim of superiority over these methods is made. Under the corrected protocol, the proposed model achieves 93.84 ± 1.73% on KDEF and 93.07 ± 1.22% on CK+.

**Table 9:** Performance comparison with state-of-the-art FER methods.

| Method | Dataset | #Classes | Accuracy (%) |
|---|---|---|---|
| Fine-tuned MobileNetV2 [32] | CK+ | 8 | 92.39 |
| CNN with Efficient Channel Attention [33] | KDEF | 7 | 88.2 |
| DCNN-BiLSTM with dual attention [34] | KDEF | 7 | 95.7 |
| Ensemble Learning [35] | CK+ | 8 | 92.0 |
| Proposed Dual-Stream FERNet | KDEF | 7 | 93.84 ± 1.73 |
| Proposed Dual-Stream FERNet | CK+ | 7 | 93.07 ± 1.22 |

### *4.7 Grad-CAM Interpretability Analysis*

This study computed Grad-CAM on the EfficientNetV2-S branch of the proposed model for KDEF. Grad-CAM needs a spatial (2D) activation map to compute gradients. After each backbone's features are pooled and passed through the sigmoid gate and fusion layer, the fused representation is a flat vector with no spatial structure left. So Grad-CAM cannot be applied to the fused model directly. Standard practice for multi-branch models is to compute Grad-CAM separately at each backbone's own last spatial layer. This study reports the EfficientNetV2-S branch. Fig. 9 shows the resulting heatmaps. Fear, sad, and surprise all concentrate on the central nose/mouth region. Surprise's hotspot sits directly over the open mouth, a plausible match since the mouth is that expression's most visible feature. Neutral gives a broad, diffuse activation over most of the central face, not a sharp hotspot, consistent with neutral having no single dominant action unit. Angry is the exception. Its hotspot sits near the eye/cheek boundary, not the central mouth/nose region the other four share, and it is more diffuse than the fear/sad/surprise maps.

### *4.8 Computational Efficiency*

Table 10 reports the full computational profile of the proposed Dual-Stream FERNet, measured on an NVIDIA GeForce RTX 4070 GPU (12 GB VRAM) with CUDA 12.1, PyTorch 2.5.1+cu121, in FP32 (single-precision floating point). Five warm-up batches are executed before timing begins to ensure GPU memory is fully allocated and caching effects have stabilised. The reported inference time includes image preprocessing (resize to 224 × 224 pixels and ImageNet-mean normalisation) but excludes face detection. In

a complete end-to-end pipeline, a lightweight face detector such as RetinaFace would add approximately 5–15 ms per frame depending on image resolution. Batch-16 throughput alone does not establish single-frame latency, so batch-size-1 latency (median and 95th-percentile, CUDA-synchronized, repeated measurement) is reported. On KDEF, batch-1 median latency is 18.96 ms (95th percentile 21.99 ms), about 53 single-frame inferences per second. This is well below the batch-16 throughput figure and is the more relevant number for a real-time single-camera use case. Per-image throughput does not depend on test-set size. This is checked directly with a fixed number (20) of synthetic batches of the same size, across two independent sessions. In session 1, Run A reached 456.3 FPS and Run B reached 456.0 FPS (0.06% difference). In session 2, Run A reached 459.7 FPS and Run B reached 457.9 FPS (0.40% difference). Both sessions show ordinary measurement noise, not an effect of dataset size. This study reports peak GPU memory (via `torch.cuda.max_memory_allocated`) and FLOPs verified with the named tool `thop`. The full profile is computed the same way for the proposed model and every fair (SSL-matched) baseline (Table 10). The proposed model needs 14.17 GFLOPs per image (2×-MACs convention) and has a total size of 222.82 MB. Peak GPU memory is 2.15 GB during batch-16 inference, well within the RTX 4070's 12 GB VRAM, and within reach of most consumer and workstation GPUs, though higher than any single-backbone baseline (1.45–1.69 GB). On KDEF, the model reaches 459.67 FPS at batch size 16.

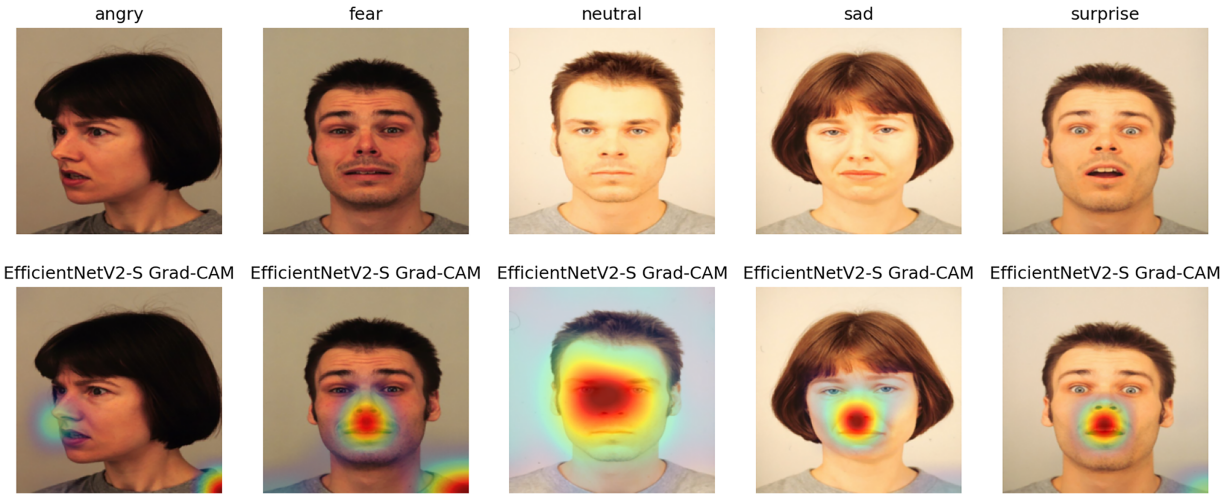


**Figure 9:** KDEF Grad-CAM visualisations (EfficientNetV2-S branch).

**Table 10:** Computational efficiency of the proposed Dual-Stream FERNet.

| Model | Params (M) | GFLOPs | Size (MB) | Peak Mem (MB) | B1 Med. (ms) | B1 p95 (ms) | FPS (B16) |
|---|---|---|---|---|---|---|---|
| Proposed FERNet | 53.97 | 14.17 | 222.82 | 2151.74 | 18.96 | 21.99 | 459.67 |
| ResNet50 | 24.56 | 8.27 | 102.70 | 1513.74 | 4.10 | 4.62 | 1171.36 |
| EfficientNet-B0 | 4.62 | 0.77 | 21.55 | 1449.84 | 4.89 | 5.97 | 2756.33 |
| Swin-Small | 49.19 | 17.09 | 198.78 | 1687.26 | 12.73 | 14.33 | 439.36 |

The proposed dual-stream model is more expensive than any single-backbone baseline on every axis measured. Its batch-1 median latency (18.96 ms) is roughly 1.6–4.6× that of the baselines, and its peak GPU memory (2.15 GB) is genuinely higher than every baseline (1.45–1.69 GB). This is the direct computational cost of the dual-stream architecture and the fusion/EDL head, traded against the added calibrated-uncertainty capability discussed in Section 4.3 that none of the baselines provide.

## 5 Conclusion

This paper combines EfficientNetV2-S and Swin-Tiny through sigmoid-gated fusion, SimCLR-style dual-branch self-supervised pre-training, and Evidential Deep Learning for uncertainty-aware classification. On KDEF, the model reaches 93.84 ± 1.73% accuracy at 459.67 FPS (batch-1 median latency 18.96 ms). On CK+, it reaches 93.07 ± 1.22% accuracy. Grad-CAM on both datasets shows attention on facial regions tied to FACS action units. This is qualitative evidence only. It does not show the model causally relies on those regions. The model's highest-uncertainty predictions fall on inter-class boundary cases, ones that are genuinely ambiguous even to people. CK+ performance also reflects the added difficulty of imbalanced minority classes such as contempt. These results position Dual-Stream FERNet as a controlled-setting baseline for future FER work on backbone fusion, SSL pre-training, and uncertainty modeling.

This study is limited to controlled, posed expressions. In-the-wild testing (RAF-DB, FER-2013, AffectNet, SFEW) is needed before any claim about specific deployments, such as driver monitoring or clinical mental-health screening. KDEF and CK+ are both controlled datasets with limited demographic diversity. Performance may not transfer across age, ethnicity, gender presentation, disability, or culture. Given this paper's motivating discussion of clinical and driver-monitoring use, larger-scale evaluation is needed before any deployment. Future work has three priorities. First, evaluation on large-scale, unconstrained FER benchmarks (RAF-DB, AffectNet, FER-2013) to test generalization beyond controlled lab conditions. Second, continual learning methods for online adaptation without catastrophic forgetting. Third, extending SSL pre-training to 50–100 epochs on the unlabeled target domain. Fourth, testing SSL and the EDL head separately to see how much each one actually contributes.

**Acknowledgement:** Not applicable.

**Funding Statement:** This work is supported by Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia through the Researchers Supporting Project PNURSP2026R333.

**Author Contributions:** The authors confirm contribution to the paper as follows. Rashid Jahangir handled formal analysis, investigation, methodology, and writing the original draft. Nazik Alturki handled visualization, methodology, writing review and editing, and funding acquisition. Mohammed Alreshoodi handled analysis, methodology, writing review and editing, and software. All authors reviewed and approved the final version of the manuscript.

**Availability of Data and Materials:** KDEF is available by request from Karolinska Institutet at https://kdef.se/download-2/. CK+ is available by request from the original maintainers at https://ckplus.jeffcohn.net.

**Ethics Approval:** KDEF was obtained directly from Karolinska Institutet. CK+ was obtained directly from the official CK+ distribution portal, request ID CKP-000027. Neither dataset involved new data collection from any person.

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

## References

1. Kaur M, Kumar M. Facial emotion recognition: a comprehensive review. Expert Syst. 2024;41(10):e13670.
2. Kopalidis T, Solachidis V, Vretos N, Daras P. Advances in facial expression recognition: a survey of methods, benchmarks, models, and datasets. Information. 2024;15(3):135.
3. Fei Z, Zhang B, Zhou W, Li X, Zhang Y, Fei M. Global multi-scale extraction and local mixed multi-head attention for facial expression recognition in the wild. Neurocomputing. 2025;622(3):129323. doi:10.1016/j.neucom.2024.129323.
4. Li N, Huang Y, Wang Z, Fan Z, Li X, Xiao Z. Enhanced hybrid vision transformer with multi-scale feature integration and patch dropping for facial expression recognition. Sensors. 2024;24(13):4153. doi:10.3390/s24134153.

5. Li Q, Fu K, Liu J, Li Y, Ren Q, Xu K, et al. Optimizing class imbalance in facial expression recognition using dynamic intra-class clustering. Biomimetics. 2025;10(5):296. doi:10.3390/biomimetics10050296.
6. Nemavhola A, Chibaya C, Viriri S. A systematic review of CNN architectures, databases, performance metrics, and applications in face recognition. Information. 2025;16(2):107. doi:10.3390/info16020107.
7. Song D, Liu C. A facial expression recognition network using hybrid feature extraction. PLoS One. 2025;20(1):e0312359. doi:10.1371/journal.pone.0312359.
8. Vatcharaphrueksadee A, Maliyaem M, Sawakchart P. Hybrid models for facial emotion recognition and intensity detection: generalization across human and cartoon faces using CNN and vision transformer. J Adv Inform Technol. 2025;16(4):478–90.
9. Shao Z, Chen B, Zhou Y, Shi X, Li C, Ma L, et al. Constrained and directional ensemble attention for facial action unit detection. Pattern Recognit. 2026;169(11):111904. doi:10.1016/j.patcog.2025.111904.
10. Ezzameli K. Vision transformer-based facial emotion recognition. IAENG Int J Comput Sci. 2026;53(1):410.
11. Wei LLX, Sani NS. Enhanced facial expression recognition based on ResNet50 with a convolutional block attention module. Int J Adv Comput Sci Appl. 2025;16(1):695–711.
12. Balachandran G, Ranjith S, Chenthil T, Jagan G. Facial expression-based emotion recognition across diverse age groups: a multi-scale vision transformer with contrastive learning approach. J Comb Optim. 2025;49(1):11. doi:10.1007/s10878-024-01241-8.
13. Yan L, Yang J, Xia J, Gao R, Zhang L, Wan J, et al. Self-supervised extracted contrast network for facial expression recognition. Multimed Tools Appl. 2025;84(15):14977–96. doi:10.1007/s11042-024-19556-3.
14. Zhu A, Jia X, Yang L, Zhou H, Su W. DUAL: a dual-stage approach for facial expression recognition based on contrastive learning. Int J Intell Syst. 2025;2025:7401168.
15. Yoon T, Kim H. Uncertainty estimation by density aware evidential deep learning. In: Proceedings of the 41st International Conference on Machine Learning; 2024 Jul 21–27; Vienna, Austria. p. 57217–43.
16. Li J, Zhou H, Qian Y, Dong Z, Wang SJ. Micro-expression recognition using dual-view self-supervised contrastive learning with intensity perception. Neurocomputing. 2025;619:129142. doi:10.2139/ssrn.4882306.
17. Kim JH, Kim BG, Roy PP, Jeong DM. Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. IEEE Access. 2019;7:41273–85. doi:10.1109/access.2019.2907327.
18. Jain N, Kumar S, Kumar A, Shamsolmoali P, Zareapoor M. Hybrid deep neural networks for face emotion recognition. Pattern Recognit Lett. 2018;115(2):101–6. doi:10.1016/j.patrec.2018.04.010.
19. Jain DK, Shamsolmoali P, Sehdev P. Extended deep neural network for facial emotion recognition. Pattern Recognit Lett. 2019;120:69–74. doi:10.1016/j.patrec.2019.01.008.
20. Minaee S, Minaei M, Abdolrashidi A. Deep-emotion: facial expression recognition using attentional convolutional network. Sensors. 2021;21(9):3046.
21. Shao J, Qian Y. Three convolutional neural network models for facial expression recognition in the wild. Neurocomputing. 2019;355:82–92. doi:10.1016/j.neucom.2019.05.005.
22. Umer S, Rout RK, Pero C, Nappi M. Facial expression recognition with trade-offs between data augmentation and deep learning features. J Ambient Intell Humaniz Comput. 2022;13(2):721–35. doi:10.1007/s12652-020-02845-8.
23. Georgescu MI, Ionescu RT, Popescu M. Local learning with deep and handcrafted features for facial expression recognition. IEEE Access. 2019;7:64827–36. doi:10.1109/access.2019.2917266.
24. Jeong M, Ko BC. Driver's facial expression recognition in real-time for safe driving. Sensors. 2018;18(12):4270. doi:10.3390/s18124270.
25. Khan M, Khan U, Awad M, Zaki N, Son G, Kwon S. Real-time emotion recognition system using adaptive distillation technique. Comput Model Eng Sci. 2026;147(1):1–10. doi:10.32604/cmes.2026.079697.
26. Baghel N, Dubey SR, Singh SK. UpAttTrans: upscaled attention based transformer for facial image super-resolution. Image Vis Comput. 2025;163:105731.
27. Khan M, El Saddik A, Deriche M, Gueaieb W. STT-Net: simplified temporal transformer for emotion recognition. IEEE Access. 2024;12:86220–31.
28. Shu Y, Gu X, Yang GZ, Lo B. Revisiting self-supervised contrastive learning for facial expression recognition. arXiv:2210.03853. 2022.

29. Gao J, Chen M, Xiang L, Xu C. A comprehensive survey on evidential deep learning and its applications. IEEE Trans Pattern Anal Mach Intell. 2026;48(3):2118–38. doi:10.1109/tpami.2025.3625258.
30. Goeleven E, De Raedt R, Leyman L, Verschuere B. The Karolinska directed emotional faces: a validation study. Cogn Emot. 2008;22(6):1094–118. doi:10.1080/02699930701626582.
31. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended Cohn-Kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops; 2010 Jun 13–18; San Francisco, CA, USA. p. 94–101.
32. Kaur M, Kumar M, Dasi S. Fine-tuning MobileNetV2 for lightweight facial emotion recognition: FER-2013 and CK+ datasets. Natl Acad Sci Lett. 2026:1–8.
33. Ramirez-Quintana JA, Muñoz-Pacheco JJ, Ramirez-Alonso G, Medrano-Hermosillo JA, Corral-Saenz AD. Lightweight convolutional neural network with efficient channel attention mechanism for real-time facial emotion recognition in embedded systems. Sensors. 2025;25(23):7264. doi:10.3390/s25237264.
34. Jayaraman S, Mahendran A. An interactive information based DCNN-BiLSTM model with dual attention mechanism for facial expression recognition. Sci Rep. 2025;15(1):26287. doi:10.1038/s41598-025-09709-1.
35. Manal A, Alsulaiman FA. An ensemble learning approach for facial emotion recognition based on deep learning techniques. Electronics. 2025;14(17):3415. doi:10.3390/electronics14173415.