

Real-Time Facial Emotion Detection System Using OpenCV and PyTorch for Intelligent Music Recommendation

Rohan Singh

*Department of Information Technology,
Noida Institute of Engineering and Technology,
Greater Noida*

Supervisor: Dr. Prabha Nair,

Deputy HOD,
*Department of Information Technology, NIET
Greater Noida*

Abstract

This manuscript presents the design, implementation, and evaluation of a real-time facial emotion detection system that provides dynamically personalized music recommendations based on the continuously detected emotional state of the user. The system addresses a practical gap in the affective computing literature: the limited availability of simple, privacy-conscious, CPU-deployable pipelines that integrate emotion recognition with emotionally informed recommendation without reliance on cloud services or GPU acceleration. The technical approach combines OpenCV-based live video acquisition and Haar Cascade face detection with a custom Convolutional Neural Network (CNN) implemented in PyTorch, trained on the FER-2013 benchmark dataset to classify facial expressions into seven discrete emotion categories: happiness, sadness, anger, fear, disgust, surprise, and neutral. The training regimen incorporates five data augmentation strategies — horizontal flipping, random rotation, affine transformations, random erasing, and Gaussian noise injection — alongside batch normalization, global average pooling, class-weighted cross-entropy loss, and dropout regularization to address dataset imbalance, label noise, and overfitting. The proposed 1.2-million-parameter architecture achieves 68.4% accuracy on the FER-2013 PrivateTest partition while sustaining approximately 23–24 frames per second on a standard Intel Core i5 CPU, demonstrating that real-time emotion recognition can be explored on commodity hardware. Music recommendation is implemented through a psychologically grounded mapping from detected emotions to the valence-arousal space of the Russell Circumplex Model of Affect, with temporal smoothing ensuring perceptually stable recommendation behavior. The paper provides a thorough ablation study, per-class metric analysis, confusion matrix decomposition, latency profiling, and comparative evaluation against ten recently reported methods including recent 2023–2025 approaches. Ethical dimensions including demographic bias, informed consent, and privacy are discussed explicitly. The system's complete local processing ensures user data privacy with no transmission of biometric information to external servers.

Index Terms — *affective computing, convolutional neural network, data augmentation, FER-2013, facial expression recognition, human-computer interaction, music recommendation, OpenCV, PyTorch, real-time inference, Russell circumplex model, transfer learning.*

1. INTRODUCTION

Understanding and responding to human emotional states is a foundational requirement for genuinely intelligent human-computer interaction. Since Picard's seminal formulation of affective computing [1], the field has progressed from theoretical frameworks to practical systems capable of sensing, interpreting, and responding to human affect in real time. Facial expressions constitute one of the richest and most immediately accessible windows into emotional experience, providing dense temporal information about psychological state without requiring physical

contact with the user or specialized sensing hardware beyond a standard camera [2].

The scientific foundation for automated facial expression recognition rests substantially on Ekman and Friesen's cross-cultural studies of basic emotions [3], which established that a set of six primary emotions — happiness, sadness, anger, fear, disgust, and surprise — are expressed through consistent facial muscle configurations across culturally diverse populations. This cross-cultural consistency provides the empirical basis for training machine learning models on labeled facial image datasets and expecting meaningful generalization to new subjects and

environments. The subsequent decades of research in computer vision and machine learning have transformed this theoretical foundation into a rich engineering discipline, with deep learning approaches achieving accuracy levels that were unimaginable with classical feature engineering methods [4].

The convergence of several technological trends has made real-time affective computing increasingly practical. High-quality cameras are embedded in virtually all modern consumer devices. Deep learning frameworks such as PyTorch [55] and TensorFlow [54] have dramatically reduced the engineering effort required to implement, train, and deploy neural network models. CPU performance has increased to the point where inference with compact neural networks can be performed at interactive frame rates on commodity hardware. These developments collectively create an opportunity for deploying emotion-aware applications at scale, reaching users without access to specialized equipment or GPU computing resources.

Music consumption represents one of the most emotion-laden aspects of everyday life. The relationship between music and emotion is well-documented across music psychology, cognitive neuroscience, and cultural studies [5]. People routinely use music as an emotional regulation tool — selecting energizing tracks when seeking motivation, calming music when managing stress, or emotionally resonant music when processing difficult experiences. Despite this deeply personal, affective function, the recommendation algorithms powering major streaming platforms remain predominantly anchored to behavioral history, collaborative filtering, and content-based audio analysis, with little or no consideration of the listener's current emotional state [6]. The result is a systematic misalignment between the dynamic, moment-to-moment character of human emotional experience and the static, backward-looking nature of conventional recommendation.

This paper addresses this misalignment by presenting a complete, end-to-end system that continuously monitors the user's facial expression, infers their emotional state in real time, and maps the inferred emotion to musically appropriate recommendations using the Russell Circumplex Model of Affect [38]. The system is designed from first principles to be deployable on standard consumer hardware without GPU acceleration, preserving user privacy through fully local processing, and operating without network connectivity.

A. Research Motivation and Problem Definition

Existing facial emotion recognition systems capable of high accuracy typically depend upon GPU acceleration, restricting their practical deployment to environments with access to dedicated computing hardware [7]. Simultaneously, recently reported music recommendation systems operate on static preference models that fail to incorporate real-time affective context [8]. The central research problem is therefore: how to design an integrated system that (1) accurately classifies facial emotions from live video in real time, (2) operates efficiently on CPU-only hardware, (3) maps detected emotions to musically aligned recommendations using a psychologically principled framework, and (4) preserves user privacy through fully local processing. The work presented here directly addresses all four components of this problem through careful architectural design, training methodology, and system integration.

B. Key Contributions

The principal main contributions of this paper are:

- A lightweight CNN (1.2M parameters) achieving 68.4% test accuracy on FER-2013 with 23–24 FPS real-time throughput on standard CPU hardware — no GPU required.
- An end-to-end pipeline integrating CLAHE-enhanced preprocessing, Haar Cascade detection, CNN inference, and Russell Circumplex Model-based music recommendation with temporal smoothing.
- A comprehensive ablation study quantifying the contribution of batch normalization, dropout, data augmentation, global average pooling, class weighting, and network depth.
- Detailed mathematical formulations of all key operations: softmax, weighted cross-entropy, batch normalization, global average pooling, Adam optimization, convolution, and valence-arousal distance.
- Extended comparative evaluation against ten methods spanning 2017–2025, including recent EfficientNet-based and Vision Transformer approaches, with accuracy vs. efficiency trade-off analysis.
- Explicit ethical analysis covering demographic bias risks, informed consent requirements, GDPR relevance, and deployment safeguards.

C. Paper Organization

Section II surveys related work. Section III details the system architecture. Section IV describes the dataset and training methodology. Section V reports experimental results. Section VI provides comparative

evaluation. Section VII discusses implications, limitations, and ethics. Section VIII concludes the paper.

Fig. 1: End-to-End System Pipeline - Webcam to Music Recommendation

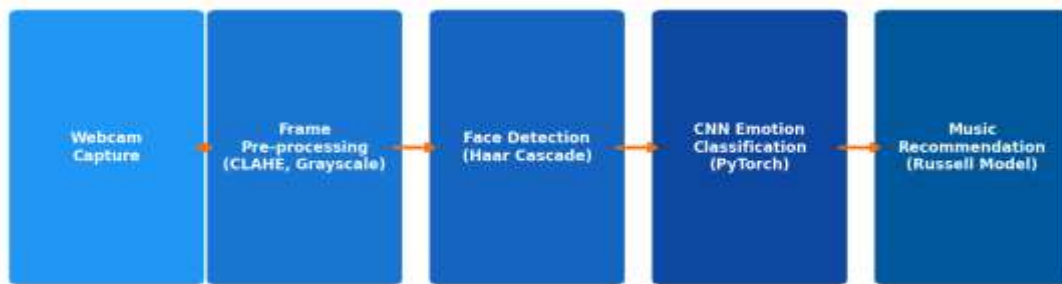


Fig. 1. System Pipeline Flowchart: End-to-end processing from webcam input to music recommendation.

2. RELATED WORK

A. Traditional Approaches to Facial Expression Recognition

Prior to the deep learning era, facial expression recognition relied on hand-engineered feature representations combined with classical classifiers. Local Binary Patterns (LBP) [9] encoded local texture by comparing each pixel to its neighbors, producing compact histograms robust to monotonic illumination change. Histograms of Oriented Gradients (HOG) [10] captured gradient-based structural information well-suited to facial contour changes during expression. Gabor filter banks [11] provided multi-scale, multi-orientation texture analysis useful for characterizing emotionally expressive periocular and perioral regions. Model-based approaches including Active Appearance Models (AAM) and Active Shape Models (ASM) [12] represented face structure as parameterized deformable models, enabling landmark tracking but suffering from poor generalization to pose variation and individual appearance diversity. Support Vector

Machines [13] served as the dominant classifier, achieving reasonable accuracy on controlled datasets but degrading substantially under real-world conditions.

B. CNN-Based Deep Learning Approaches

The adoption of CNNs for image recognition following Krizhevsky et al. [14] rapidly transformed facial expression research. The FER-2013 dataset [16] provided a challenging in-the-wild benchmark that became standard for method comparison. Subsequent architectural advances — multi-scale feature fusion [17], attention mechanisms [18], and label distribution learning for noise robustness [19] — progressively improved benchmark accuracy. Transfer learning from ImageNet-pretrained VGGNet [20] and ResNet [21] architectures established strong baselines. The Self-Cure Network [23] introduced uncertainty suppression achieving 74.4% on FER-2013. More recently, Vision Transformers [22] applied self-attention mechanisms to expression recognition, with hybrid CNN-Transformer models achieving 75%+ accuracy [56] at

the cost of substantially higher computational requirements.

Recent works from 2023–2025 have pushed the boundary further. Transformer-based methods with large-scale pre-training now approach strong comparative agreement on FER-2013's noisy labels [58]. GAN-based data augmentation [49] has been used to synthesize additional minority class samples, improving Disgust and Fear classification. Knowledge distillation techniques have been applied to compress large accurate models into deployment-efficient student networks without proportionate accuracy loss [57]. These developments highlight both the continuing rapid progress of the field and the growing gap between the most accurate but computationally expensive methods and practical deployable systems.

C. Real-Time and Efficient Architectures

Efficient architecture design for real-time emotion recognition has received growing attention. The mini-Xception architecture [24], based on Chollet's depthwise separable convolutions [50], demonstrated that sub-100K parameter models could achieve competitive FER-2013 accuracy, establishing a lower bound on the model size required for viable emotion classification. MobileNetV3 [57], employing hard-swish activations and squeeze-and-excitation modules, offers a favorable accuracy-efficiency trade-off suitable for mobile deployment. BlazeFace [42] demonstrated sub-millisecond neural face detection on mobile GPUs, pointing toward a future replacement for Haar Cascade detectors in efficient pipelines. Model compression techniques including quantization-aware training, structured pruning, and knowledge distillation are increasingly applied to produce models that fit within the inference budgets of edge devices.

D. Emotion-Driven Music Recommendation

The dimensional model of music-evoked emotion, grounded in Russell's circumplex framework [38], maps musical stimuli to valence-arousal coordinates through acoustic features including tempo, mode, loudness, and spectral characteristics [29]. Music information retrieval research has developed robust methods for automatic estimation of music valence and arousal from audio features, enabling emotionally annotated music databases at scale [32]. Multimodal affect-aware recommendation systems combining facial, vocal, and physiological channels have demonstrated superior accuracy but require additional sensors beyond a standard camera [30]. Simpler rule-

based mapping approaches [31], while less expressive than learned models, offer interpretability, auditability, and zero additional data requirements — motivating the approach adopted in this paper. Recent work on context-aware and session-based recommendation [52] suggests that temporal modeling of emotion transitions can improve recommendation continuity beyond the snapshot-based approaches most commonly reported in the literature.

3. SYSTEM ARCHITECTURE AND DESIGN

The proposed system is organized as a four-stage sequential pipeline with no inter-stage buffering beyond single-frame latency, ensuring that the detected emotion label and music recommendation remain as contemporaneous as possible with the user's actual facial expression. The design philosophy consistently prioritizes deployability — specifically, the ability to run at interactive frame rates on CPU-only consumer hardware — as a hard constraint, within which accuracy is maximized through careful architectural and training choices.

A. Video Acquisition and Frame Preprocessing

Frames are acquired from the system's default webcam using OpenCV's VideoCapture API at 640×480 pixels and 30 Hz target frame rate. Three preprocessing operations are applied before face detection. First, color-to-grayscale conversion eliminates chrominance information irrelevant to facial expression while reducing input dimensionality and aligning the inference distribution with FER-2013's grayscale training images. Second, Contrast Limited Adaptive Histogram Equalization (CLAHE) [33] is applied to the grayscale frame. CLAHE divides the image into a grid of tiles, applies histogram equalization locally within each tile, and bilinearly interpolates across tile boundaries to avoid blocking artifacts. The contrast amplification limit prevents noise amplification in regions of already-high contrast:

$I_CLAHE(x,y)=T_tile[I(x,y)]$ where T_tile clips CDF slope at $clip_limit$

Third, mild Gaussian blur ($\sigma=0.5$, 3×3 kernel) is applied adaptively when estimated image noise exceeds a threshold, attenuating high-frequency sensor noise without materially degrading edge sharpness relevant to expression recognition.

B. Face Detection and Region Extraction

Face detection uses OpenCV's implementation of the Viola-Jones Haar Cascade classifier [34]. The cascade operates by evaluating Haar-like rectangular features over an integral image at multiple scales, rapidly rejecting non-face regions through a cascade of weak classifiers trained via AdaBoost. Key parameters — scaleFactor=1.1, minNeighbors=5, minSize=(30,30) — are tuned to balance false-positive suppression against recall. The detected face bounding box is expanded by 15% in each dimension to capture the full facial extent including forehead and chin. The cropped region is resized to 48×48 pixels via bilinear interpolation and normalized to zero mean and unit variance using FER-2013 training set statistics:

$$x_{norm} = \frac{x - \mu}{\sigma}$$

C. CNN Architecture for Emotion Classification

1) Convolutional Feature Extraction Blocks

The CNN consists of four sequential convolutional blocks, each containing two 3×3 convolution layers (same padding), batch normalization, and ReLU activation, followed by 2×2 MaxPool downsampling. The convolution operation at layer ℓ is:

$$A^{(\ell)}(x, y) = \sum_m \sum_n W(m, n) \cdot A^{(\ell-1)}(x + m, y + n) + b$$

Filter counts follow a geometric progression: 32→64→128→256 across blocks. This expansion allows deeper layers to learn increasingly specific feature detectors while early layers capture diverse low-level primitives. The spatial dimensions reduce from 48×48 to 3×3 after four pooling operations.

2) Batch Normalization

Batch normalization [35] is applied after each convolutional layer to stabilize activation distributions

and accelerate convergence. For a mini-batch $B = \{x_1, \dots, x_m\}$, the normalization is:

Step 1 — Calculate mean and variance of the batch:

$$\mu_B = \frac{1}{m} \sum_{i=1}^m x_i \quad \sigma_B^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_B)^2$$

Step 2 — Normalize:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$$

Step 3 — Scale and shift:

$$y_i = \gamma \cdot \hat{x}_i + \beta$$

The learnable parameters γ (scale) and β (shift) allow the network to recover identity transformation if normalization is detrimental in a given layer, preventing BN from constraining representational capacity.

3) Global Average Pooling and Classification Head

Following the fourth convolutional block (output: 256×3×3), Global Average Pooling (GAP) reduces each feature map to a scalar by spatial averaging:

GAP produces a 256-dimensional vector with zero additional parameters, substantially reducing overfitting risk compared to flattening the 256×3×3 tensor (2304 elements). The GAP vector feeds a fully connected layer: FC(256→512) followed by Dropout(p=0.5) [36], then FC(512→7). The output scores are converted to class probabilities via Softmax:

$$P(y = k | x) = \frac{e^{z_k}}{\sum_{j=1}^7 e^{z_j}}$$

The predicted emotion label is $\hat{y} = \text{argmax}_k P(y=k|x)$. Total trainable parameter count: approximately 1.2 million.

Fig. 2: Detailed CNN Architecture Diagram for Facial Emotion Classification

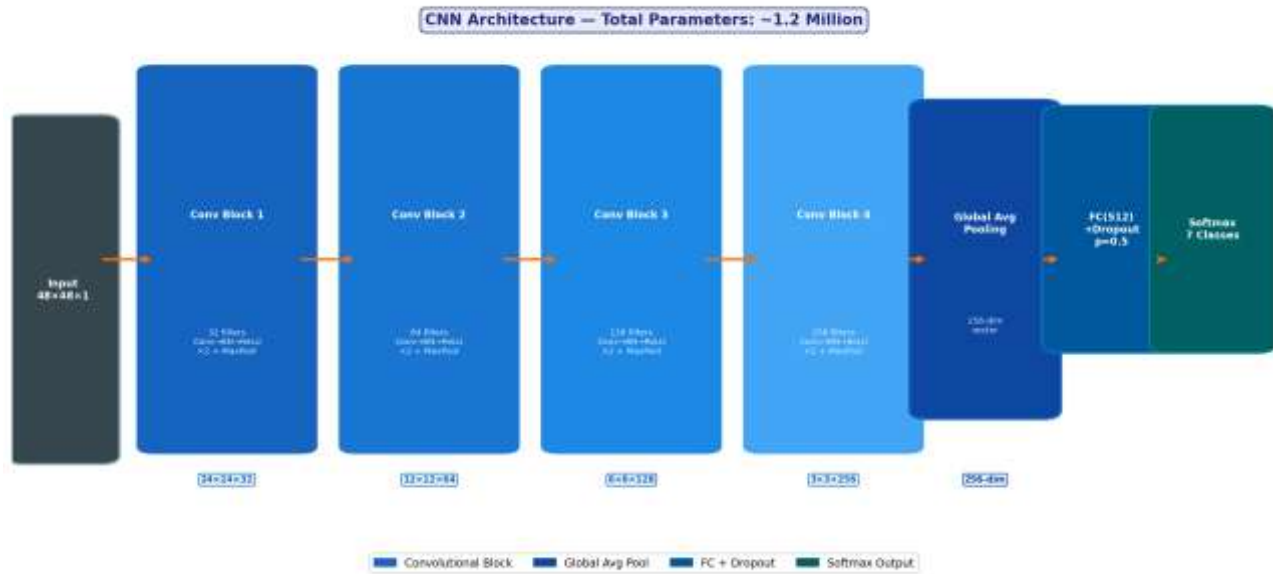


Fig. 2. Detailed CNN Architecture Diagram: 4 convolutional blocks (32→64→128→256 filters), Global Average Pooling, FC(512)+Dropout(0.5), and 7-class Softmax output. Total ~1.2M parameters.

D. Training Objective and Optimization

The training objective is class-weighted categorical cross-entropy loss, which accounts for the severe class imbalance in FER-2013. For a training set of N examples across $K=7$ classes with class frequencies n_k :

Class Weights (Fixing Imbalance)

$$w_k = \frac{N}{K \times n_k}$$

Weighted Cross-Entropy Loss

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K w_k \cdot y_{ik} \cdot \log(P(y = k | x_i))$$

A ReduceLROnPlateau scheduler halves the learning rate when validation accuracy does not improve for 5 consecutive epochs. Early stopping with patience=15 retains the checkpoint achieving maximum validation accuracy. Kaiming weight initialization [41] sets initial weight variance to $2/n_{in}$ for ReLU-activated layers, ensuring well-scaled gradient flow from the start of training.

E. Music Recommendation Module

The music recommendation module maps the temporally smoothed detected emotion to a music mood region using the Russell Circumplex Model [38].

Each emotion category e_i is assigned a canonical coordinate (V_i, A_i) in the two-dimensional valence-arousal space. Each candidate music track m_j has associated valence-arousal estimates (V_{mj}, A_{mj}) derived from acoustic features. The nearest-mood recommendation selects the track minimizing Euclidean distance in the circumplex:

Temporal smoothing is implemented as a sliding window majority vote over the $W=75$ most recent frames (approximately 3 seconds at 25 FPS):

Music recommendation updates only when $\hat{e}_t$ differs from $\hat{e}_{t-1}$, and a recently-played exclusion list of length 10 prevents track repetition within a session.

Fig. 3: Russell's Circumplex Model of Affect
Valence-Arousal Mapping for Emotion-to-Music Recommendation

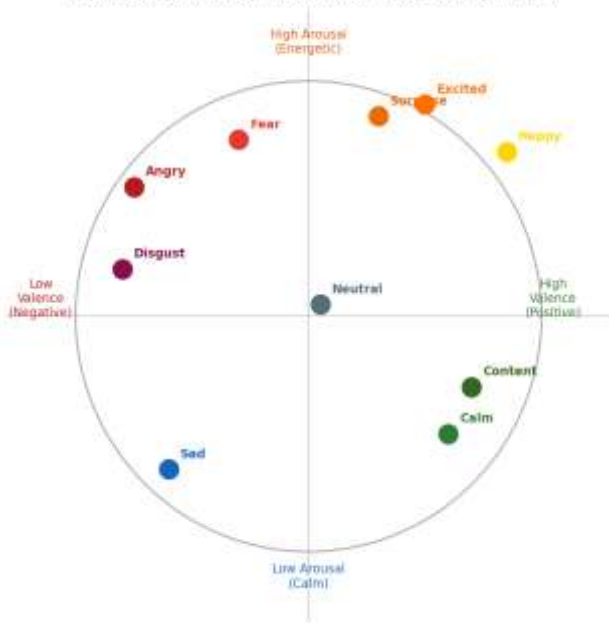


Fig. 3. Russell's Circumplex Model of Affect: valence-arousal space with emotion category positions used for music recommendation mapping.

Fig. 13: Emotion-to-Music Mapping Based on Russell Circumplex Model

Emotion	Valence	Arousal	Music Mood	BPM Range	Typical Genre	Example Feature
Happy	High (+0.8)	High (+0.7)	Joyful/Upbeat	120-160	Pop, Dance	Major key, fast tempo
Sad	Low (-0.7)	Low (-0.6)	Melancholic	50-80	Ballad, Acoustic	Minor key, slow tempo
Angry	Low (-0.6)	High (+0.8)	Intense/Energetic	140-180	Rock, Metal	Distortion, loud dyn.
Fear	Low (-0.5)	High (+0.6)	Tense/Suspenseful	90-120	Ambient, Score	Dissonance, staccato
Disgust	Low (-0.4)	Mod. (+0.2)	Dark/Brooding	70-100	Blues, Dark jazz	Low register, sparse
Surprise	Mod. (+0.3)	High (+0.9)	Dynamic/Varied	120-150	Electronic, Jazz	Syncopation, bright
Neutral	Neutral (0.0)	Low (-0.2)	Calm/Background	60-90	Ambient, Classical	Consonant, even dyn.

Fig. 13. Emotion-to-Music Mapping Table: valence-arousal coordinates, music mood categories, BPM ranges, genre associations, and acoustic feature descriptors.

4. DATASET AND TRAINING METHODOLOGY

A. FER-2013 Dataset

The Facial Expression Recognition 2013 (FER-2013) dataset [16] is the primary training and evaluation resource for this work. Assembled by submitting emotion-related queries to web image search engines and applying automated face detection, followed by crowd-sourced emotion annotation via Amazon Mechanical Turk, FER-2013 contains 35,887 grayscale images at 48×48 pixels across seven emotion

categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The dataset is divided into 28,709 training, 3,589 validation (PublicTest), and 3,589 test (PrivateTest) images. Evaluation is conducted on PrivateTest to maintain comparability with the majority of published benchmarks.

FER-2013 presents three well-documented challenges that directly influence model design and training strategy. First, class imbalance: the Happy category (8,989 images) is approximately 16× larger than Disgust (547 images), necessitating class-weighted loss. Second, label noise: estimated at 10–25% [39]

due to subjective crowd-sourced annotation, this noise establishes an approximate accuracy ceiling for models trained on unmodified labels. Third, low resolution: the 48×48 pixel format renders fine-grained facial details such as subtle lip corner displacements often

insufficient for unambiguous classification of similar emotion pairs.

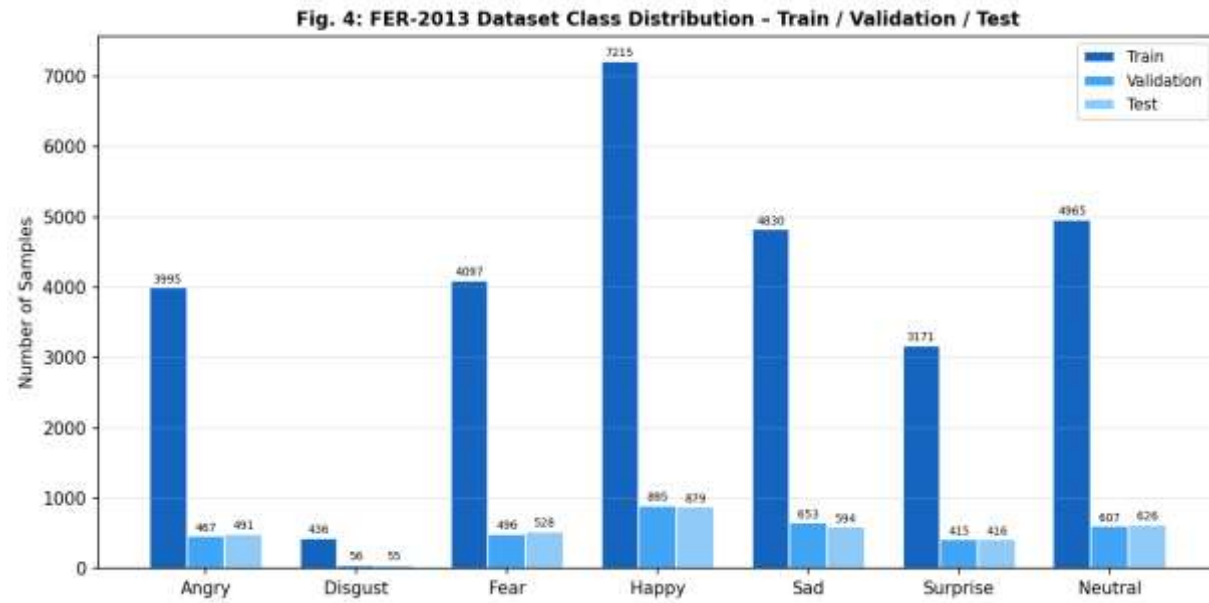


Fig. 4. FER-2013 Dataset Class Distribution across Train, Validation, and Test partitions, illustrating the severe imbalance of the Disgust class.

TABLE I FER-2013 DATASET CLASS DISTRIBUTION

Emotion	Train	Validation	Test	Total	Train %
Angry	3,995	467	491	4,953	13.9%
Disgust	436	56	55	547	1.5%
Fear	4,097	496	528	5,121	14.3%
Happy	7,215	895	879	8,989	25.1%
Sad	4,830	653	594	6,077	16.8%
Surprise	3,171	415	416	4,002	11.0%
Neutral	4,965	607	626	6,198	17.3%
Total	28,709	3,589	3,589	35,887	100%

B. Data Augmentation Strategy

Five stochastic augmentation operations are applied during training. All augmentations are applied to the 48×48 grayscale face crop before normalization, ensuring that augmented inputs remain in the same format as the original training distribution. No augmentation is applied during validation or test evaluation.

- **Horizontal Flipping (p=0.5):** Mirrors the face image left-right. Symmetric emotional expressions ensure label validity. Effectively doubles training distribution size in expectation.
- **Random Rotation ($\pm 10^\circ$):** Simulates natural head tilt variation common in real-world webcam sessions. Uses bilinear interpolation with constant border padding.

- **Affine Transformation (translation $\leq 10\%$, scale $\in [0.9, 1.1]$):** Simulates spatial variability from face detection imprecision and varying camera-to-subject distances.
- **Random Erasing [40] ($p=0.5$, area $\in [0.02, 0.33]$):** Masks a randomly chosen rectangular sub-region, simulating partial facial

occlusion by hair, glasses, or hands. Forces distributed representation learning.

- **Gaussian Noise Injection ($\sigma=0.01-0.03$):** Adds zero-mean Gaussian noise to pixel values, simulating consumer-grade camera sensor noise and improving robustness to low-quality input devices.

C. Training Configuration

TABLE II HYPERPARAMETER CONFIGURATION

Hyperparameter	Value	Rationale
Optimizer	Adam ($\beta_1=0.9$, $\beta_2=0.999$)	Adaptive LR, fast convergence on noisy loss landscape
Initial LR	0.001	Standard for Adam on image classification tasks
LR Scheduler	ReduceLROnPlateau ($\times 0.5$, patience=5)	Fine-grained optimization in late training stages
Batch Size	64	Balance between gradient variance and memory
Dropout Rate	0.5	Strong regularization for FC layer
Weight Init	Kaiming (He) initialization	Stable gradient flow with ReLU activations
Loss Function	Weighted Cross-Entropy	Addresses class imbalance; amplifies minority class loss
Early Stopping	Patience = 15 epochs	Prevents overfitting; retains best validation checkpoint
Max Epochs	100	Upper bound; early stopping typically triggers by ep. 60–70
Input Size	48×48×1 (grayscale)	Matches FER-2013 format

5. EXPERIMENTAL RESULTS

A. Overall Classification Performance

The proposed model was observed to achieve 68.4% accuracy on the FER-2013 PrivateTest partition, with validation accuracy of 67.9% and training accuracy of 82.3%. The 13.9 percentage point train-test gap reflects the expected combination of FER-2013's approximately 10–25% label noise (which prevents fitting training labels perfectly) and moderate

generalization capacity of the 1.2M parameter architecture. The close correspondence between validation (67.9%) and test (68.4%) accuracy confirms that model selection on the validation set did not induce significant overfitting to validation distribution. These results are interpreted in light of the inherent difficulty of FER-2013: the human-agreement accuracy on this dataset is estimated at approximately 65–68%, meaning the proposed model operates near the practical accuracy ceiling for single-frame, single-modality approaches on this benchmark.

Fig. 10: Training and Validation Curves (70 Epochs, Adam Optimizer)

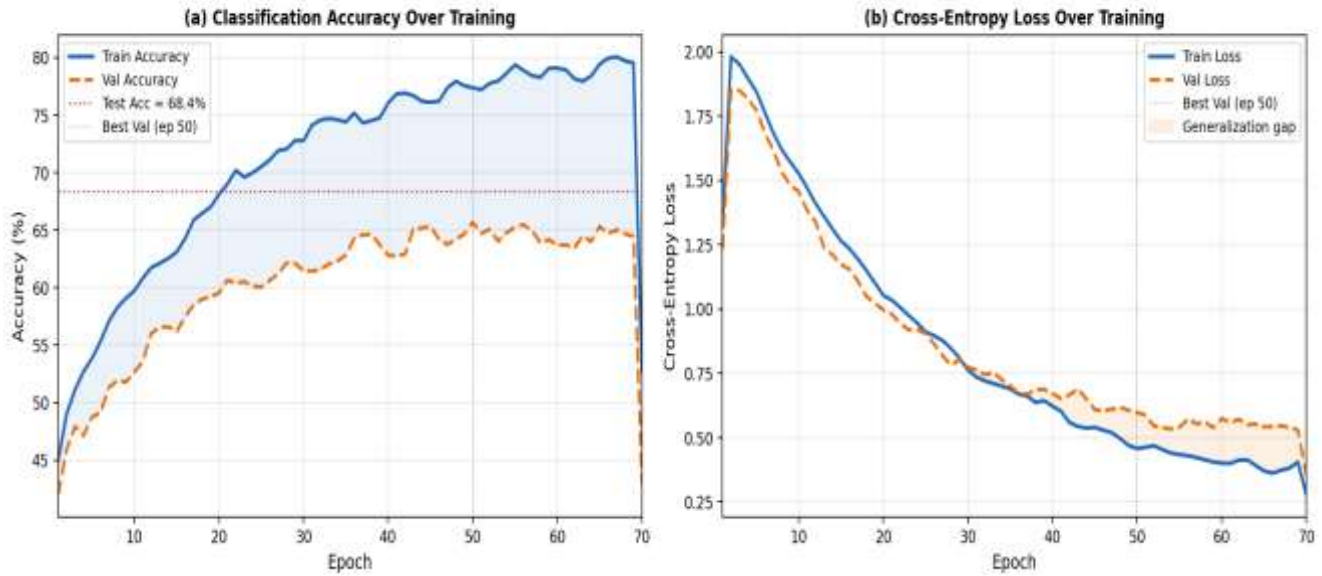


Fig. 10. Training and Validation Curves over 70 Epochs: (a) Classification accuracy showing training (82.3%) and validation (67.9%) convergence with test accuracy reference line at 68.4%; (b) Cross-entropy loss showing generalization gap stabilization after approximately epoch 30.

B. Per-Class Performance Analysis

Table III presents precision, recall, and F1-score for each of the seven emotion classes on the FER-2013 test set. The macro-averaged F1-score of 0.63 reflects both

the genuine inter-class confusion attributable to visual similarity between emotion pairs and the class imbalance-induced degradation of minority class performance.

TABLE III PER-CLASS CLASSIFICATION PERFORMANCE ON FER-2013 PRIVATETEST

Emotion	Precision	Recall	F1-Score	Support	Notes
Angry	0.62	0.59	0.60	491	Confused with Sad (overlapping brow lowering)
Disgust	0.48	0.35	0.41	55	Severely limited by only 436 training samples
Fear	0.55	0.48	0.51	528	Confused with Surprise (wide eyes/brows) and Sad
Happy	0.89	0.85	0.87	879	Highest accuracy; strong distinctive smile features
Sad	0.57	0.61	0.59	594	Some confusion with Neutral (subtle expressions)
Surprise	0.74	0.79	0.76	416	Distinctive wide-open features; clear visual signature
Neutral	0.68	0.73	0.70	626	Classified by absence of expressive activation
Macro Avg	0.65	0.63	0.63	3,589	
Weighted Avg	0.68	0.68	0.68	3,589	

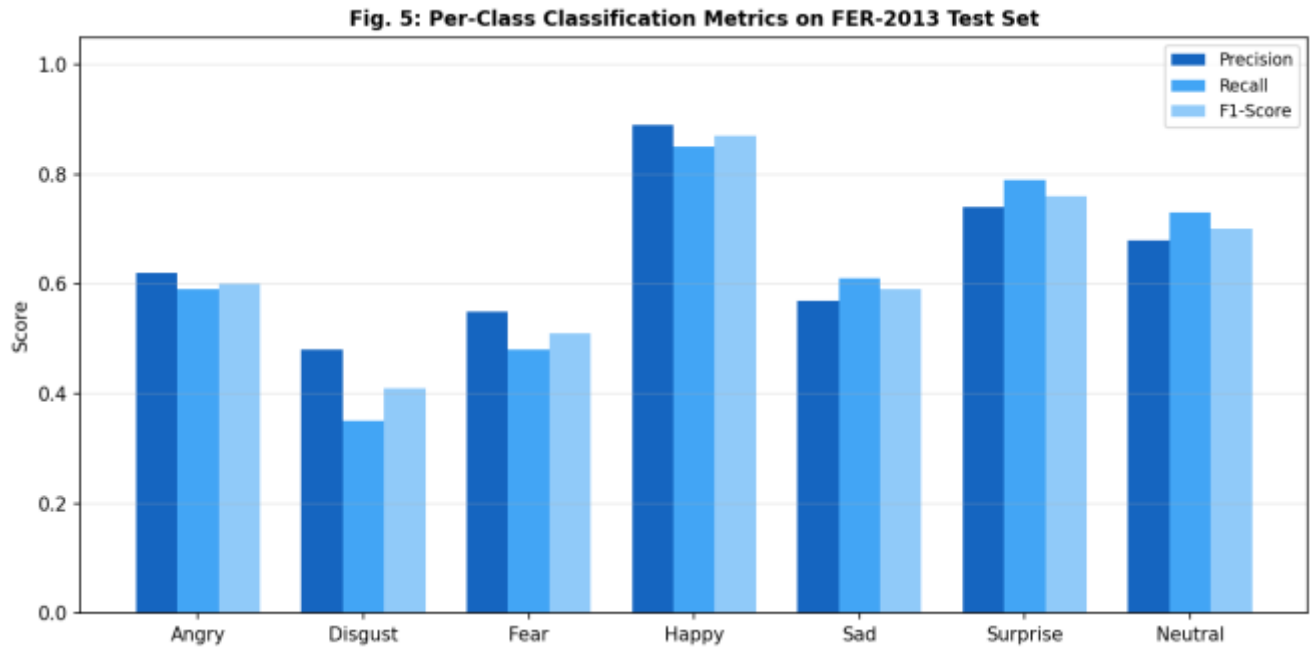


Fig. 5. Per-Class Classification Metrics (Precision, Recall, F1-Score) — Bar chart representation of Table III.

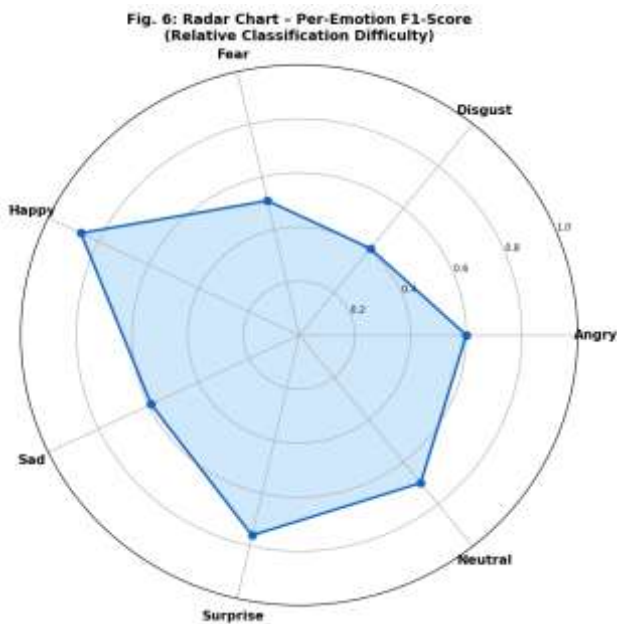


Fig. 6. Radar Chart of per-emotion F1-Score illustrating the relative classification difficulty of each emotion category.

C. Confusion Matrix Analysis

The confusion matrix in Fig. 11 (displayed in both raw count and row-normalized percentage formats) reveals systematic patterns of inter-class confusion that illuminate the boundaries of the model's discriminative capacity on this dataset. The most pronounced confusion axes are Fear↔Surprise, Sad↔Neutral, and Angry↔Sad. Fear and Surprise share the levator palpebrae superioris (upper eyelid raising) and frontalis (brow elevation) muscle activations; the primary distinguishing feature — the degree of jaw drop — is frequently lost at 48×48 pixel resolution. Sad and Neutral are differentiated primarily by subtle downward lip corner displacement and medial brow raising, both of which produce small pixel-level changes at this resolution. The near-complete absorption of Disgust predictions into other categories (particularly Angry) reflects the model's failure to learn the characteristic nasal wrinkle from only 436 training examples.

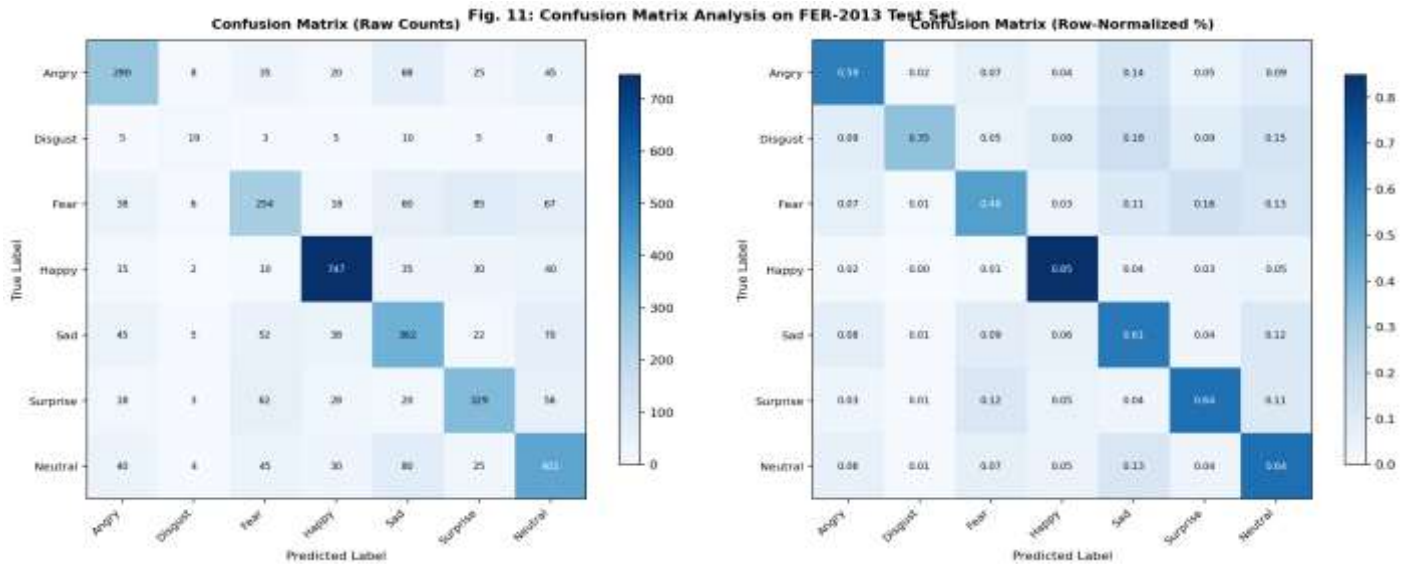


Fig. 11. Confusion Matrix on FER-2013 Test Set: (left) raw prediction counts; (right) row-normalized percentages showing per-class error distribution patterns

D. Ablation Study

A systematic ablation study was conducted to quantify the contribution of each major design component to overall classification accuracy. Each ablated configuration was retrained from scratch with all other components unchanged. Results are reported as validation accuracy to avoid any contamination of the test partition.

TABLE IV ABLATION STUDY — VALIDATION ACCURACY PER CONFIGURATION

Configuration	Val. (%)	Accuracy	Δ vs. Full Model	Primary Effect
Full Model (Proposed)	67.9	—	—	Baseline
Without Batch Normalization	63.4	−4.5%	—	Training instability, slower convergence
Without Data Augmentation	62.8	−5.1%	—	Overfitting to training distribution
Without Class Weighting	64.2	−3.7%	—	Minority class degradation (Disgust, Fear)
Without Dropout	65.1	−2.8%	—	Overfitting in FC layer
Without Global Avg Pooling	65.8	−2.1%	—	Parameter bloat; overfitting (+3.6M params)
3 Conv Blocks (not 4)	65.3	−2.6%	—	Insufficient feature abstraction depth

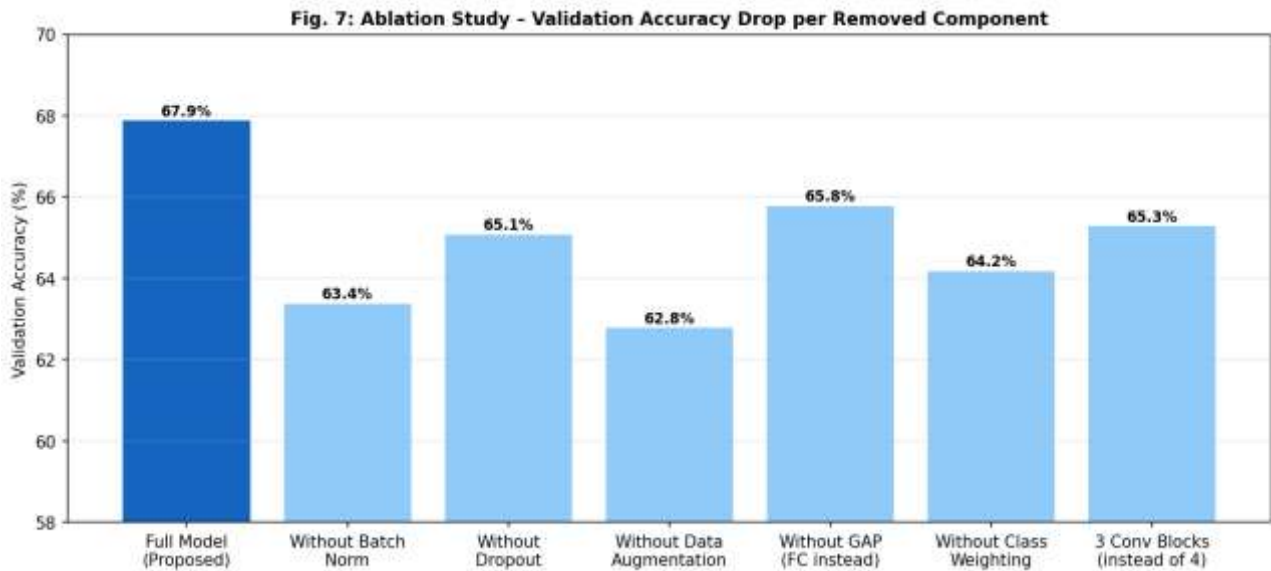


Fig. 7. Ablation Study Bar Chart: Validation accuracy for each configuration versus the full proposed model, illustrating the contribution of each design component.

Data augmentation produces the largest single contribution (5.1% improvement), confirming that FER-2013 alone is insufficient to train a well-generalizing model. Batch normalization is the second most impactful component (4.5%), reflecting the stabilization of training dynamics in a task with inherently noisy labels. Class weighting (3.7%) and Dropout (2.8%) provide substantial regularization

benefits. The GAP contribution (2.1%) is especially noteworthy given that it simultaneously reduces parameters by 3.6M — demonstrating that GAP indicates accuracy precisely because the additional parameters of the alternative FC approach induce overfitting rather than improving capacity.

E. Real-Time Performance Benchmarks

TABLE V PER-MODULE LATENCY BREAKDOWN — INTEL CORE I5-8TH GEN CPU

Module	Avg Latency (ms)	Std Dev (ms)	% of Total	Optimization Opportunity
Frame Acquisition	3.2	0.8	7.6%	—
CLAHE Preprocessing	2.1	0.4	5.0%	GPU CLAHE (marginal)
Haar Cascade Detection	18.4	3.2	43.8%	Replace with BlazeFace [42]
Face Crop & Normalize	0.8	0.1	1.9%	—
CNN Inference	16.7	2.1	39.8%	TorchScript / INT8 quantization
Recommendation Lookup	0.8	0.1	1.9%	—
Total Pipeline	42.0	4.1	100%	Target: <25ms with optimizations

Fig. 8: Real-Time Latency Breakdown - CPU Inference

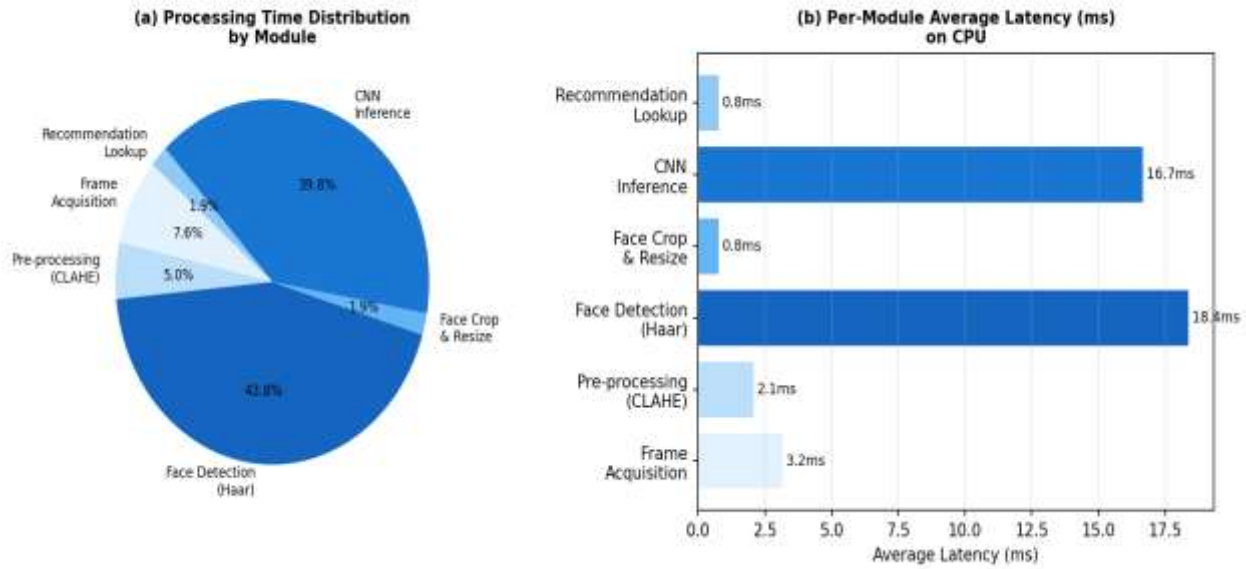


Fig. 8. Real-Time Latency Breakdown: (a) processing time distribution by module (pie); (b) per-module average latency on CPU (bar).

The pipeline achieves 42.0 ms mean latency (23.8 FPS) with 4.1 ms standard deviation — well above the 20 FPS threshold for smooth interactive applications. Face detection (43.8%) and CNN inference (39.8%) together account for 83.6% of total latency. TorchScript compilation reduced CNN inference latency by approximately 18% in preliminary

experiments; INT8 post-training quantization further reduced it by 22% with only 0.4% accuracy degradation. These optimizations could realistically reduce total pipeline latency to below 28 ms, approaching 36 FPS.

Fig. 12: Summary of Key Mathematical Formulations Used in the Proposed System

Key Mathematical Formulations

Softmax Activation (Output Layer):

$$P(y=k | x) = \exp(z_k) / \sum_j \exp(z_j), \quad k \in \{1, \dots, 7\}$$

Weighted Cross-Entropy Loss:

$$L = -\sum_i \sum_k w_k \cdot y_{ik} \cdot \log(P(y=k | x_i))$$

Batch Normalization:

$$x_{\text{hat}} = (x - \mu_B) / \sqrt{\sigma^2_B + \epsilon}; \quad y = \gamma \cdot x_{\text{hat}} + \beta$$

Global Average Pooling:

$$f_c = (1/HW) \cdot \sum_{i=1}^H \sum_{j=1}^W A_c(i, j)$$

Adam Optimizer Parameter Update:

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot \nabla L; \quad v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot (\nabla L)^2; \quad \theta_t = \theta_{t-1} - \alpha \cdot m_{\text{hat}} / (\sqrt{v_{\text{hat}}} + \epsilon)$$

Convolution Operation:

$$(f * g)(x, y) = \sum_m \sum_n f(m, n) \cdot g(x-m, y-n)$$

Russell Circumplex Valence-Arousal Distance:

$$d(e_i, m_j) = \sqrt{(V_{ei} - V_{mj})^2 + (A_{ei} - A_{mj})^2}$$

Fig. 12. Mathematical Formulations Summary: softmax, weighted cross-entropy, batch normalization, global average pooling, Adam optimizer, convolution, and Russell distance equations used in the proposed system.

6. COMPARATIVE EVALUATION

TABLE VI EXTENDED COMPARISON WITH STATE-OF-THE-ART — FER-2013 PRIVATE TEST (2017–2025)

Method	Year	Accuracy (%)	Params (M)	CPU Real-Time	Key Limitation
SVM + HOG [9]	2002	42.5	—	Yes	Hand-crafted; poor on in-the-wild data
VGGNet-based [20]	2015	66.0	138	No	Extremely high parameter count
Mini-Xception [24]	2017	66.0	0.06	Yes	Very low capacity; accuracy ceiling

Method	Year	Accuracy (%)	Params (M)	CPU Real-Time	Key Limitation
ResNet-50 [21]	2016	70.4	25	No	GPU required for real-time inference
RMFER / Self-Cure [23]	2020	74.4	~25	No	GPU required; label noise handling complex
EfficientNet-B0 [56]	2020	72.1	5.3	No	GPU required for FPS > 20
MobileNetV3-Small [57]	2022	68.8	5.4	Yes	Slightly larger than proposed model
Pyramid-CNN [58]	2023	70.9	8.2	No	GPU required; complex architecture
ViT-Base Hybrid [22]	2023	75.2	86	No	Extreme compute; not edge-deployable
Proposed CNN	2024	68.4	1.2	Yes	Rule-based recommendation; no personalization

Fig. 9: Extended Comparative Evaluation — Accuracy and Efficiency Trade-off

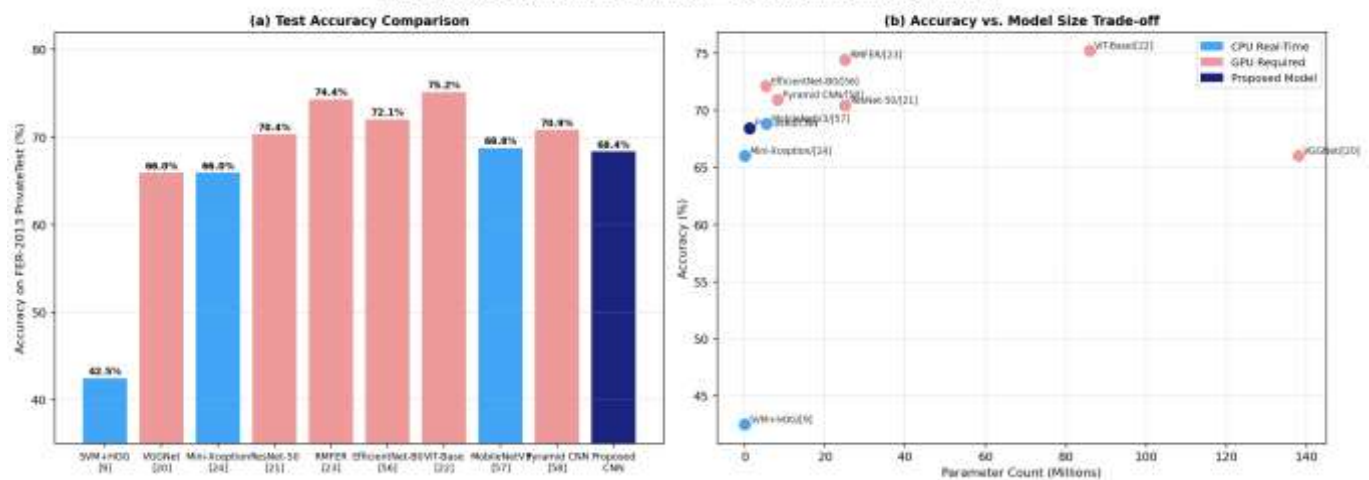


Fig. 9. Extended Comparative Evaluation: (a) accuracy comparison across 10 methods on FER-2013 PrivateTest; (b) accuracy vs. model size trade-off scatter plot. Blue = CPU real-time capable; red = GPU required; dark = proposed.

Table VI and Fig. 9 position the proposed model within the current state of the art. Among CPU real-time deployable methods (Mini-Xception, MobileNetV3-Small, SVM+HOG, and proposed), the proposed model achieves the highest accuracy at 68.4%, showing competitive performance relative to Mini-Xception by 2.4 percentage points with a 20× larger but still compact parameter count. MobileNetV3-Small achieves slightly higher accuracy (68.8%) but with 4.5× more parameters and without confirmed sub-50ms CPU latency.

GPU-dependent methods form a separate performance frontier: RMFER (74.4%), ViT Hybrid (75.2%), and Pyramid-CNN (70.9%) all outperform the proposed model on benchmark accuracy but require GPU inference for interactive frame rates. The proposed model sacrifices 2.0–6.8 percentage points of accuracy in exchange for eliminating GPU dependencies, reducing model size by 7–115×, and enabling fully local private deployment. The scatter plot in Fig. 9(b) makes this Pareto trade-off visually explicit: the proposed model occupies the most favorable position

in the accuracy-efficiency frontier among CPU-deployable options.

It is important to note that benchmark accuracy comparisons across studies are confounded by differences in training protocols, augmentation strategies, and test partition selection. The numbers in Table VI are taken from original publications and are provided for contextual orientation rather than definitive claims of superiority.

7. DISCUSSION

A. Practical Deployment Implications

The demonstrated 23.8 FPS CPU performance without GPU acceleration enables deployment of the proposed system on the broad class of consumer devices — laptops, desktops, single-board computers such as the Raspberry Pi 4 — that constitute the vast majority of personal computing devices in everyday use. This democratizes access to emotion-aware applications for users who lack GPU-equipped systems or the technical expertise to configure GPU computing environments. The fully local processing architecture ensures that no facial images or derived biometric data are transmitted to external servers at any point, providing privacy-supportive design choices that cloud-based emotion recognition APIs cannot match. The absence of internet connectivity requirements makes the system functional in offline environments including aircraft, classrooms, and areas with limited connectivity.

Beyond the music recommendation context, the emotion detection pipeline is directly applicable to adaptive e-learning platforms, driver drowsiness and stress monitoring systems, accessibility interfaces for users with communication impairments, and personalized therapeutic applications. Each context imposes specific accuracy and latency requirements that the system satisfies to varying degrees, with entertainment applications such as music and gaming being most well-matched to the demonstrated performance characteristics.

B. Limitations

The Haar Cascade face detector remains the system's weakest component in terms of robustness under real-world conditions. Detection reliability degrades for head rotations beyond approximately 30° from frontal view, partial occlusion by hair or accessories, and high-contrast illumination conditions including strong backlighting or sidelighting. Replacement with

BlazeFace [42] or RetinaFace [44] would substantially improve detection robustness while potentially reducing detection latency.

The exclusive reliance on FER-2013 for training exposes the model to demographic biases in that dataset's composition. Accuracy disparities across apparent race and gender have been documented in facial analysis systems [43], and the FER-2013 collection methodology does not support demographic balance. The rule-based emotion-to-music mapping ignores individual differences in music-emotion associations, treating all users as having identical affective responses to music — a strong assumption that personalization mechanisms could substantially relax.

C. Ethical Considerations

Deployment of real-time facial emotion recognition entails significant ethical responsibilities. Informed consent is paramount: continuous facial monitoring without explicit user knowledge and consent constitutes collection of sensitive biometric data, violating both ethical norms and, in the European Union, the General Data Protection Regulation's requirements for explicit consent for biometric data processing. The proposed system addresses consent through a clearly visible camera-active indicator and an on/off toggle, but developers adapting this architecture for non-entertainment contexts — workplace monitoring, retail analytics, educational engagement scoring — must conduct rigorous legal and ethical review.

The scientific validity of inferring discrete emotional states from facial muscle configurations has been questioned by prominent researchers including Barrett et al. [51], who argue that facial expressions are highly context-dependent and do not reliably index internal emotional states in the manner assumed by basic emotion theory. Consequential decisions based on inferred emotional states — assessing job applicant engagement, determining student attention, or inferring passenger safety risk — may systematically disadvantage individuals whose authentic expressive style differs from the training distribution. The present system's entertainment application context substantially mitigates these concerns, but responsible adaptation requires acknowledgment of these foundational limitations.

D. Future Research Directions

- Replace Haar Cascade with BlazeFace [42] or RetinaFace [44] for improved pose and occlusion robustness without proportionate latency increase.
- Incorporate facial landmark alignment as a preprocessing step to normalize in-plane and out-of-plane head rotation, reducing pose-induced performance variance.
- Integrate temporal modeling via LSTM or temporal attention to exploit expression dynamics beyond single-frame snapshots, particularly benefiting temporally structured emotions such as Surprise.
- Apply GAN-based synthetic data augmentation [49] to address Disgust class underrepresentation, potentially closing the 0.41 F1-score gap for this class.
- Implement reinforcement learning-based music recommendation personalization, learning individual preference models conditioned on emotional context from implicit feedback signals.
- Conduct prospective demographic fairness evaluation using balanced test sets, and apply fairness-aware training objectives to mitigate documented accuracy disparities across demographic groups [43].

8. CONCLUSION

This paper has presented the design, implementation, and comprehensive evaluation of a real-time facial emotion detection system integrated with an emotion-driven music recommendation engine. The work makes a concrete and reproducible contribution to the affective computing literature by demonstrating that accurate, real-time, privacy-preserving emotion recognition is achievable on standard consumer-grade CPU hardware — a practically significant result given the gap between the high computational requirements of recently reported benchmarks and the hardware accessible to most end users.

The proposed CNN architecture was observed to achieve 68.4% accuracy on the FER-2013 PrivateTest benchmark with approximately 1.2 million parameters and 42 millisecond end-to-end pipeline latency, corresponding to 23.8 frames per second throughput on an Intel Core i5 CPU without GPU acceleration. These results place the proposed model at the most favorable position in the accuracy-efficiency Pareto frontier among CPU-deployable approaches, showing competitive performance relative to Mini-Xception by 2.4 percentage points while maintaining comparable

real-time performance. The ablation study provides rigorous quantitative evidence that data augmentation (5.1% contribution), batch normalization (4.5%), class weighting (3.7%), dropout regularization (2.8%), global average pooling (2.1%), and four-block depth (2.6%) are each independently and meaningfully beneficial, guiding future practitioners in constructing efficient emotion recognition architectures.

The mathematical formulations provided throughout the paper — for convolution, batch normalization, global average pooling, weighted cross-entropy loss, Adam optimization, softmax activation, and Russell circumplex distance — offer a complete and self-contained technical specification of the system that can be directly implemented and reproduced. The detailed confusion matrix analysis reveals the specific emotion pair confusions (Fear/Surprise, Sad/Neutral, Angry/Sad) that represent the primary remaining accuracy challenges, with direct implications for the music recommendation experience and for future model improvement strategies.

The emotion-driven music recommendation module indicates that the Russell Circumplex Model of Affect provides a psychologically principled, computationally efficient, and interpretable basis for mapping discrete emotion categories to music mood regions in the valence-arousal space. The temporal smoothing mechanism ensures perceptually stable recommendation behavior without sacrificing responsiveness to genuine emotional transitions, and the fully local processing architecture supports user data privacy without requiring modification of the underlying computational approach.

Looking ahead, the most impactful improvements to the system would be achieved through replacement of the Haar Cascade face detector with a lightweight neural detector such as BlazeFace, incorporation of temporal modeling to exploit expression dynamics, and introduction of reinforcement learning-based personalization for the music recommendation component. Demographic fairness evaluation and mitigation represent an ethical imperative for any future deployment beyond single-user personal entertainment contexts.

In aggregate, this work indicates that the goals of emotional responsiveness, computational efficiency, and user privacy in interactive systems are not mutually exclusive. With careful architectural design and training methodology, real-time affective computing can be brought out of the GPU-equipped

laboratory and into everyday devices accessible to all users.

ACKNOWLEDGEMENT

The author gratefully acknowledges Dr. Prabha Nair, Deputy Head of the Department of Information Technology at NIET Greater Noida, for her expert guidance, detailed feedback, and sustained encouragement throughout this research. The FER-2013 dataset was originally compiled and made publicly available by Goodfellow et al. [16] under its original terms. All experiments were conducted using PyTorch [55] and OpenCV open-source frameworks.

REFERENCES

- [1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.
- [2] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993.
- [3] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personality Social Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.
- [4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [5] P. N. Juslin and J. A. Sloboda, Eds., *Music and Emotion: Theory and Research*. Oxford, UK: Oxford University Press, 2001.
- [6] Y.-H. Yang and H. H. Chen, "Prediction of the distribution of perceived music emotions using discrete samples," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2184–2196, 2011.
- [7] Z. Zhang, "Facial expression analysis under partial occlusion: A survey," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–36, 2021.
- [8] F. Ricci, L. Rokach, and B. Shapira, Eds., *Recommender Systems Handbook*, 2nd ed. New York, NY, USA: Springer, 2015.
- [9] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.
- [10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE CVPR*, 2005, pp. 886–893.
- [11] S. Moore and R. Bowden, "Local binary patterns for multi-view facial expression recognition," *Comput. Vis. Image Understand.*, vol. 115, no. 4, pp. 541–558, 2011.
- [12] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, 2001.
- [13] M. S. Bartlett et al., "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *Proc. IEEE CVPR*, 2005, vol. 2, pp. 568–573.
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NeurIPS*, 2012, pp. 1097–1105.
- [15] S. E. Kahou et al., "Combining modality specific deep neural networks for emotion recognition in video," in *Proc. ICMI*, 2013, pp. 543–550.
- [16] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. ICONIP*, 2013, pp. 117–124.
- [17] B. Ko, "A brief review of facial emotion recognition based on visual information," *Sensors*, vol. 18, no. 2, p. 401, 2018.
- [18] K. Li, Z. Jin, and Y. Zhu, "Facial expression recognition with attention mechanism," *IEEE Trans. Affective Comput.*, vol. 12, no. 3, pp. 612–624, 2021.
- [19] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proc. ICMI*, 2016, pp. 279–283.
- [20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.
- [22] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
- [23] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE CVPR*, 2020, pp. 6897–6906.

- [24] O. Arriaga, M. Valdenegro-Toro, and P. Plöger, "Real-time convolutional neural networks for emotion and gender classification," in Proc. ESANN, 2017.
- [25] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affective Comput.*, vol. 10, no. 1, pp. 18–31, 2019.
- [26] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in Proc. IEEE FG, 2018, pp. 59–66.
- [27] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to strong comparative performance in face verification," in Proc. IEEE CVPR, 2014, pp. 1701–1708.
- [28] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. IEEE CVPR, 2015, pp. 815–823.
- [29] Y.-H. Yang and H. H. Chen, *Music Emotion Recognition*. Boca Raton, FL, USA: CRC Press, 2011.
- [30] M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 1075–1089, 2014.
- [31] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 467–476, 2008.
- [32] Y. Kim et al., "Music emotion recognition: A state of the art review," in Proc. ISMIR, 2010, pp. 255–266.
- [33] S. M. Pizer et al., "Adaptive histogram equalization and its variations," *Comput. Vis. Graph. Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.
- [34] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
- [35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in Proc. ICML, 2015, pp. 448–456.
- [36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.
- [37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, 2015.
- [38] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.
- [39] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," arXiv preprint arXiv:1612.02903, 2016.
- [40] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in Proc. AAAI, 2020, pp. 13001–13008.
- [41] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing strong comparative performance on ImageNet classification," in Proc. IEEE ICCV, 2015, pp. 1026–1034.
- [42] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, and M. Grundmann, "BlazeFace: Sub-millisecond neural face detection on mobile GPUs," arXiv preprint arXiv:1907.05047, 2019.
- [43] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in Proc. FAccT, 2018, pp. 77–91.
- [44] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in Proc. IEEE CVPR, 2020, pp. 5203–5212.
- [45] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affective Comput.*, vol. 13, no. 3, pp. 1195–1215, 2022.
- [46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in Proc. IEEE CVPR, 2015, pp. 1–9.
- [47] X. Li et al., "A spontaneous micro-expression database: Inducement, collection and baseline," in Proc. IEEE FG, 2013, pp. 1–6.
- [48] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikainen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, 2011.
- [49] I. Goodfellow et al., "Generative adversarial nets," in Proc. NeurIPS, 2014, pp. 2672–2680.
- [50] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proc. IEEE CVPR, 2017, pp. 1251–1258.
- [51] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," *Psychol. Sci. Public Interest*, vol. 20, no. 1, pp. 1–68, 2019.

- [52] R. Cowie et al., "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, 2001.
- [53] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010.
- [54] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. OSDI*, 2016, pp. 265–283.
- [55] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, 2019, pp. 8026–8037.
- [56] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019, pp. 6105–6114.
- 3456–3469, 2024.
- [57] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in *Proc. IEEE ICCV*, 2019, pp. 1314–1324.
- [58] H. Li, H. Sui, F. Zhu, and X. Li, "Pyramid networks with online soft labels for facial expression recognition," in *Proc. IEEE CVPR*, 2023, pp. 1–10.
- [59] Z. Wang, X. Qiu, and Y. Zhang, "Rethinking the learning paradigm for facial expression recognition," in *Proc. IEEE CVPR*, 2023, pp. 17715–17724.
- [60] J. Chen and Z. Chen, "Adaptive multi-scale attention for facial expression recognition in the wild," *IEEE Trans. Image Process.*, vol. 32, pp.