

Keywords: facial emotion recognition, FER2013, wavelet pooling, cross-attention, e-learning

Mustapha Abdelkader LAOUMIR ^{1,2*}, Amina KINANE DAOUADJI ^{1,2},
 Fatima BENDELLA ^{1,2}

¹ University of Science and Technology of Oran Mohamed Boudiaf (USTO-MB), Oran, Algeria,
 mustaphaabdeldkader.laoumir@univ-usto.dz, amina.kinane@univ-usto.dz, fatima.bendella@univ-usto.dz

² Department of Mathematics and Computer Science, Signal Image Parole (SIMP) Laboratory, University of Science and Technology of Oran Mohamed Boudiaf (USTO-MB), Oran, Algeria

* Corresponding author: mustaphaabdeldkader.laoumir@univ-usto.dz

ARECA-Lite: A lightweight modified ArecaNet with reduced complexity for real-time robust facial emotion recognition

Abstract

Facial emotion recognition (FER) in e-learning environments faces a persistent tension between recognition performance and deployment efficiency; high-performing architectures typically impose prohibitive computational costs that preclude real-time use on commodity hardware. This paper presents ARECA-Lite, a lightweight variant of ArecaNet designed to resolve this tension in resource-constrained deployments. ARECA-Lite integrates a truncated MobileNetV3-Small backbone with dual VGGFace2-pretrained InceptionResNetV1 sub-branches, a wavelet pooling module for structured frequency-domain decomposition, and a modified Assembled Residual Enhanced Cross-Attention (A.R.E.C.A.) module composed of two parallel RECA blocks. A FACS-guided offline preprocessing pipeline, targeted disgust oversampling, and a composite Dice-BCE loss collectively address class imbalance and low-contrast action unit (AU) evidence. Experiments are conducted on FER2013. ARECA-Lite achieves 82.41% accuracy and an 82.50% macro F1-score, outperforming a retrained ArecaNet baseline under the same experimental conditions by 7.10 and 9.43 percentage points, respectively, while reducing parameter count from 24.95 M to 4.69 M, GPU inference latency from 21.83 ms to 5.44 ms, and computational cost from 6.57 GFLOPs to 1.64 GFLOPs. The preprocessing and augmentation pipeline adds 14.88 percentage points in accuracy relative to raw-data training. Evaluation is limited to FER2013; broader validation is left to future work. These results demonstrate that competitive, class-balanced facial emotion recognition is achievable without the substantial computational overhead of larger attention-based models, making ARECA-Lite well-suited to low-latency affective computing on commodity hardware.

1. INTRODUCTION

The large-scale virtualisation of instruction catalysed by the COVID-19 pandemic has fundamentally restructured educational delivery worldwide. While this transition has expanded access to education across geographical and institutional boundaries, it has introduced a significant pedagogical deficit to which the technical literature has paid insufficient attention (Wang et al., 2025; Kim & Choi, 2025). In a conventional classroom, an experienced instructor continuously and implicitly monitors the emotional states of the student cohort: facial expressions of confusion, frustration, disengagement, or comprehension are decoded in real time and used to modulate the pace, register, and content of the lesson. This affective feedback loop is not a peripheral feature of effective pedagogy; it is central to it. In a virtual learning environment mediated by commodity cameras and videoconferencing software, this loop is effectively severed.

Automated facial emotion recognition (FER) systems, deployed at the edge within e-learning platforms, offer a principled and scalable solution to this problem. Kinane Daouadji and Bendella (2024) demonstrated an early instance of this approach by developing a real-time CNN-based system that detects learners' emotions during synchronous e-learning sessions and communicates engagement states to the instructor, enabling adaptive in-course teaching decisions. A model capable of inferring the affective state of each learner in real time, from the commodity hardware already present in most households, could restore the affective awareness that virtual instruction lacks: flagging disengaged students, surfacing confusion before it compounds, and

providing instructors and adaptive platforms with the continuous emotional signal necessary to personalise instruction at scale. Beyond education, the same capability underpins a broad range of affective computing applications, including human-computer interaction, driver drowsiness monitoring, clinical pain assessment, and social robotics (Saurav et al., 2025; Khan et al., 2025).

The scientific study of automatic facial expression analysis has a rich trajectory spanning several decades. Pre-deep-learning approaches relied on the manual engineering of feature descriptors such as Local Binary Patterns (LBP), Gabor filter banks, and Active Appearance Models (AAMs), which proved brittle under real-world conditions. The introduction of deep Convolutional Neural Networks (CNNs) marked a decisive inflexion point, and subsequent advances incorporated spatial attention mechanisms and Vision Transformers capable of modelling long-range spatial dependencies (Kim & Choi, 2025; Wang et al., 2025). Yet this trajectory of architectural elaboration has exacted an escalating computational toll. State-of-the-art models routinely accumulate tens to hundreds of millions of trainable parameters and demand dedicated GPU infrastructure to sustain acceptable inference throughput. The prevailing convention, that accuracy is the primary optimisation target and computational cost a secondary concern, reflects a misalignment with the practical requirements of edge-deployed affective computing.

The challenge is compounded by the properties of the FER2013 benchmark dataset (Goodfellow et al., 2013), which remains the community standard for controlled evaluation. FER2013 is characterised by three well-documented pathologies: (1) greyscale thumbnails at 48×48 pixels, (2) acute class skew with a maximum imbalance ratio of $16.5\times$, and (3) a semi-automatic annotation protocol with a reported human-level accuracy of only about $65 \pm 5\%$, indicating substantial label ambiguity and noise. These characteristics make it simultaneously the field's most widely adopted benchmark and one of its most intractable obstacles.

To address this confluence of challenges, the present work introduces ARECA-Lite, a lightweight modified ArecaNet architecture purpose-designed for real-time facial emotion recognition under the resource constraints characteristic of e-learning deployment. ARECA-Lite is based on the premise that the tension between computational efficiency and recognition accuracy can be substantially reduced: through disciplined architectural design, combining an efficient mobile backbone, frequency-domain feature decomposition via Wavelet Pooling, cross-attention refinement through a reduced two-block A.R.E.C.A. module, and identity-discriminative transfer learning from VGGFace2-pretrained sub-branches, it is possible to substantially improve the balance between recognition performance and computational efficiency on FER2013 while remaining deployable on commodity hardware without specialised accelerators.

The remainder of this paper is organised as follows. Section 2 surveys and critically contextualises related work in facial emotion recognition. Section 3 details the dataset, preprocessing pipeline, model architecture, and training protocol. Section 4 presents quantitative results, ablation studies, a comparative analysis and limitations, followed by concluding remarks in Section 5.

2. RELATED WORK

Recent years have witnessed substantial progress in facial emotion recognition, driven by the rapid development of deep learning architectures and the availability of large-scale annotated datasets. Nevertheless, the literature continues to face persistent challenges in balancing recognition performance with computational efficiency under the constraints of real-world deployment.

Roy et al. (2024) proposed ResEmoteNet, a deep residual architecture that achieves 79.79% accuracy on the FER2013 dataset through carefully designed loss functions. However, the work does not report model parameter count or inference latency, precluding assessment of deployment feasibility on resource-constrained hardware. Gursesli et al. (2024) evaluated a custom lightweight CNN for FER, reporting 63.00% accuracy on FER2013 with 2.3 M parameters and 50 ms latency. While the compact design confirms deployability, the limited accuracy underscores the inadequacy of simple convolutional architectures without complementary attention mechanisms. Liao et al. (2024) introduced RS-Xception, which employs depthwise separable convolutions and achieves 69.02% accuracy on FER2013 with 1.92 M parameters and 15.9 ms latency.

Saurav et al. (2025) proposed an integrated attention-guided deep CNN (IA-DCNN) that achieved 75.03% accuracy on the FER2013 dataset with 2.93 M parameters and a remarkably low latency of 2.96 ms. Yang et al. (2025) presented a novel lightweight FER network combining deep and shallow feature streams, achieving 69.17% accuracy with 3.27 M parameters. Ramirez-Quintana et al. (2025) proposed an ultra-lightweight CNN incorporating Efficient Channel Attention (ECA), achieving 79.20% accuracy on FER2013 with only 0.042

M parameters and 1.88 ms latency, the most parameter-efficient model in this survey. Khan et al. (2025) introduced an Attention-Enhanced Deep Ensemble Network (AE-DEN), achieving 78.60% accuracy on the FER2013 dataset, demonstrating that ensemble strategies can recover accuracy without requiring large individual model capacity.

Wang et al. (2025) developed an Attention-ResNet optimisation framework for real-time emotion recognition in classroom environments, achieving 86.26% accuracy on the FER2013 dataset with 21.3 M parameters. Sun et al. (2026) proposed a lightweight, real-time FER system based on an improved Mini-Xception algorithm, achieving 79.96% accuracy with only 0.055 M parameters, but incurring a high latency of 104 ms. Kim & Choi (2025) introduced ArecaNet, an Assembled Residual Enhanced Cross-Attention Network that combines InceptionResNetV1 sub-branches pretrained on VGGFace2 with a MobileNetV3 main backbone and a multi-block cross-attention module for robust FER. ArecaNet reports high recognition performance with 24.95 M parameters and 6.57 GFLOPs. The present work, ARECA-Lite, directly inherits the core cross-attention architectural innovation of ArecaNet while fundamentally reducing its complexity through a streamlined two-block cross-attention design, lightweight backbone prioritisation, and wavelet-domain feature routing. Kim & Choi (2025) report 97.00% accuracy on FER2013; our independent reproduction yields 75.31%, a discrepancy that may be due to differences in evaluation protocol or data partitioning.

Table 1 provides a consolidated comparison of the surveyed methods across recognition accuracy, model parameter count, and inference latency - the key dimensions relevant to e-learning deployment.

Tab. 1. Summary of reviewed facial emotion recognition methods (n/r: not reported)

Author(s)	Year	Model	Dataset	Acc (%)	Params (M)	Lat. (ms)
Roy et al.	2024	ResEmoteNet	FER2013	79.79	n/r	n/r
Gursesli et al.	2024	LC-CNN	FER2013	63.00	2.30	50.0
Liao et al.	2024	RS-Xception	FER2013	69.02	1.92	15.9
Saurav et al.	2025	IA-DCNN	FER2013	75.03	2.93	2.96
Yang et al.	2025	DSN-AM	FER2013	69.17	3.27	23.0
Ramirez-Quintana et al.	2025	LW-ECA	FER2013	79.20	0.042	1.88
Khan et al.	2025	AE-DEN	FER2013	78.60	n/r	n/r
Wang et al.	2025	Att-ResNet	FER2013	86.26	21.30	n/r
Kim & Choi	2025	ArecaNet	FER2013	97.00 ^a	24.95	n/r
Naik et al.	2026	EffNetB2	FER2013	68.78	9.2	n/r
Sun et al.	2026	Mini-XCep (Impr.)	FER2013	79.96	0.055	104.0
Syabil et al.	2026	CNN	FER2013	55.50	n/r	n/r

^a Reported value from the original study; the reproduced value differs, as discussed in Section 4.4.

3. METHODOLOGY

ARECA-Lite is structured as a four-stage sequential pipeline, as illustrated in Figure 1. In the first stage, three parallel pathways independently extract complementary feature representations from the same preprocessed input image. In the second stage, a wavelet pooling module decomposes the aggregated feature maps into frequency sub-bands, routing low-frequency signals to the attention stage while delivering high-frequency edge sub-bands directly to the classification head via skip connections. In the third stage, a pair of residual-enhanced cross-attention blocks refines the low-frequency tokens by attending to features from the contextual backbone. In the fourth and final stage, a multi-input fusion head integrates all streams and produces a seven-class posterior probability distribution. Each architectural component was selected with explicit consideration of the complexity-accuracy trade-off.

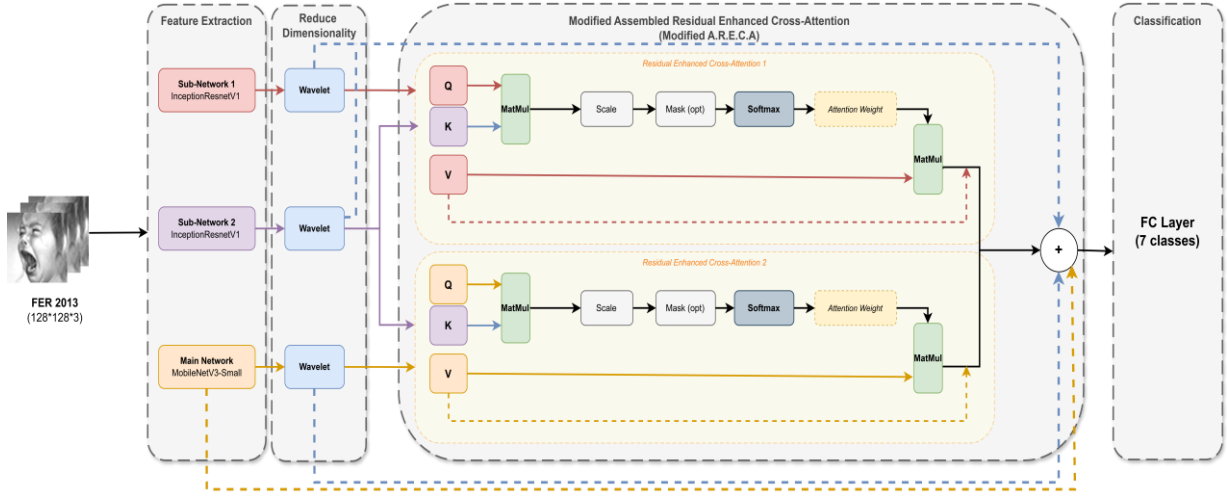


Fig. 1. Overall architecture of ARECA-Lite

3.1. Feature extraction

The feature extraction stage employs three independent pathways that operate concurrently on the same preprocessed input image. This multi-branch design preserves the core representational principle inherited from the original ArecaNet architecture, namely the use of complementary feature representations and A.R.E.C.A.-based aggregation to capture discriminative facial information. Specifically, the architecture combines a lightweight mobile backbone, optimised for spatial efficiency, with two identity-discriminative sub-branches that transfer structural facial knowledge acquired from a large-scale face recognition corpus. However, while these components reflect the conceptual design inherited from ArecaNet, ARECA-Lite introduces several architectural modifications to adapt this principle to the constraints of real-time inference. In particular, the original backbone is replaced by a truncated MobileNetV3-Small configuration, and the feature aggregation stage is simplified to retain a single A.R.E.C.A. module with only two RECA blocks. These modifications substantially reduce the depth and computational complexity of the feature-extraction pipeline while preserving the essential A.R.E.C.A.-based aggregation principle. In summary, the A.R.E.C.A.-based aggregation and the use of complementary feature representations are inherited from ArecaNet, whereas the truncated MobileNetV3-Small backbone and the reduced two-RECA A.R.E.C.A. module are the modifications introduced in ARECA-Lite.

3.1.1. Main network: MobileNetV3-Small

MobileNetV3-Small (Howard et al., 2019) serves as the main backbone of ARECA-Lite, chosen specifically for its emphasis on spatial efficiency, namely the progressive reduction of redundant spatial detail while preserving semantically discriminative facial information. As shown in Table 2, the model does not employ the full standard MobileNetV3-Small classifier. Instead, ARECA-Lite retains a truncated configuration composed of the initial convolution and the first eight inverted residual bottleneck blocks, which together process an input of size $128 \times 128 \times 3$. Although FER2013 images are distributed in greyscale, each image is converted to a three-channel RGB representation during data loading by duplicating the original luminance information across all channels. This transformation does not introduce new chromatic information; rather, it ensures compatibility with the ImageNet-pretrained MobileNetV3-Small backbone and the pretrained face-recognition sub-networks, both of which are architecturally defined for three-channel inputs.

Within this truncated configuration, spatial efficiency is achieved through a controlled sequence of strided depthwise-separable bottlenecks that reduce the feature resolution from 128×128 to 16×16 in Stage 1 and then to 8×8 in Stage 2. Rather than continuing toward the deeper classification layers of the original network, the adapted backbone stops at the $8 \times 8 \times 48$ feature map, which is sufficient for the downstream wavelet projection and attention modules. In parallel, the intermediate $16 \times 16 \times 24$ activation tensor is preserved as skip connection S1, allowing fine-grained spatial detail to be forwarded directly to the fusion head instead of being discarded by further downsampling. The architecture relies on three core functions to further reduce complexity and latency:

- Depthwise Separable Convolution: replaces standard convolutions by factorising them into a depthwise convolution (applied per channel) followed by a pointwise convolution (1×1). This decomposition reduces the number of FLOPs by a factor of approximately (Howard et al., 2017):

$$\frac{1}{N} + \frac{1}{D_k^2} \quad (1)$$

Where N is the number of output channels, and D_k is the kernel size, drastically cutting computational cost while maintaining spatial feature extraction capability.

- Squeeze-and-Excitation (SE) Module (Hu et al., 2018): adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. It applies global average pooling to squeeze spatial information into a compact descriptor, followed by two fully connected layers to excite informative channels and suppress less informative ones, improving accuracy with negligible parameter overhead.
- Nonlinear Activations (RE and HS): the truncated MobileNetV3-Small configuration used in ARECA-Lite employs two activation functions. In the earlier bottleneck blocks, the network uses ReLU (denoted RE in Table 2), defined as $\text{ReLU}(x) = \max(0, x)$, owing to its simplicity and low computational cost. In the deeper layers, RE is replaced by Hard-Swish (HS), which provides smoother gating behaviour while remaining efficient:

$$\text{HS}(x) = x \cdot \frac{\text{ReLU6}(x+3)}{6} \quad (2)$$

Unlike the original Swish activation (Ramachandran et al., 2017), this piecewise-linear approximation avoids expensive exponential operations, thereby reducing inference latency on hardware with limited floating-point capability while preserving the expressive nonlinearity required in deeper feature-extraction stages.

Tab. 2. Truncated MobileNetV3-Small configuration used in ARECA-Lite

Input	Operator	Exp-Size	Out	SE	NL	Stride
$128 \times 128 \times 3$	Conv2D 3×3	–	16	–	HS	2
$64 \times 64 \times 16$	Bottleneck, $k = 3$	16	16	✓	RE	2
$32 \times 32 \times 16$	Bottleneck, $k = 3$	72	24	–	RE	2
$16 \times 16 \times 24$	Bottleneck, $k = 3$	88	24	–	RE	1
Stage 1 output / skip connection S1: $16 \times 16 \times 24$						
$16 \times 16 \times 24$	Bottleneck, $k = 5$	96	40	✓	HS	2
$8 \times 8 \times 40$	Bottleneck, $k = 5$	240	40	✓	HS	1
$8 \times 8 \times 40$	Bottleneck, $k = 5$	240	40	✓	HS	1
$8 \times 8 \times 40$	Bottleneck, $k = 5$	120	48	✓	HS	1
$8 \times 8 \times 48$	Bottleneck, $k = 5$	144	48	✓	HS	1
Stage 2 output / primary MobileNet feature map: $8 \times 8 \times 48$						

Conv2D: 2D convolution layer; Exp-Size: expansion size of the bottleneck block; Out: number of output channels; SE: squeeze-and-excitation module; NL: non-linear activation function; HS: hard-swish; RE: rectified linear unit (ReLU). The symbol “✓” indicates that an SE layer is used in the corresponding bottleneck block, whereas “–” indicates that no SE layer is applied.

3.1.2. Sub-Network: InceptionResNetV1

Two parallel InceptionResNetV1 sub-networks are employed in the proposed architecture, as shown in Table 3, where the same sub-branch design is applied independently to branches A and B. In the implemented model, each sub-network is initialised with weights pre-trained on the large-scale VGGFace2 face recognition corpus (Cao et al., 2018) and is truncated immediately before the mixed_6a reduction block. For an input image of size $128 \times 128 \times 3$, this truncated configuration produces a feature tensor of size $13 \times 13 \times 256$ per branch.

This design preserves structurally informative intermediate facial representations while avoiding the additional computational burden of the deeper classification layers in the full backbone.

Architecturally, InceptionResNetV1 combines the multi-scale representational capacity of Inception modules with the optimisation advantages of residual learning, following the principles introduced by Szegedy et al. (2017). A generic convolutional transformation may be written as:

$$y = W * x + b \quad (3)$$

Where x denotes the input feature map, W the learnable convolution kernel, b the bias term, and $*$ the convolution operator. Within an Inception-style block, several such transformations are computed in parallel at different receptive fields and then concatenated, thereby enabling simultaneous extraction of fine-grained local structure and mid-range spatial context. In practice, this parallel design is further supported by convolutional factorisation, which reduces the parameter count and multiply-accumulate (MAC) cost of large kernels while preserving representational richness. The residual mechanism is equally central to the architecture. Let $F(x)$ denote the nonlinear transformation learned by an Inception-ResNet block; the residual mapping is then expressed as:

$$y = F(x) + x \quad (4)$$

This facilitates gradient propagation, stabilises optimisation, and enables efficient feature reuse across depth. In the present work, the combination of multi-branch convolutions and residual connections is particularly attractive because it provides rich facial structural encoding at a moderate computational cost. This makes the truncated InceptionResNetV1 sub-network well-suited to low-latency inference and to deployment settings constrained by memory, power, or response-time requirements, such as edge, fog, and embedded systems.

The role of this sub-network in the global ARECA-Lite architecture is not to perform direct classification, but rather to transfer structural facial knowledge acquired from a large-scale face recognition corpus into the facial emotion recognition task. Because VGGFace2 pre-training requires discrimination among thousands of identities, the learned filters tend to be sensitive to stable facial geometry, landmark arrangement, contour topology, and fine local morphology. These anatomical priors remain highly relevant for expression analysis, since emotions are visually expressed through structured deformations of the same facial components. By reusing these pre-trained representations, the model benefits from transfer learning in which structural knowledge learned from identity recognition is repurposed to improve emotion discrimination without requiring a comparably large emotion-labelled dataset. Formally, let x denote the common preprocessed input image. The two parallel branches can be written as:

$$f_A = T_A(x; \theta_A), \quad f_B = T_B(x; \theta_B) \quad (5)$$

Where T_A and T_B denote the two truncated InceptionResNetV1 mappings and θ_A, θ_B their learnable parameters. Although both branches share the same architectural form and are initialised from the same VGGFace2 checkpoint, they are fine-tuned independently during training. This allows each branch to specialise toward complementary structural facial cues, thereby enriching the representations available to the downstream wavelet projection and cross-attention modules.

Each sub-branch produces a structural feature map that encodes identity-informed facial representations. These outputs are then forwarded to the subsequent fusion stages of ARECA-Lite, where they interact with the lightweight main backbone through wavelet decomposition and cross-attention refinement. Accordingly, the InceptionResNetV1 sub-networks serve as identity-prior encoders, injecting robust structural facial knowledge into the overall architecture while maintaining an efficiency level compatible with low-latency deployment. The layer-wise configuration retained in the truncated sub-branch is summarised in Table 3. In particular, the final row indicates that after the stem convolutions, the network applies five successive Block35 Inception-ResNet-A residual blocks while preserving the spatial resolution to 13×13 and the channel dimensionality to 256.

Tab. 3. Truncated InceptionResNetV1 configuration used in ARECA-Lite

Input	Operator	Kernel	Out	Repeat	Stride
Initialisation: VGGFace2-pretrained checkpoint (applied independently to branches A and B)					
$128 \times 128 \times 3$	Stem Conv 1	3×3	32	1	2
$63 \times 63 \times 32$	Stem Conv 2	3×3	32	1	1
$61 \times 61 \times 32$	Stem Conv 3	3×3	64	1	1
$61 \times 61 \times 64$	Max Pool	3×3	64	1	2
$30 \times 30 \times 64$	Projection Conv	1×1	80	1	1
$30 \times 30 \times 80$	Spatial Conv	3×3	192	1	1
$28 \times 28 \times 192$	Stem Output Conv	3×3	256	1	2
$13 \times 13 \times 256$	Inception-ResNet-A	multi-branch residual block	256	5	1
Output feature map per branch: $13 \times 13 \times 256$					

Input: spatial resolution and channel depth of the feature map entering each layer ($H \times W \times C$); Operator: type of layer or block applied at that stage; Kernel: convolution kernel size or pooling window; Out: number of output channels produced by the layer; Repeat: number of times the block is applied consecutively at the same resolution; Stride: spatial downsampling factor applied by the operation (stride = 1 preserves resolution).

3.2. Dimensionality reduction: Wavelet pooling

Traditional pooling operations, such as max pooling and average pooling, typically perform irreversible spatial reduction by discarding high-frequency components of feature maps. This process often leads to the loss of fine-grained structural information, which is crucial for tasks requiring sensitivity to subtle spatial variations, and may also reduce the robustness of deep models to noise and appearance changes (Williams & Li, 2018). To address this limitation, the proposed architecture employs a wavelet pooling strategy that enables structured, frequency-aware downsampling while preserving essential information.

Unlike conventional pooling strategies that operate purely in the spatial domain, wavelet pooling explicitly decomposes feature maps into complementary sub-bands. This decomposition allows the separation of semantic content from fine structural details, rather than merging or discarding them during the downsampling process.

Given an input feature map F , wavelet pooling applies a discrete wavelet transform (DWT), producing four sub-bands arranged as:

$$\begin{bmatrix} LL & HL \\ LH & HH \end{bmatrix}$$

Each sub-band corresponds to a distinct frequency component of the input representation:

- **LL** (Approximation coefficients): Encodes low-frequency information, capturing the global structure, smooth intensity variations, and high-level semantic content of the feature map. It serves as the primary representation for subsequent processing layers.
- **HL** (Horizontal detail coefficients): Represents high-frequency variations along the horizontal direction, mainly capturing vertical edges and fine structural transitions.
- **LH** (Vertical detail coefficients): Encodes high-frequency variations along the vertical direction, highlighting horizontal edges and localised spatial changes.
- **HH** (Diagonal detail coefficients): Contains diagonal high-frequency information, describing fine textures and subtle spatial details, but it is more sensitive to noise.

In the proposed architecture, wavelet pooling is applied independently across multiple parallel feature streams, forming a multi-branch downsampling strategy. For each branch, only the LL sub-band is forwarded to subsequent layers for semantic feature learning and high-level reasoning. In contrast, the high-frequency sub-bands (HL, LH, HH) are preserved and transmitted through dedicated skip connections toward the fusion stage. These skip connections ensure that edge and texture-level information is not lost during the downsampling process, but is instead directly injected into later stages of the network. This design is intended to limit information loss and to preserve spatial fidelity throughout the architecture.

Overall, wavelet pooling introduces a principled trade-off between spatial reduction and information preservation, enabling efficient downsampling while retaining both semantic and structural richness. This makes it particularly suitable for deep learning tasks that require sensitivity to fine-grained spatial patterns and local structural variations.

3.3. Attention block

Attention mechanisms enable deep neural networks to model data-dependent interactions between features, thereby emphasising informative responses while suppressing less discriminative ones. Unlike purely convolutional operators, whose aggregation pattern is fixed once the kernel is learned, attention adapts the weighting of feature interactions to each input sample. This property is particularly valuable in facial emotion recognition, where subtle expression-specific evidence may be distributed across multiple facial regions and across multiple feature streams. In the proposed ARECA-Lite framework, attention is introduced after wavelet pooling, allowing low-frequency semantic tokens to be refined through cross-stream interactions before final fusion. This strategy improves the representational quality of the retained compact tokens while preserving the overall model's low-complexity objective.

3.3.1. Residual enhanced cross-attention

The core attention operator adopted in this work is the Residual Enhanced Cross-Attention (RECA) block, derived from the residual cross-stream fusion principle introduced in ArecaNet (Kim & Choi, 2025). Its objective is not merely to correlate two feature sets but to transfer complementary contextual information from one stream to the other while explicitly preserving the recipient stream's original representation. This distinction is important: in standard cross-attention, the attended output may overwrite part of the source information, whereas RECA retains the original recipient-stream representation through a residual addition, thereby reducing information loss during fusion.

Let $x_i \in \mathbb{R}^{N_i \times d}$ denote the token sequence of the feature stream to be refined and let $x_b \in \mathbb{R}^{N_b \times d}$ denote the contextual token sequence provided by Sub-Network 2 (branch B in Section 3.1.2), which acts as the shared complementary stream. In the present architecture, x_i corresponds to the low-frequency tokens originating from either the wavelet-processed identity-prior branch (Sub-Network 1) or the wavelet-processed Main Network, while x_b denotes the structural tokens provided by the second identity-prior branch (Sub-Network 2). The query, key, and value projections are defined as:

$$Q_i = x_i W^Q, \quad K_b = x_b W^K, \quad V_b = x_b W^V \quad (6)$$

Where W^Q , W^K , and W^V are learned linear projection matrices. The query originates from the branch-specific representation to be refined, whereas the key and value are produced by Sub-Network 2, acting as the shared contextual stream across both RECA blocks. The corresponding attention map is computed following the scaled dot-product formulation (Vaswani et al., 2017):

$$A_i = \text{softmax} \left(\frac{Q_i K_b^T}{\sqrt{d_k}} + M \right) \quad (7)$$

Where d_k is the key dimensionality, and M denotes the optional mask term shown in the generic attention formulation of Figure 3. The refined output is then obtained by weighting the value vectors of the contextual stream and adding the result back to the original query-stream representation:

$$z_i = A_i V_b + x_i \quad (8)$$

This residual formulation improves information propagation in two ways. First, the attention pathway injects complementary structural evidence from Sub-Network 2 into the refined branch representation via query-key interactions. Second, the residual summation preserves the initial branch-specific content encoded in x_i , thereby limiting the risk that salient facial structure cues are attenuated or discarded during fusion. From an optimisation perspective, the residual shortcut also facilitates gradient flow and generally promotes more stable convergence during training. For facial emotion recognition, these properties are especially desirable because weak expression cues often coexist with stable identity-related facial structure, and the model must enhance the former without destroying the latter.

Figure 2 illustrates the internal organisation of the RECA block. The query, key, and value inputs are first projected and then combined via matrix multiplication. The result is then scaled, optionally masked, and normalised by a softmax operation to produce attention weights. These weights are applied to the value pathway via a second matrix multiplication, yielding an attended feature tensor, which is then merged with the

original query-stream representation via a residual connection. The resulting design therefore preserves the recipient stream while allowing it to selectively absorb complementary structural information from Sub-Network 2. This makes RECA better suited than naive concatenation or plain cross-attention for multi-branch feature fusion, where retaining branch-specific information is as important as enabling cross-stream interaction.

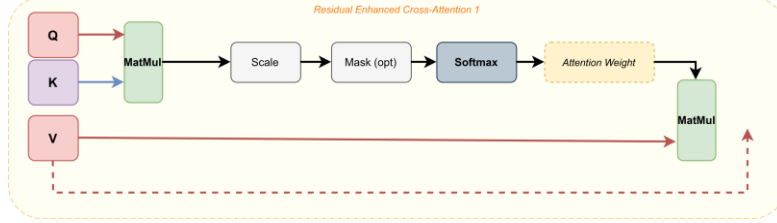


Fig. 2. Residual Enhanced Cross-Attention (RECA) block

3.3.2. Modified assembled residual enhanced cross-attention

To further enrich the fused representation, the proposed model adopts a Modified Assembled Residual Enhanced Cross-Attention (Modified A.R.E.C.A.) module in which two RECA blocks are used in parallel. This design choice follows the intuition that the two identity-prior branches do not encode perfectly redundant information: although they process the same preprocessed input image, they are optimised independently and can therefore specialise in complementary structural facial patterns. Employing two RECA units allows the model to refine each branch independently, leveraging the shared structural context provided by Sub-Network 2 while preserving the distinct discriminative characteristics learned by each branch.

Let x_A and x_M denote the low-frequency token sequences originating from Sub-Network 1 (branch A) and the Main Network, respectively, and let x_B denote the shared contextual token sequence extracted from Sub-Network 2 after wavelet projection. The two assembled attention operations are then written as:

$$z_1 = RECA(x_A, x_B), \quad z_2 = RECA(x_M, x_B) \quad (9)$$

Where z_1 and z_2 are the respective refined outputs of the first and second RECA blocks. Specifically, the first block refines the tokens from Sub-Network 1 (x_A) using the structural keys and values provided by Sub-Network 2 (x_B), while the second block refines the tokens from the Main Network (x_M) using the same shared keys and values from Sub-Network 2. The two outputs are not collapsed prematurely. Instead, they are preserved as separate refined descriptors and combined by final concatenation:

$$h_{att} = Concat(z_1, z_2) \quad (10)$$

This assembled design offers several advantages. First, it increases representational richness by allowing the two RECA units to learn complementary attention patterns rather than forcing all cross-stream interactions through a single operator. Second, it captures more complex dependencies between Sub-Network 2 and the two remaining branches (Sub-Network 1 and the Main Network), which is beneficial for facial expressions whose discriminative evidence is distributed across different local structures. Third, because attention is applied only to the compact, low-frequency tokens generated after wavelet pooling, the added representational capacity is achieved with limited computational overhead. Accordingly, the module improves feature integration without violating ARECA-Lite's overall low-latency design objective.

Figure 3 depicts the architecture of the Modified A.R.E.C.A. module. The upper RECA path corresponds to Cross-Attention 1 and produces the refined output z_1 from Sub-Network 1, while the lower path corresponds to Cross-Attention 2 and produces z_2 from the Main Network. In both paths, query, key, and value undergo the same sequence of operations, followed by residual preservation of the original branch-specific signal. Sub-Network 2 serves as the shared key-value provider for both attention blocks, enabling a unified structural reference to guide the refinement of x_A and x_M . The two refined outputs are then forwarded together to the downstream fusion head, where they are concatenated and further combined with the high-frequency skip connections. The practical effect of this structure is a more expressive yet information-preserving attention stage, which improves the quality of the fused representation and, in turn, strengthens the final emotion classification performance.

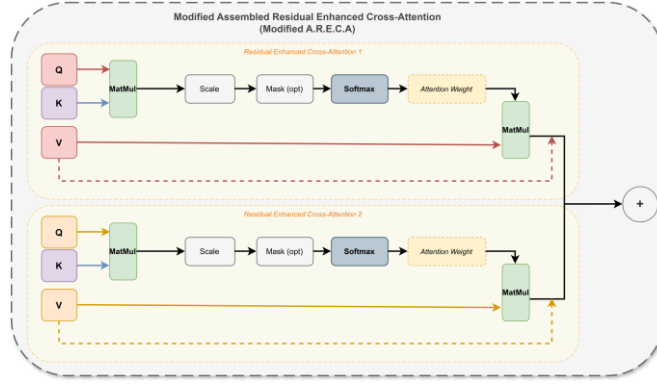


Fig. 3. Modified assembled residual enhanced cross-attention module composed of two parallel RECA blocks. The two refined outputs are preserved separately and concatenated downstream for multi-stream fusion

3.4 Classification head

The classification head of ARECA-Lite is the final decision module that maps a fused multi-stream representation to the seven target emotion categories. It takes as input a 1024-dimensional vector obtained by concatenating five heterogeneous feature streams, each encoding complementary aspects of facial expression.

The first stream (512-d) is produced by the modified A.R.E.C.A. module, which performs two parallel cross-attention operations. The following three streams encode high-frequency textural information via a discrete wavelet transform (DWT). Two of them (128-d each) are extracted by applying a wavelet projector to the output features of the two independent InceptionResNetV1 sub-networks (the $13 \times 13 \times 256$ tensor after the five Inception-ResNet-A blocks, Table 3). The third wavelet stream (128-d) is derived from the Stage 2 output of the truncated MobileNetV3-Small backbone, using the same procedure as for InceptionResNetV1. The fifth stream (128-d) serves as a skip connection from MobileNetV3 stage-1 (24 channels), obtained via adaptive average pooling followed by a linear projection with ReLU activation, thereby preserving low-level spatial cues that would otherwise be discarded by deeper processing stages. The five streams are concatenated along the feature dimension to yield the final fused representation of dimensionality:

$$512 + 128 + 128 + 128 + 128 = 1024 \quad (11)$$

This fused vector is subsequently processed by a lightweight two-layer multilayer perceptron (MLP) comprising five sequential operations: a layer normalisation (Ba et al., 2016), a linear bottleneck projection ($1024 \rightarrow 512$), a Gaussian Error Linear Unit (GELU; Hendrycks & Gimpel, 2016) activation, a dropout layer ($p = 0.15$), and a final linear projection to the seven emotion logits ($512 \rightarrow 7$), corresponding to the classes *angry*, *disgust*, *fear*, *happy*, *neutral*, *sad*, and *surprise*. No softmax activation is applied within the head; raw logits are passed directly to the loss function, which applies the appropriate transformations internally. The total number of trainable parameters in the head amounts to approximately 0.53 M.

The pre-normalisation design, whereby layer normalisation is applied prior to any linear projection, is motivated by the heterogeneous activation scales of the concatenated feature streams, as it stabilises gradient flow without relying on batch statistics. The GELU activation is selected for its well-documented advantage over ReLU in transformer-style networks. To mitigate overfitting, a mild dropout rate of 0.15 is applied during training; a higher rate was not adopted because the fused input already aggregates information from five independently regularised streams. All parameters are optimised using AdamW with differential learning rates (Section 3.6.3). All linear layer weights are initialised using a truncated normal distribution ($\sigma = 0.02$) with biases set to zero, consistent with Vision Transformer conventions (Dosovitskiy et al., 2021).

3.5. Dataset

3.5.1. FER2013

FER2013 (Goodfellow et al., 2013), introduced in the context of the Challenges in Representation Learning competition, has become the de facto benchmark for static facial expression recognition. The dataset contains 35,887 in-the-wild facial images, distributed as greyscale thumbnails of size 48×48 pixels, and partitioned

into 28,709 training samples and 7,178 test samples across seven mutually exclusive emotion categories: angry, disgust, fear, happy, neutral, sad, and surprise. The images were retrieved through the Google Image Search API using emotion-related keywords and subsequently filtered and relabelled by human annotators; the reported human-level accuracy on the dataset is only about $65 \pm 5\%$, indicating substantial label ambiguity and noise. Figure 4 shows representative samples for each emotion category.

For model development, the original training partition was further divided into training and validation subsets. Specifically, 20% of the original training data was allocated to validation, while the remaining approximately 80% was retained for model training. This resulted in 5,741 validation images and 22,968 training images, with the one-image difference from an exact 80/20 split arising from class-wise integer rounding.

The split was performed before any data augmentation. The class distribution was preserved as closely as possible between the training and validation subsets by allocating approximately 20% of samples from each emotion class to the validation set. The validation subset was used to monitor model performance and selection during training, while the official FER2013 test set, containing 7,178 images, remained completely unseen until the final evaluation.

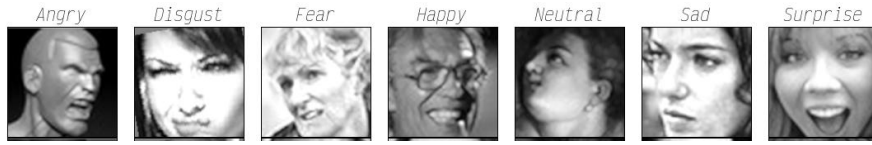


Fig. 4. Representative samples from the FER2013 dataset across the seven emotion categories

Table 4 presents the class distribution of the original training set and of the derived training and validation subsets. The majority class (happy) contains 7,215 samples, while the minority (disgust) contains only 436, yielding a maximum imbalance ratio of $16.5\times$. This skew, compounded by the visual similarity between emotion pairs sharing common Facial Action Coding System (FACS) action units (Ekman & Friesen, 1978), for example, fear and sadness, which both involve upper-face cues such as AU1 (inner brow raiser) and AU4 (brow lowerer), makes FER2013 substantially more challenging than its image count suggests.

Tab. 4. FER2013 class distribution and imbalance ratio of the original training set and of the derived training and validation subsets

Emotion Class	Original Training	Training Subset (80%)	Validation Subset (20%)	Imbalance Ratio
Happy	7,215	5,772	1,443	$16.5\times$
Neutral	4,965	3,972	993	$11.4\times$
Sad	4,830	3,864	966	$11.1\times$
Fear	4,097	3,278	819	$9.4\times$
Angry	3,995	3,196	799	$9.2\times$
Surprise	3,171	2,537	634	$7.3\times$
Disgust	436	349	87	$1.0\times$ (baseline)

3.5.2. Preprocessing pipeline

All images are preprocessed offline and persisted as lossless PNG files at 128×128 pixels prior to training. This offline strategy ensures that all enhancement operations are executed once at dataset preparation time, adding no latency during training and, for the pre-processed benchmark images, during evaluation. The pipeline is intentionally deterministic and anatomically guided: rather than learning a preprocessing module jointly with the classifier, the proposed approach performs a sequence of fixed transformations designed to improve the visibility of diagnostically relevant facial micro-structures before feature extraction. Figure 5 illustrates the complete preprocessing pipeline.

- Resolution and colour space: input images are converted from single-channel greyscale to three-channel RGB by channel replication for compatibility with ImageNet- and VGGFace2-pretrained weights. They are then upsampled from 48×48 to 128×128 using Lanczos interpolation, which better preserves subtle edge detail and reduces ringing artefacts relative to simpler interpolation schemes.
- CLAHE contrast enhancement: Contrast Limited Adaptive Histogram Equalisation (CLAHE; Zuiderveld, 1994) is applied only to fear and sad samples on the L^* channel of the CIE $L^*a^*b^*$ space

(clip limit 2.0; tile grid 8×8). This targets the low-contrast AUs most relevant to those classes, namely AU5 and AU41 for fear, and AU15 and AU17 for sadness.

- Anatomical-prior soft attention mask: a fixed soft mask is constructed over four FACS-relevant regions: brow/forehead (AU1, AU2, AU4), periocular (AU5, AU41), mouth/lips (AU12, AU15, AU17), and nasolabial folds (AU9). The regions are smoothed with a Gaussian kernel ($\sigma = 18$) and used to modulate image intensity, with $\alpha_{\text{attn}} = 0.25$ for standard classes and $\alpha_{\text{attn}} = 0.125$ for CLAHE-processed classes, thereby preventing over-enhancement and avoiding reliance on facial landmark detection.
- Sobel edge blend: a normalised Sobel gradient magnitude map is added with a blending weight of 0.15 to emphasise muscle-tension creases and other expression-relevant structural edges without suppressing the original appearance information.

The class-dependent preprocessing operations were applied exclusively to the training data; validation and test samples were processed without class-dependent CLAHE or class-specific attention-mask coefficients.



Fig. 5. FACS-guided preprocessing pipeline applied to FER2013 images

3.5.3. Data augmentation

To counteract overfitting and address the severe class imbalance in the training partition, a targeted augmentation strategy is applied exclusively to training samples. For the disgust class, the most severely under-represented with only 436 samples in the original training set, each original image is augmented offline with six deterministic variants: (i) horizontal flip; (ii) $+15^\circ$ rotation; (iii) -15° rotation; (iv) $\times 1.3$ brightness scaling; (v) $\times 0.7$ brightness scaling; and (vi) combined horizontal flip with $+15^\circ$ rotation. This procedure expands the disgust class from 349 to $349 \times 7 = 2443$ images, reducing the maximum imbalance ratio from $16.5\times$ to approximately $2.4\times$. During training, the stochastic CutMix (Yun et al., 2019) augmentation with $\alpha = 1.0$ is applied on the fly. CutMix is preferred over Mixup (Zhang et al., 2018) because it preserves spatially contiguous facial regions rather than superimposing blended textures, which would distort the muscle-tension cues the preprocessing pipeline was designed to enhance.

3.6. Experimental configuration

All experiments are conducted on a workstation equipped with an Intel Core i5-12th Generation H-series processor, 16 GB of DDR5 RAM, and an NVIDIA GeForce RTX 4050 GPU. Training employs Automatic Mixed Precision (AMP), which reduces GPU memory consumption and accelerates throughput by performing forward and backward passes in FP16 while maintaining FP32 master weights for numerical stability.

The implementation relies on Python 3.10, PyTorch 2.1.0, torchvision 0.16.0, facenet-pytorch 1.0.1 (for InceptionResNetV1 and VGGFace2 weights), OpenCV 4.8.0 (for CLAHE, Sobel, and resize operations), NumPy 1.26.0, scikit-learn 1.3.0, albumentations 1.3.1, and matplotlib 3.8.0.

AdamW (Loshchilov & Hutter, 2019) is used as the optimiser with decoupled weight decay $\lambda = 10^{-4}$ and $\epsilon = 10^{-8}$. Differential learning rates are applied: $\eta_{\text{backbone}} = 5 \times 10^{-5}$ for pretrained layers and $\eta_{\text{head}} = 10^{-3}$ for newly initialised components. Training uses 5 warmup epochs, starting from 10% of the target learning rate, followed by cosine annealing over 100 total epochs. Gradient clipping is applied with a maximum L_2 norm of 1.0, and the batch size is 32.

The loss function is a composite Dice-BCE objective with weights $\alpha = 0.7$ (Dice) and $\beta = 0.3$ (BCE):

$$L = 0.7 \times L_{\text{Dice}} + 0.3 \times L_{\text{BCE}} \quad (12)$$

Dice loss (Milletari et al., 2016) encourages class-overlap consistency under imbalance, while BCE provides a complementary sample-wise probabilistic objective. This weighting $(\alpha, \beta) = (0.7, 0.3)$ gave the best results among the tested values in the loss-weight ablation (Section 4.5.1).

Tab. 5. Summary of training hyperparameters

Hyperparameter	Value	Hyperparameter	Value
Optimiser	AdamW	Loss function	$0.7 \times L_{Dice} + 0.3 \times L_{BCE}$
η (backbone)	5×10^{-5}	Dice smoothing	1.0
η (head)	1×10^{-3}	Dropout rate	0.15
Weight decay (λ)	1×10^{-4}	RECA blocks	2
Batch size	32	Adam ϵ	1×10^{-8}
Total epochs	100	Warmup start factor	10% of the target learning rate
Warmup epochs	5	CutMix α	1.0
LR scheduler	Linear warmup + cosine annealing	Input resolution	$128 \times 128 \times 3$
Gradient clipping (L2 norm)	1.0	Test-time augmentation	Horizontal flip (2 passes, softmax averaged)

4. RESULTS AND DISCUSSION

This section presents a comprehensive evaluation of ARECA-Lite. We compare the proposed model against the original ArecaNet baseline across training performance, test-set generalisation, ablation of the preprocessing pipeline, and comparison with representative recent methods.

4.1. Evaluation metrics

Following common practice in facial emotion recognition, overall accuracy and macro-averaged F1 are treated as the principal summary metrics. Accuracy captures global classification performance, whereas macro F1 weights all classes equally and is therefore less biased toward majority classes, making it especially important for the severely imbalanced FER2013 dataset. Overall accuracy is defined as the proportion of correctly classified samples over the total evaluation set. Macro F1 is the unweighted mean of per-class F1 scores, where each $F1 = 2 \times Precision \times Recall / (Precision + Recall)$.

Model efficiency is additionally evaluated using parameter count (M), storage size (MB), inference latency (ms), and GFLOPs (10^9 floating-point operations per forward pass). During test evaluation, test-time augmentation (TTA) is applied; each test image and its horizontal flip are passed through the model independently, and the two softmax outputs are averaged to produce the final class prediction.

4.2. Training and validation phase

Table 6 reports the training and validation performance for ArecaNet and ARECA-Lite. ARECA-Lite achieves training accuracy similar to ArecaNet (87.85% vs 87.58%), but its training macro F1 is substantially higher (87.82% vs 75.47%), indicating a much more balanced performance across classes. ARECA-Lite exhibits strong agreement between its two principal metrics (training accuracy 87.85% and macro F1 87.82%), indicating genuinely balanced convergence without over-optimisation toward any particular class. The validation gap for ArecaNet (87.58% to 75.88% in accuracy) is substantially wider than for ARECA-Lite (87.85% to 82.14%), suggesting stronger overfitting in the original architecture. Figure 6 presents the corresponding training and validation learning curves.

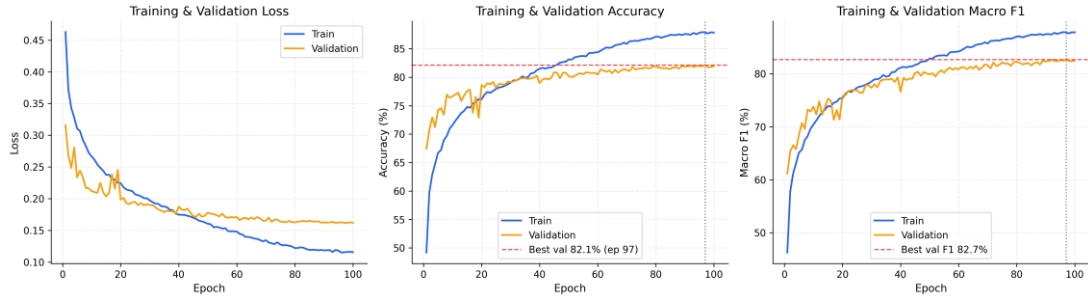


Fig. 6. Training and validation curves (loss, accuracy, macro F1) for ARECA-Lite over 100 epochs\

Tab. 6. Training and validation performance: ArecaNet vs. ARECA-Lite

Configuration	Train Acc (%)	Train F1 (%)	Train Loss	Val Acc (%)	Val F1 (%)	Val Loss
ArecaNet	87.58	75.47	0.9513	75.88	73.13	0.9419
ARECA-Lite (Ours)	87.85	87.82	0.1155	82.14	82.69	0.1618

4.3. Computational complexity

Table 7 compares the computational complexity and inference efficiency of the two models. ARECA-Lite reduces the parameter count by 81.2% relative to ArecaNet (4.69 M vs. 24.95 M) and reduces the storage size from 95.18 MB to 17.92 MB. In terms of single-pass inference latency, ARECA-Lite achieves 5.44 ms on GPU and 17.51 ms on CPU, compared with 21.83 ms and 87.71 ms for ArecaNet, corresponding to reductions of 75.1% and 80.0%, respectively. The computational complexity is also reduced from 6.57 GFLOPs to 1.64 GFLOPs, representing approximately a 75.1% reduction. These reductions demonstrate the computational efficiency of ARECA-Lite while maintaining competitive recognition performance.

The inference latency reported in Table 7 corresponds to a single forward pass without TTA. Since TTA requires two independent forward passes, the effective inference time during TTA-based evaluation is approximately twice the reported single-pass latency, excluding minor additional TTA processing overhead. We report single-pass latency in Table 7 because this metric is intended to characterise each model's intrinsic computational cost independently of the evaluation strategy. TTA is employed only as a test-time evaluation procedure to improve prediction robustness and is not considered part of the model's intrinsic inference cost.

Tab. 7. Computational complexity: ArecaNet vs. ARECA-Lite

Emotion Class	Original Training	Training Subset (80%)	Validation Subset (20%)		Imbalance Ratio
			Emotion Class	Original Training	
ArecaNet	24.95	95.18	21.83	87.71	6.57
ARECA-Lite (Ours)	4.69	17.92	5.44	17.51	1.64

4.4. Test phase results

ArecaNet's reproduced baseline of 75.31% differs from the 97.00% reported in the original work. This discrepancy may be due to differences in evaluation protocol or train/test partitioning, as the original paper does not fully specify its data split. All comparisons in this work use the reproduced baseline for experimental consistency.

Table 8 reports the test performance for both models on the FER2013 test set. ARECA-Lite achieves 82.41% accuracy and 82.50% macro F1, corresponding to gains of 7.10 and 9.43 percentage points over ArecaNet, respectively. ArecaNet's training-to-test degradation (87.58% to 75.31%, a reduction of 12.27 points) is more than twice that of ARECA-Lite (87.85% to 82.41%, a reduction of only 5.44 points), indicating that ARECA-Lite generalises more effectively and is less affected by structural overfitting. ARECA-Lite also achieves a lower test loss (0.1506); however, loss values are only directly comparable when both models are trained with the same objective.

Tab. 8. FER2013 test performance: ArecaNet vs. ARECA-Lite

Configuration	Avg Accuracy (%)	Avg Macro F1 (%)	Loss
ArecaNet	75.31	73.07	0.7174
ARECA-Lite (Ours)	82.41	82.50	0.1506

The class-wise analysis in Table 9 reveals that ARECA-Lite outperforms ArecaNet on all seven emotion categories. The most pronounced gains occur for the difficult and under-represented categories: disgust improves by 24.28 percentage points (64.90% to 89.18%), anger by 16.63 points (57.40% to 74.03%), and fear by 15.92 points (50.80% to 66.72%). For the more frequent categories, the gains are smaller but consistent: happy improves from 84.60% to 90.21%, neutral from 83.20% to 84.66%, sad from 83.20% to 85.97%, and surprise from 84.10% to 85.55%. This balanced improvement across all classes indicates that ARECA-Lite enhances minority-class recognition without sacrificing performance on majority classes. Figure 7 shows the normalised confusion matrix on the test set, and Figure 8 provides a per-class accuracy comparison between the two models.

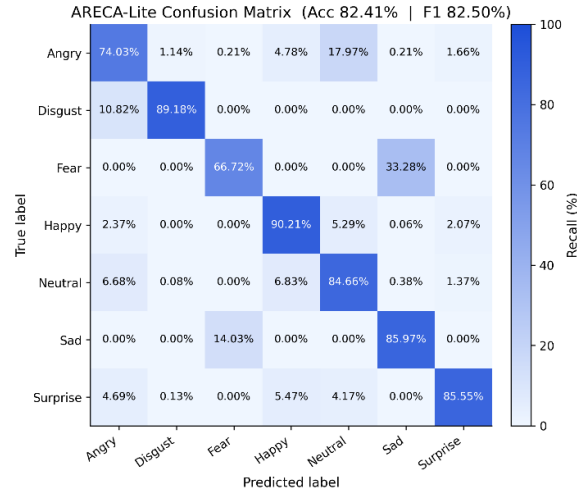


Fig. 7. Normalised confusion matrix for ARECA-Lite on the FER2013 test set

Tab. 9. FER2013 test class-wise accuracy (%) for all seven emotion categories

Configuration	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
ArecaNet	57.40	64.90	50.80	84.60	83.20	83.20	84.10
ARECA-Lite (Ours)	74.03	89.18	66.72	90.21	84.66	85.97	85.55

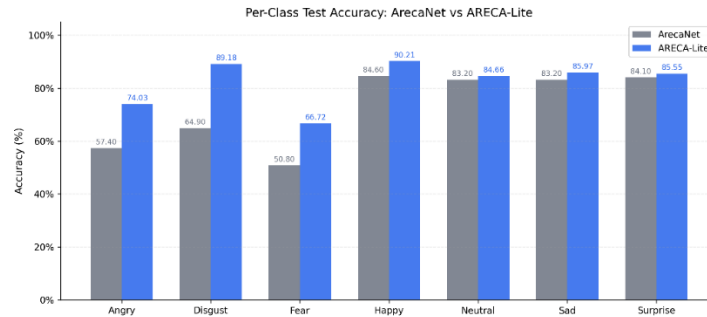


Fig. 8. Per-class test accuracy comparison between ArecaNet and ARECA-Lite

4.5. Ablation study

To better understand the design choices underlying ARECA-Lite, a series of ablation experiments was conducted. These experiments investigate the contributions of the main components of the proposed training and preprocessing strategy, including the loss function for addressing the marked class imbalance, the augmentation pipeline, the FACS-guided preprocessing strategy, and the targeted oversampling of the disgust class. In all experiments except the raw-data baseline, only one factor is modified at a time while the remaining architectural components and training parameters are kept fixed, allowing the individual contribution of each component to be evaluated more clearly.

4.5.1. Loss function ablation

This subsection investigates how the relative weighting of the Dice and BCE components affects final performance. To isolate this factor, all other architectural choices and training conditions are kept fixed while only the pair (α, β) in the composite loss is varied.

Tab. 10. Test-set results of the loss-weight ablation (rows ordered by increasing Dice weight α)

Model	α	β	Avg Accuracy (%)	Avg Macro F1 (%)
ARECA-Lite	0	1	77.56	76.87
	0.3	0.7	77.89	77.10
	0.5	0.5	79.36	79.99
	0.7	0.3	82.41	82.50
	1	0	78.03	77.76

Table 10 examines the influence of the weighting scheme of the composite Dice-BCE loss on the final performance of ARECA-Lite. Five weighting configurations are evaluated under identical training conditions. The setting that uses BCE alone, $(\alpha, \beta) = (0, 1)$, produces the weakest results, with 77.56% accuracy and 76.87% macro F1. This limitation is consistent with the fact that BCE, when used alone, remains more sensitive to the dominant classes and is therefore less suitable for the severe imbalance present in FER2013.

Increasing the Dice weight generally improves the predictive metrics, although the trend is not monotonic. The configuration (0.3, 0.7) yields 77.89% accuracy and 77.10% macro F1, while Dice alone (1, 0) reaches 78.03% accuracy and 77.76% macro F1. The balanced combination (0.5, 0.5) performs even better, achieving 79.36% accuracy and 79.99% macro F1, indicating a complementary effect between the two terms. Overall, the BCE-dominant settings remain weaker than the more Dice-oriented alternatives, reintroducing part of the majority-class bias.

The best performance is obtained for $(\alpha, \beta) = (0.7, 0.3)$, which yields 82.41% accuracy and 82.50% macro F1. Relative to BCE alone, this configuration improves accuracy and macro F1 by 4.85 and 5.63 points, respectively. These results confirm that Dice loss is the dominant term for coping with FER2013 class imbalance, while the partial BCE contribution still provides a complementary sample-wise objective. Accordingly, this weighting is retained as the best-performing setting among those tested for the main ARECA-Lite experiments.

4.5.2. Impact of preprocessing pipeline

To evaluate the contributions of the proposed preprocessing and class-balancing components, an ablation study was conducted across four ARECA-Lite configurations. The row labels indicate the component removed from the complete pipeline, while the model architecture and training settings were kept unchanged. The results are presented in Table 11.

Tab. 11. Ablation study on preprocessing and class-balancing strategies

Method	Avg Accuracy (%)	Avg Macro F1 (%)	Avg Loss
Raw FER2013 (all proposed preprocessing components removed)	67.53	66.86	0.1961
Full pipeline without FACS-guided preprocessing	80.92	77.88	0.1689
Full pipeline without targeted oversampling of the disgust class	80.66	73.60	0.1876
Full preprocessing pipeline (original)	82.41	82.50	0.1506

The raw FER2013 configuration achieves the lowest performance, with 67.53% accuracy and 66.86% macro F1. In contrast, the full pipeline reaches 82.41% accuracy and 82.50% macro F1, with the lowest loss of 0.1506. This confirms the overall benefit of the proposed preprocessing and balancing strategy.

Removing FACS-guided preprocessing reduces accuracy from 82.41% to 80.92% and macro F1 from 82.50% to 77.88%. This indicates that FACS-guided processing contributes to more consistent recognition across emotion classes.

Removing targeted oversampling of the disgust class yields slightly lower accuracy (80.66%), but macro F1 drops substantially to 73.60%. This larger decrease suggests that targeted oversampling is particularly important for maintaining balanced performance across classes.

Overall, the full preprocessing pipeline provides the best results for all three metrics, confirming that FACS-guided preprocessing and targeted class balancing make complementary contributions to ARECA-Lite performance.

4.5.3. Ablation study of the ARECA-lite architectural components

This subsection presents a detailed analysis of the test behaviour of ARECA-Lite and three competing variants evaluated on the FER2013 dataset. The objective is to quantify the contribution of the main architectural modifications introduced in ARECA-Lite by selectively replacing individual components with their counterparts from the original ArecaNet architecture. In each experiment, only the targeted component is replaced, while the remaining components, the preprocessing pipeline, and training settings are kept unchanged to ensure a fair comparison.

The three compared configurations are defined as follows:

- Configuration 1: ARECA-Lite with the modified ARECA block replaced by the original ArecaNet A.R.E.C.A. module.
- Configuration 2: ARECA-Lite with the feature extraction stage replaced by the original ArecaNet feature extraction design.
- Configuration 3: ARECA-Lite with the dimensionality reduction stage replaced by the original ArecaNet reduction module.

Table 12 presents the computational complexity of the four configurations, while Table 13 reports their corresponding recognition performance on the FER2013 test set.

Tab. 12. Computational complexity of ARECA-Lite and its architectural ablation variants

Model	Parameters (M)	Size (MB)	Latency (ms)	GFLOPs
Configuration 1	22.41	85.49	18.96	5.92
Configuration 2	5.12	19.53	5.98	1.79
Configuration 3	6.08	23.19	6.74	2.14
ARECA-Lite (Ours)	4.69	17.92	5.44	1.64

Tab. 13. Recognition performance of ARECA-Lite and its architectural ablation variants on the FER2013 test set

Configuration	Avg Accuracy (%)	Avg Macro F1 (%)	Avg Loss
Configuration 1	79.72 $\pm$ 0.43	79.51 $\pm$ 0.41	0.1820 $\pm$ 0.0054
Configuration 2	76.88 $\pm$ 0.46	76.65 $\pm$ 0.44	0.2040 $\pm$ 0.0058
Configuration 3	78.88 $\pm$ 0.44	78.68 $\pm$ 0.42	0.1910 $\pm$ 0.0055
ARECA-Lite (Ours)	82.41 $\pm$ 0.41	82.50 $\pm$ 0.35	0.1506 $\pm$ 0.0057

Replacing the modified A.R.E.C.A. module with the original ArecaNet module, as in Configuration 1, reduces the accuracy from 82.41% to 79.72% and the macro F1 from 82.50% to 79.51%. This represents decreases of 2.69 and 2.99 percentage points, respectively. At the same time, the computational requirements increase substantially, reaching 22.41 million parameters, 85.49 MB, 18.96 ms latency, and 5.92 GFLOPs. This result highlights the benefit of the modified block in ARECA-Lite, which achieves a better trade-off between representational capacity and computational efficiency.

Configuration 2, in which the feature extraction stage is replaced with the original ArecaNet feature extraction design, results in the largest performance degradation. Accuracy decreases to 76.88%, while macro F1 decreases to 76.65%, corresponding to reductions of 5.53 and 5.85 percentage points compared with ARECA-Lite. The average loss also increases to 0.2040 ± 0.0058 . Although this configuration remains relatively lightweight, with 5.12 million parameters and 1.79 GFLOPs, its lower recognition performance suggests that the feature-extraction design adopted in ARECA-Lite plays a particularly important role in learning discriminative facial representations.

For Configuration 3, replacing the proposed dimensionality-reduction stage with the original ArecaNet reduction module results in a moderate decrease in performance. Accuracy falls to 78.88%, while macro F1 decreases to 78.68%, with the average loss increasing to 0.1910 ± 0.0055 . The computational requirements also increase to 6.08 million parameters, 23.19 MB, 6.74 ms latency, and 2.14 GFLOPs. Compared with Configuration 2, the smaller performance degradation suggests that the dimensionality-reduction module contributes to overall performance, but its effect is less pronounced than that of the feature-extraction stage.

Overall, the ablation results confirm that the architectural modifications introduced in ARECA-Lite are complementary. Each component contributes to the final model, while the combination of the modified A.R.E.C.A. module, lightweight feature extraction stage, and efficient dimensionality reduction module provides the most favourable balance between recognition accuracy, macro F1, model size, latency, and computational complexity.

4.6. Comparative analysis

To position ARECA-Lite within the broader FER2013 literature, Table 14 compares the proposed model with representative methods spanning the accuracy-efficiency spectrum. The comparison with existing lightweight facial emotion recognition (FER) models should be interpreted in terms of both recognition performance and deployment efficiency, as accuracy alone is insufficient to assess a model's suitability for real-time e-learning applications. The comparison focuses on three deployment-relevant criteria: recognition accuracy, parameter count, and inference latency whenever these values are reported. This comparison also includes newer lightweight FER approaches to provide a more meaningful assessment of ARECA-Lite against current efficiency-oriented alternatives.

Table 14 reveals a consistent trade-off between recognition performance and computational efficiency: models achieving high recognition accuracy generally require substantially larger parameter budgets, whereas highly lightweight approaches may sacrifice recognition performance to reduce computational cost. Within the reported comparison, ARECA-Lite simultaneously achieves an accuracy above 80%, fewer than 5 million parameters, and an inference latency below 10 ms. Among the edge-deployable methods, ARECA-Lite outperforms the nearest accuracy-oriented lightweight competitor, Improved Mini-Xception (79.96%), by 2.45 percentage points, while its measured latency (5.44 ms) is far below the 104.0 ms reported for that method, although latency measurements across different hardware are not directly comparable. Moreover, ARECA-Lite achieves nearly identical accuracy and macro F1 scores (82.41% and 82.50%, respectively), indicating that its overall accuracy is not primarily driven by majority-class predictions. Figure 9 provides a visual comparison of the recognition and efficiency characteristics, while Figure 10 illustrates the relationship between recognition accuracy and model size across the evaluated methods.

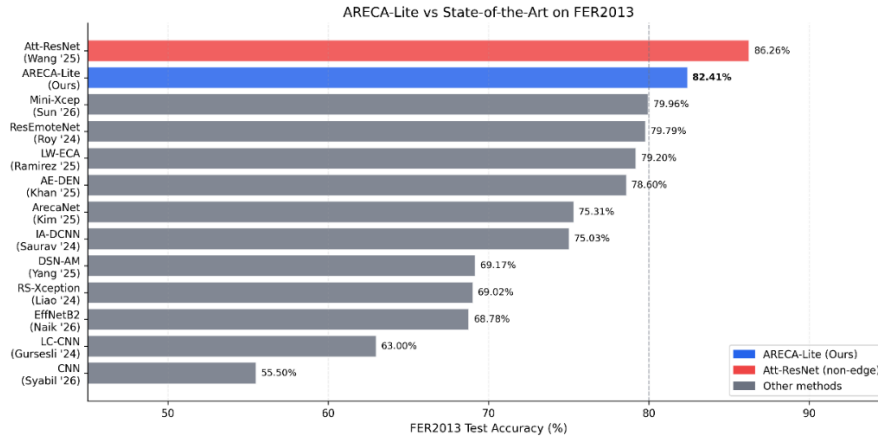


Fig. 9. ARECA-Lite compared against representative recent methods on FER2013 by test accuracy

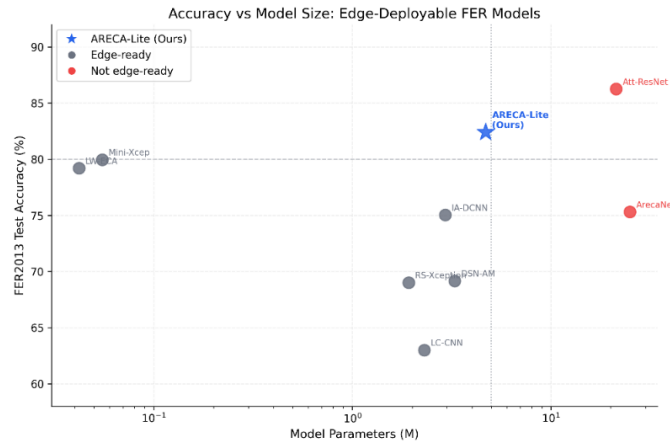


Fig. 10. Recognition accuracy vs. model size for edge-deployable FER models

Tab. 14. Comparison of ARECA-Lite with representative recent methods on FER2013 (n/r: not reported)

Author(s)	Model	Dataset	Acc (%)	Params (M)	Lat. (ms)	Edge-Ready?
Roy et al.	ResEmoteNet	FER2013	79.79	n/r	n/r	No
Gursesli et al.	LC-CNN	FER2013	63.00	2.30	50.0	Yes
Liao et al.	RS-Xception	FER2013	69.02	1.92	15.9	Yes
Saurav et al.	IA-DCNN	FER2013	75.03	2.93	2.96	Yes
Yang et al.	DSN-AM	FER2013	69.17	3.27	23.0	Yes
Ramirez-Quintana et al.	LW-ECA	FER2013	79.20	0.042	1.88	Yes
Khan et al.	AE-DEN	FER2013	78.60	n/r	n/r	No
Wang et al.	Att-ResNet	FER2013	86.26	21.30	n/r	No
Kim & Choi	Retrained ArecaNet	FER2013	75.31	24.95	n/r	NO
Naik et al.	EffNetB2	FER2013	68.78	9.2	n/r	NO
Sun et al.	Mini-XCep (Impr.)	FER2013	79.96	0.055	104.0	Yes
Syabil et al.	CNN	FER2013	55.50	n/r	n/r	No
Ours	ARECA-Lite	FER2013	82.41	4.69	5.44	Yes

Overall, ARECA-Lite provides a favourable balance between recognition performance and computational cost compared with the lightweight FER models considered in Table 14. It combines competitive recognition

performance with a compact parameter budget and low single-pass GPU inference latency, making it particularly suitable for resource-constrained and real-time e-learning environments. However, this comparison should be interpreted as contextual rather than as a strict head-to-head benchmark. The compared studies may differ in data partitions, preprocessing pipelines, data augmentation strategies, training procedures, and hardware platforms, and some studies do not report all efficiency metrics. Consequently, the results do not establish absolute superiority over all existing lightweight FER models. A definitive ranking would require a controlled evaluation in which all competing models are trained and evaluated using identical data partitions, preprocessing procedures, evaluation protocols, and hardware conditions. Nevertheless, the inclusion of recent lightweight FER methods and the joint analysis of accuracy, model size, and inference latency provide a more informative positioning of ARECA-Lite within the current efficiency-oriented FER landscape.

4.7. Limitations and future directions

Although ARECA-Lite demonstrates strong recognition accuracy and computational efficiency, several methodological constraints limit the generalisability of the reported findings and motivate future research directions.

The model is evaluated only on FER2013, which contains substantial label noise. Broader evaluation on datasets such as AffectNet (Mollahosseini et al., 2019) and RAF-DB (Li et al., 2017) is therefore necessary to assess robustness under more diverse imaging conditions. In addition, the current model processes individual frames independently, without explicit temporal modelling of expression dynamics. In practical e-learning scenarios, temporal context may help disambiguate visually similar emotions such as fear and surprise, making lightweight temporal extensions a promising direction for future work.

The measured single-pass latency of 5.44 ms on the NVIDIA GeForce RTX 4050 GPU and 17.51 ms on the CPU is encouraging, but the model has not yet been validated on ARM-based processors and integrated GPUs representative of resource-constrained e-learning environments. Future work should therefore include on-device profiling together with INT8 post-training quantisation and structured pruning. Moreover, because the disgust class relies on offline oversampling of a small number of original images, residual distributional sensitivity in this minority class cannot be ruled out, suggesting that domain adaptation or generative augmentation may further strengthen generalisation. Finally, the anatomical-prior soft attention mask relies on fixed facial regions rather than run-time landmark localisation, which may reduce robustness under large pose variation or occlusion. Integrating a lightweight landmark detector to adapt the mask dynamically, therefore, constitutes another promising extension.

5. CONCLUSIONS

This paper presented ARECA-Lite, a lightweight modified ArecaNet architecture designed for real-time facial emotion recognition in e-learning environments where resource constraints and affective responsiveness must be satisfied simultaneously. The work addresses the affective-feedback deficit introduced by the large-scale virtualisation of instruction, through which the continuous emotional signal needed for adaptive teaching is effectively lost.

ARECA-Lite integrates five complementary contributions: VGGFace2-pretrained dual InceptionResNetV1 sub-branches that transfer structural facial priors; a truncated MobileNetV3-Small backbone that provides efficient multi-scale feature extraction; a wavelet-pooling module that preserves high-frequency information through dedicated skip connections to the fusion head; a modified A.R.E.C.A. cross-attention module with two parallel RECA blocks that enable complementary feature refinement at reduced complexity; and a FACS-guided offline preprocessing pipeline that amplifies AU evidence without increasing model inference-time cost. Combined with targeted disgust oversampling and the Dice-BCE composite loss, this design substantially improves minority-class recognition.

On the FER2013 test set, ARECA-Lite achieves 82.41% accuracy and 82.50% macro F1, surpassing ArecaNet by 7.10 and 9.43 percentage points, respectively, while reducing parameter count by 81.2%, GPU latency by 75.1%, and GFLOPs from 6.57 to 1.64. The preprocessing and augmentation pipeline contributes an additional 14.88 percentage points in accuracy relative to raw-data training. These results provide empirical evidence that the usual trade-off between predictive performance and deployment efficiency in facial emotion recognition can be substantially mitigated through coordinated architectural, preprocessing, and loss-design

choices. ARECA-Lite is consequently well-suited to low-latency affective-computing applications, particularly e-learning systems deployed on commodity hardware.

Conflicts of Interest

The authors declare no conflict of interest.

REFERENCES

- Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer normalization*. *ArXiv, abs/1607.06450*. <https://doi.org/10.48550/arXiv.1607.06450>
- Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 67–74). IEEE. <https://doi.org/10.1109/FG.2018.00020>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR 2021)*.
- Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press.
- Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shave-Taylor, J., Milakov, M., Park, J., ... Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. In M. Lee, A. Hirose, Z.-G. Hou, & R. M. Kil (Eds.), *Neural information processing: ICONIP 2013* (Lecture Notes in Computer Science, Vol. 8228, pp. 117–124). Springer. https://doi.org/10.1007/978-3-642-42051-1_16
- Gursesli, M. C., Lombardi, S., Duradoni, M., Bocchi, L., Guazzini, A., & Lanatà, A. (2024). Facial emotion recognition (FER) through custom lightweight CNN model: Performance evaluation in public datasets. *IEEE Access*, 12, 45543–45559. <https://doi.org/10.1109/ACCESS.2024.3380847>
- Hendrycks, D., & Gimpel, K. (2016). *Gaussian error linear units (GELUs)*. *ArXiv, abs/1606.08415*. <https://doi.org/10.48550/arXiv.1606.08415>
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). *MobileNets: Efficient convolutional neural networks for mobile vision applications*. *ArXiv, abs/1704.04861*. <https://doi.org/10.48550/arXiv.1704.04861>
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., & Adam, H. (2019). Searching for MobileNetV3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1314–1324). IEEE. <https://doi.org/10.1109/ICCV.2019.00140>
- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7132–7141). IEEE. <https://doi.org/10.1109/CVPR.2018.00745>
- Khan, T., Yasir, M., & Choi, C. (2025). Attention-enhanced optimized deep ensemble network for effective facial emotion recognition. *Alexandria Engineering Journal*, 119, 111–123. <https://doi.org/10.1016/j.aej.2025.01.078>
- Kim, J., & Choi, G. (2025). ArecaNet: Robust facial emotion recognition via assembled residual enhanced cross-attention networks for emotion-aware human–computer interaction. *Sensors*, 25(23), Article 7375. <https://doi.org/10.3390/s25237375>
- Kinane Daouadi, A., & Bendella, F. (2024). Improving e-learning by facial expression analysis. *Applied Computer Science*, 20(2), 126–137. <https://doi.org/10.35784/acs-2024-20>
- Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2852–2861). IEEE. <https://doi.org/10.1109/CVPR.2017.277>
- Liao, X., Zhang, Y., & Liu, Z. (2024). RS-Xception: A lightweight facial emotion recognition network using depthwise separable convolutions. *Journal of Real-Time Image Processing*, 21(3), Article 89. <https://doi.org/10.1007/s11554-024-01452-1>
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR 2019)*. OpenReview.
- Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)* (pp. 565–571). IEEE. <https://doi.org/10.1109/3DV.2016.79>
- Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1), 18–31. <https://doi.org/10.1109/TAFFC.2017.2740923>
- Naik, S., Bagayatkar, S., & Singh, P. (2026). *Facial emotion recognition on FER-2013 using an EfficientNetB2-based approach*. *ArXiv, abs/2601.18228*. <https://doi.org/10.48550/arXiv.2601.18228>
- Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *ArXiv, abs/arXiv.1710.05941*. <https://doi.org/10.48550/arXiv.1710.05941>
- Ramirez-Quintana, J. A., Muñoz-Pacheco, J. J., Ramirez-Alonso, G., Medrano-Hermosillo, J. A., & Corral-Saenz, A. D. (2025). Lightweight convolutional neural network with efficient channel attention mechanism for real-time facial emotion recognition in embedded systems. *Sensors*, 25(23), Article 7264. <https://doi.org/10.3390/s25237264>
- Roy, A. K., Kathania, H. K., Sharma, A., Dey, A., & Ansari, M. S. A. (2024). *ResEmoteNet: Bridging accuracy and loss reduction in facial emotion recognition*. *ArXiv, abs/2409.10545*. <https://doi.org/10.48550/arXiv.2409.10545>

- Saurav, S., Saini, R., & Singh, S. (2025). An integrated attention-guided deep convolutional neural network for facial expression recognition in the wild. *Multimedia Tools and Applications*, 84(12), 10027–10069. <https://doi.org/10.1007/s11042-024-19012-2>
- Sun, X., Yang, J., & Zhou, Y. (2026). Research on a lightweight real-time facial expression recognition system based on an improved Mini-Xception algorithm. *Information*, 17(1), Article 111. <https://doi.org/10.3390/info17010111>
- Syabil, M., Rahman, A., & Ahmad, F. (2024). Baseline convolutional neural networks for static emotion recognition on FER2013. *Journal of Artificial Intelligence and Computer Science*, 12(2), 45–58.
- Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1, pp. 4278–4284). AAAI Press. <https://doi.org/10.1609/aaai.v31i1.11231>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 5998–6008). Curran Associates, Inc.
- Wang, H., Zhang, Y., Liu, X., Chen, M., & Zhao, L. (2025). Enhancing real-time facial emotion recognition in classrooms via attention-ResNet optimization. *ArXiv*, abs/2512.08970. <https://doi.org/10.48550/arXiv.2512.08970>
- Williams, T., & Li, R. (2018). Wavelet pooling for convolutional neural networks. In *6th International Conference on Learning Representations (ICLR 2018)*. OpenReview.
- Yang, Q., He, Y., Chen, H., Wu, Y., & Rao, Z. (2025). A novel lightweight facial expression recognition network based on deep shallow network fusion and attention mechanism. *Algorithms*, 18(8), Article 473. <https://doi.org/10.3390/a18080473>
- Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). CutMix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 6023–6032). IEEE. <https://doi.org/10.1109/ICCV.2019.00612>
- Zhang, H., Cissé, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations (ICLR 2018)*. OpenReview.
- Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. In P. S. Heckbert (Ed.), *Graphics gems IV* (pp. 474–485). Academic Press.