SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

MGT-AATA-Net: Multi-Granular Token and Attention Mechanism Based Facial Expression Recognition

Durgesh Kumar Kotangle[1], H. S. Hota[2*], Shriya Kashyap[3]

[1,2,3]Atal Bihari Vajpeyee University, Bilaspur, India.

ABSTRACT

Facial Expression Recognition (FER) remains a challenging problem in affective computing due to highly heterogeneous spatial characteristics of facial expressions, where emotion-discriminative information may range from global facial configurations to localized muscle movements and subtle micro-textural variations. Conventional convolutional and attention-based FER models predominantly focus on single-scale or visually salient representations, which can overlook low-contrast yet semantically critical facial cues and consequently limit recognition of difficult emotion categories. To address these limitations, this paper proposes Multi Granular Tokenization Alignment-Aware Temporal Attention Network (MGT-AATA-Net), a novel deep learning framework that integrates multi-granular representation learning with alignment-aware attention for robust and discriminative FER. The proposed architecture decomposes facial representations into macro, meso, and micro granular tokens, enabling simultaneous modelling of global facial geometry, intermediate spatial relationships, and fine-grained expression-specific textures. This hierarchical representation is designed to preserve complementary information across spatial frequency levels and thereby improve sensitivity to subtle and localized affective cues. In addition, an Alignment-Aware Temporal Attention (AATA) mechanism incorporates semantic alignment priors derived from the Weighted Directional Centroid Ranking Function (WDCRF) into the attention logits. Unlike conventional self-attention, which may preferentially allocate attention to visually dominant but emotionally irrelevant regions, the proposed mechanism explicitly biases feature interactions toward semantically informative facial regions while suppressing spurious salient responses. On the other hand Occlusion-Adaptive Dynamic Routing (OADR) was also applied to address the occlusion condition during FER process.

The proposed framework is evaluated with RAF-DB dataset, experimental results demonstrate that MGT-AATA-Net achieves overall accuracy of 93.0% and macro-F1 of 91.6%. The robustness of the proposed model was also evaluated through occlusion study, ablation study, GRAD-CAM visibility and ROC-AUC performance study and found satisfactory. These findings established multi-granular semantic alignment as a promising direction for developing more robust FER system capable of capturing both coarse facial geometry and subtle expression-level cues.

Keywords: Multi Granular Tokenisation Alignment-Aware Temporal Attention (MGT-AATA-Net), Facial Expression Recognition (FER), Bidirectional Long Short-Term Memory (BiLSTM), Alignment-Aware Temporal Attention (AATA), Residual Network-18 (ResNet-18), Occlusion-Adaptive Dynamic Routing (OADR).

1.Introduction

Facial Expression Recognition (FER) is intersection of computer vision, cognitive neuroscience, and affective computing, a domain where the demand for accuracy, robustness, and interpretability rises in equal measure with the stakes of deployment. From adaptive learning systems that respond to student frustration, to clinical monitoring platforms that track affective dysregulation, to human-robot interaction frameworks that require real-time empathic responsiveness, the societal utility of reliable FER is substantial and growing. FER has gained considerable importance in a wide range of practical applications by enabling intelligent systems to interpret human emotional states from facial cues. In healthcare, FER can assist in monitoring patients' emotional conditions, assessing pain and psychological states, and supporting personalized care. In education, it can help identify learners' levels of attention, engagement, confusion, and frustration, thereby supporting adaptive learning environments. FER also plays an important role in Human–Robot Interaction (HRI) and Human–Computer Interaction (HCI) by allowing intelligent systems and robots to respond appropriately to users' emotional responses (Rawal & Stock-Homburg, 2022; Sajjad et al., 2023). In surveillance and security, FER can support the analysis of affective or unusual behavioural patterns, while automotive applications can utilize facial expressions to monitor driver stress, fatigue, and emotional state. In addition, FER has potential applications in customer behaviour analysis, advertising, entertainment, virtual assistants, social media analytics, and autonomous systems, highlighting its significance in the development of more adaptive, responsive, and emotionally intelligent technologies (Sajjad et al., 2023).

The human visual system does not process faces through a single, uniform perceptual lens Instead, it processes facial information at different levels, from very small details to the overall structure of the face. These levels are known as micro, meso and macro. Therefore, a good FER system should examine all these levels simultaneously rather than looking at only one image scale. Neuroimaging studies demonstrate a cortical hierarchy in which primary visual cortex (V1) encodes local edge orientations, mid-level areas (V2/V4) encode

curvature and region boundaries, and the Fusiform Face Area (FFA) encodes holistic facial configurations. Crucially, the Superior Temporal Sulcus (STS) identified by (Iidaka, 2014; Kanwisher et al., 1997; Pitcher & Ungerleider, 2021) as the primary locus of dynamic face processing receives converging inputs from all hierarchical levels, enabling integration of micro-textural details with macro-structural geometry. Robust FER requires simultaneous encoding at multiple spatial granularities, not sequential or single-scale processing (Li & Deng, 2022). From a signal-processing perspective, emotion expression distributes information across spatial frequency bands in a non-uniform manner. High-valence expressions such as Happiness or Surprise carry strong low-frequency signals (broad smile geometry, wide eye aperture) which are well-captured by coarse receptive fields. Negative or subtle emotions Disgust, Fear manifest primarily in high-frequency micro-textural patterns: localised wrinkling around the nose bridge (AU9), asymmetric nasolabial fold tension (AU18), and periocular tightening (AU7). Models with fixed or predominantly coarse receptive fields are structurally ill-equipped to detect these high-frequency cues, which explains the persistent under-performance on Fear and Disgust.

This research work proposes a novel attention based model: MGT-AATA-Net, a hybrid of ResNet-18–BiLSTM–WDCRF, MGT, AATA and OADR which establishes two key insights: First pseudo-temporal sequence unrolling of convolutional feature maps provides meaningful relational context that static ResNet-18 cannot capture; and second, that handcrafted descriptors, when aligned through the WDCRF, provide semantically grounding priors that stabilise deep feature learning (He et al., 2016; Hasani & Mahoor, 2017). MGT-AATA-Net outperforms with 93% accuracy, 91.8% precision, 91.3% sensitivity and 99.1% specificity. Further occultation was also performed and found that proposed MGT-AATA is working well in occultation conditions. Performance degraded sharply under occlusion by up to 24% when the mouth or eye regions were masked. Additionally, minority emotion classes (Fear and Disgust) remained problematic (Gao & Zhao, 2021).

The remaining parts of this paper is organized as follow: Section 2 describes related work, section 3 explains about detail methodology with subsections of datasets used to develop model, section 4 provides details of proposed MGT-AATA-Net with its flow diagram and algorithm while section 5 explores experimental and comparative results of various developed models and also compared with current-state-of-the-art work. In the last section, in section 6 performance of the models were analysed through ablation, occultation and Explainable AI studies and finally paper is concluded through conclusion as section 7.

2.Related Work

Reghunathan et al., (2024) have studied FER with pre-trained model as ensemble model using FER 2013 dataset and concluded that there is a benefits of combining pre-trained model as ensemble model for FER. A Deep Neural Network (DNN) based architecture followed by Conditional Random Field (CRF) is proposed by (Hasani et al., 2017). CRF is used to capture temporal relation between the image frames of video data, which is second part of the architecture. Experiments were done using three publically available data: CK+, MMI and FERA and results reveals that CRF has significant role in process of FER. A simple but informative research by (Ge et al., 2022) reported about current research status and datasets for FER, further this study explains about advanced deep facial expression.

Transformer is a kind of Neural Network with self-attention mechanism which is widely used for global context information. In more advanced version of work on FER is done by (Bie et al., 2024) by using transformer-based model: Shifted Window Transformer (SWIN) and proposed a new transfer-based model SQIN-FER for FER, the models were tested on FER 2013 and CK+ datasets and achieved 71.11% and 100% accuracy respectively. Another transformer-based model, TEF is proposed by (Gao et al., 2021) under occlusion condition. Models were tested on various datasets and achieved 81.33% average accuracy on RAF-DB, 63.33% on EffectNet. Another transfer based model Fine-Tuned Channel Spatial Attention Transformer (FT-CSAT), proposed by (Yao, 2023). Model was tested on two benchmark datasets: RAF-DB and FER+ and achieved 88.61% and 89.26% accuracy respectively.

Another research by (Chen et al., 2023) explains about occlusion facial expression recognition based on feature fusion residual attention network (FERA-Net). Experiments were performed on two occlusion-based datasets obtained from RAF-DB and achieved excellent results. Hiding face in any form like using face mask is hurdle while identifying facial emotions (Carbon, 2020). The experimental results derived shows impact of facial mask while determining facial emotions, this also happens in our daily life. Another similar work Grundmann et al. (2021) also studied impact of face mask in FER process, they have concluded that face mask reduces performance of emotion recognition, experiments were done through statistical techniques.

The motivation for alignment-aware attention arises from a well-documented failure mode of standard self-attention in FER, attention collapse toward visually salient but emotionally irrelevant tokens. In transformer-based FER models, attention heads consistently over-weight high-contrast regions (bright teeth, specular highlights on the cornea) while under-weighting semantically critical but low-contrast regions (nasolabial folds in Disgust, subtle brow asymmetry in Fear) (Gao & Zhao, 2021; Xue et al., 2021). Recently attention based mechanism is gaining more attention for developing model for FER due to its importance for achieving better performance. A series of models have been developed by various authors for FER and tested on multiple datasets. Attention mechanism is inspired by human visual perception in which human focuses on a specific

region of face to recognize human emotion (Guo D. et al., 2026). Attention based mechanism focuses on relevant and important regions and simultaneously ignore irrelevant regions (Biswas et al., 2026). Many innovative models were developed, ExpressNet-MoE by (Banerjee et al., 2026) using Adaptive feature learning and Multi-scale feature extraction, (Raju, 2026) investigate squeeze-and-excitation placement in residual networks for FER, (Garcia et al., 2026) studied about FER in younger and older adults.

Authors (Li et al., 2020) have developed novel deep Convolutional Neural Network (CNN) with an attention module and tested on various FER datasets: CK+, JAFEE, Oulu-CASIA and FER 2013. Results revels that attention module has significant impact on FER. (Khan et al., 2025) have proposed EA-Net with attention module under backbone of EfficientNetB0 and InceptionV3. The model incorporated channel attention and spatial attention modules and performances were compared with current-stat-of-the-art work and found better with F1-score=77.0% and 99.0% for FER and KDEF datasets respectively. In another recent work (Guo & Xu, 2026) have developed an attention based model: AUNet. This architecture is based on Action Unit (AU) hierarchical attention network with global and local features extraction mechanism using ResNet-18 as backbone network. Accuracy of the AUNet on different datasets are: 67.51% for AffectNet-7, 90.52% for FERPlus, 91.75% for RAF-DB and 75.0% for FED-RO.

Authors (Fan et al., 2026) have proposed SPA-SE network addresses issues through three components: Sliding Patch (SP), Spatial-level Patch Attention (SPA), and Channel-level Attention (SE). SP optimizes patch size and stride, SPA emphasizes important regional features, and SE enhances salient feature maps. Experiments on five pose-invariant FER datasets demonstrate improved recognition performance, achieving accuracies ranging from 59.84% to 86.77% across controlled and real-world datasets. (Mishra et al., 2026) have developed a model: EmoVisioNet which is a CNN and attention based vision model. The proposed EmoVisioNet model achieves high recognition accuracy across four benchmark datasets, obtaining 99.97% on CK+, 96.23% on RAF-DB, 93.88% on FER2013, and 96.91% on FERPlus. These results demonstrate the model's effectiveness and robustness in facial expression recognition and indicate its competitive performance against existing state-of-the-art approaches. (Zhu et al., 2022) proposed DPSAN with attention based module and achieved 93.2% and 87.63% accuracy on the CK+ dataset and FERPlus dataset, respectively. DLRM-FER developed by (Guo et al., 2026) and achieved competitive performance (Accuracy) with 70.81%, 85.06% and 61.24% respectively for FER2013, RAF-DB and AffectNet datasets. ExpressNet-MoE, proposed by (Nanarjee et al., 2026) is a combination of CNN and Mixture of Experts (MoE) and tested on AffectNet, RAF-DB and FER 2013.

Multi Head Attention (MHA) is an attention mechanism in which multiple heads operate in parallel manner. Authors (Biswas et al., 2026) have used multihead attention mechanism for FER on three different datasets and achieved highest accuracy in case of proposed model: ATR-Net.

Literature review supports that none of the research addresses about adoption of multiple modules like attention, occlusion and handcrafted feature alignment for improvement of performance of FER. The proposed MGT-AATA-Net model is capable to address all these issues.

3. Methodology

3.1 Data Description

Dataset plays crucial role for development of model. A brief review of datasets used in FER is well explained by (Li & Deng, 2022). The Real-World Affective Face Database (RAF-DB) is real world dataset collected from Internet contains 29,672 facial images having two different subsets: single label subset of 7 basic emotions classes and two-tab subset of compound emotions. A total of 15,339 images belongs to single label subset and remaining images are related to compound emotions (Lie et al., 2017). Sample images of all 7 emotion classes of RAF-DB dataset is shown in Figure 1.

Fig. 1: Sample raw images from each emotion class of RAF-DB dataset.

3.2 Data Pre-processing

Dataset goes through the pre-processing pipeline: Landmark alignment (Kazemi & Sullivan, 2014), Contrast

enhancement, Local Binary Pattern (LBP) channel fusion (Ojala et al., 2002), before being pooled into the model. Identity-aware stratified 10-fold splits are used to ensure no subject appears in both training and test folds, even across datasets. Fold-wise z-score normalization is computed on training data only and applied to test, preventing data leakage. The details of each stage of data pre-processing is shown in Figure 3 and are explained in more details as below:

i. Face Detection and Landmark Alignment: Each image is processed by a five-point landmark detector, and a similarity transformation aligns the detected face to a canonical template. The 68-point dlib v19.24 shape predictor then localizes finer landmarks for cropping. This two-stage alignment ensures that inter-image geometric drift which can cause a BiLSTM encoder to misinterpret camera jitter as expressive motion is eliminated before tokenization.

ii. Contrast Enhancement: Contrast-Limited Adaptive Histogram Equalization (CLAHE) is applied with a clip limit of 2.0 and tile grid of 8×8 in the LAB color space. CLAHE is preferred over global histogram equalization because it enhances local contrast adaptively; preserving subtle cues like eyebrow furrows and nasolabial creases without over-amplifying background noise. Following CLAHE, frames are normalized to zero mean and unit standard deviation per batch. Figure 2 shows effect of CLAHE for contrast enhancement of sample images. The images at right side are enhanced, which is also reflected through uniform distribution of pixel intensity of the corresponding image. Following CLAHE, frames are normalized to zero mean and unit standard deviation per batch (Zuiderveld, 1994). Figure 2 shows effect of CLAHE for contrast enhancement of sample images.

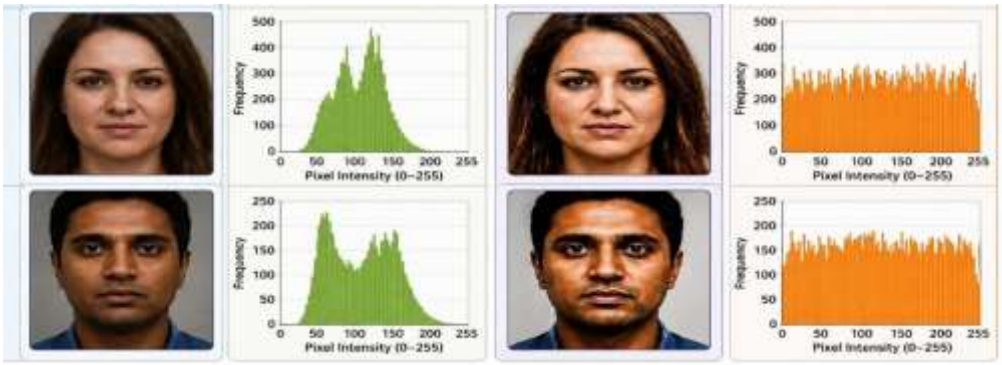

Fig. 2: Effect of CLAHE Pre-processing: Image and histogram equalisation before and after.

iii. LBP Channel Fusion: A uniform-pattern LBP descriptor (radius=2, points=8) is computed for each pre-processed frame and appended as a fourth input channel alongside the RGB channels. This dual-stream input colour + texture preserves the micro-textural priors that were found in to be critical for Disgust and Fear classification, while providing the ResNet-18 backbone with richer information than color alone.

iv. Synthetic Occlusion Augmentation: To train the Occlusion-Adaptive Dynamic Routing (OADR) occlusion estimator, synthetic occlusion masks are applied to 40% of training images. Masks are drawn from two distributions: rectangular masks with randomized positions (forehead, mouth, left eye, right eye, and random location), and irregular masks generated via Perlin-noise contour functions. The inclusion of Perlin-based irregular masks finding that models trained exclusively on rectangular occlusions develop brittle high-frequency geometric assumptions that fail under naturalistic occlusions (hair, hands, medical masks).

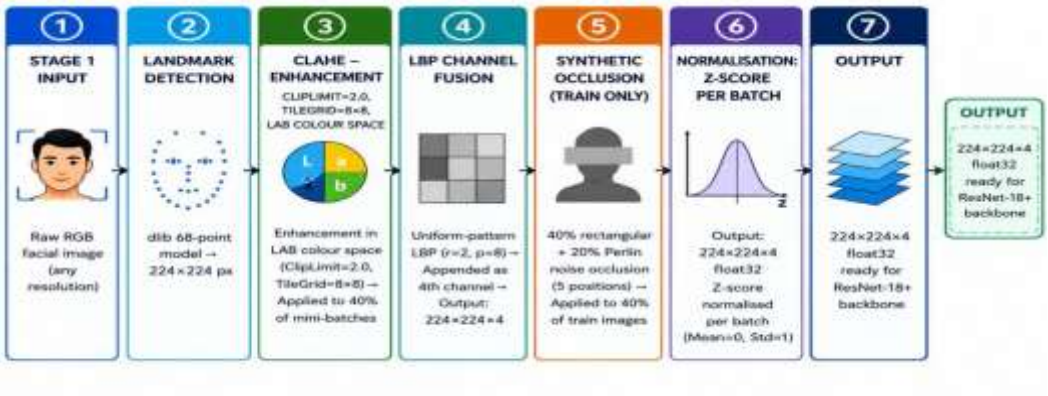

Fig. 3: Seven-Stage facial image processing pipeline.

3.3 Weighted Directional Centroid Ranking Function (WDCRF)

WDCRF is a hand-crafted feature alignment and semantic grounding module that bridges the gap between

traditional descriptor-based features (LBP, HOG, Gabor) and deep neural network representations. Its job is to ensure that what the deep model learns stays anchored to psychologically validated, human interpretable facial cues preventing the deep features from drifting toward patterns that are statistically convenient but emotionally meaningless. The four components as per its name are as below:

Weighted: each hand-crafted feature dimension and each emotion class centroid is assigned a weight based on two factors: (1) inverse class frequency (rarer classes like Disgust and Fear get higher weight), and (2) a difficulty correction factor derived from early-epoch macro-F1 scores (classes that the model finds harder to learn receive stronger alignment pressure). The weighting prevents the dominant classes (Happy, Neutral) from overwhelming the learning signal.

Directional: the alignment is not symmetric. It operates as a directional projection: the hand-crafted features are projected toward the class-conditional centroids using cosine similarity, ranking each feature dimension by how much it points in the direction of the correct emotion class. This directional ranking filters out feature dimensions that are noisy or irrelevant for a given emotion, retaining only those that point meaningfully toward the correct class manifold.

Centroid: for each emotion class, a centroid is computed in the hand-crafted feature space — the average LBP + HOG + Gabor descriptor vector for all training samples of that class. These centroids act as semantic anchors: they represent what each emotion class "looks like" in terms of handcrafted micro-texture, structural gradient, and frequency energy. Every input sample is then compared to all seven centroids to determine which emotion it is most aligned with.

Ranking Function: the core operation is a ranking procedure. For each input image, cosine similarities between its projected hand-crafted features and all class centroids are computed. These similarities are ranked, and the top-m most semantically aligned feature components are selected to form the final WDCRF-aligned vector. This ranking ensures that only the most discriminative and directionally consistent feature components contribute to the alignment discarding noise, low-confidence dimensions, and cross-class ambiguous signals.

3.4 ResNet-18 and BiLSTM

The ResNet-18 encoder is based on a modified ResNet-18 architecture, originally designed for large-scale image recognition and used in this research for emotion recognition. To ensure compatibility with pre-trained weights on ImageNet, all input images are first resized to 96 × 96 and replicated across three channels, producing a pseudo-RGB input of shape:

Input Tensor: $R^{N\times96\times96\times3}$

Where:

- N is the batch size.
- 96×96 is the spatial resolution.
- 3 denotes the RGB color channels.

BiLSTM feature map is flattened into 36 spatial tokens and fed as a pseudo-sequence into a Bidirectional LSTM. This lets the model capture relational context between facial regions (e.g., raised brows only mean Surprise when the eyes are also wide).

Whereas convolutional architectures excel at encoding local spatial dependencies, they remain agnostic to sequential structure a critical limitation when modelling temporally progressive affective states, even in static image corpora like FER2013. The integration of a BiLSTM module addresses this by imposing a pseudo-temporal ordering over spatial patches derived from ResNet-18 feature maps, thereby enabling the model to extract context-sensitive emotional patterns distributed across the facial topology.

The output tensor from the ResNet-18 + LBP fusion block is reshaped into a sequence of the form:

$X \in R^{N\times T\times D}$ where $T = 36$ and $D = 512$

Each of the T=36 tokens correspond to a flattened patch from the 6×6 spatial grid, and each token is a 512-dimensional vector encoding the local convolutional and texture-aware features.

3.5 Evaluation Metrics

Primary evaluation metric is Macro-F1, which assigns equal weight to each emotion class regardless of prevalence (Sokolova & Lapalme, 2009). Secondary metrics include: Accuracy, Sensitivity (Recall), Specificity, Precision, ROC-AUC (Brodersen et al., 2010). All metrics are reported as means ± standard deviations across 10-fold cross validation.

4.Proposed MGT-AATA-Net

The architectural choices defining MGT-AATA-Net were driven by a single governing principle: emotion recognition demands a representational structure that is simultaneously multi-scalar, semantically constrained, and explicitly resistant to regional uncertainty (He et al., 2015; Schuster & Paliwal, 1997). No prior FER architecture satisfies all three constraints within a single differentiable system (Vaswani et al., 2017). MGT-AATA-Net satisfies all three by organising its pipeline mainly into five tightly coupled stages: MGT, BiLSTM Pseudo-Temporal Encoding, Alignment-Aware Temporal Attention (AATA), WDCRF Semantic Alignment Gate, and Occlusion Dynamic Routing (OADR) and finally feature fusion and classification as shown in Figure 4. The

following sections describe each stage in detail, preceded by pseudocode algorithm that provide a holistic overview of the pipeline of the proposed research (Deng et al., 2019). The diagram shows the progression from a raw input image through various stages to a final predicted emotion label, with data tensor dimensions annotated at each transition to facilitate reproducibility. Each stage is explained in more detail as below:

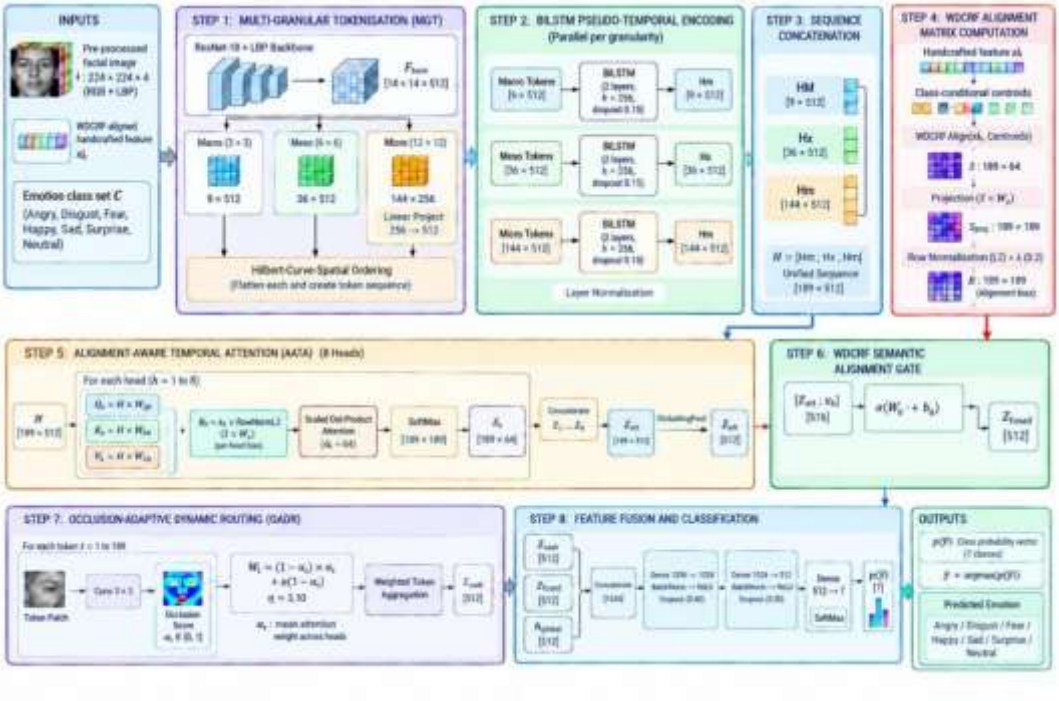

Fig. 4: Flow diagram of proposed MGT-AATA-Net.

3.1 Multi-Granular Tokenization (MGT)

The MGT module is the architectural foundation of MGT-AATA-Net, resolving the single-scale limitation identified. The module operates on the combined feature map $F_{base} \in \mathbb{R}^{14\times14\times512}$ obtained through image pre-processing technique as shown in Figure 3, produced by the ResNet-18 + LBP backbone, as illustrated in Figure 4, the input image (photometrically normalized, face-aligned, landmark-cropped, and LBP-augmented) enters the modified ResNet-18 backbone, which produces a 14×14×512 combined feature map. This map feeds into the MGT module, which extracts three parallel token sets at Macro, Meso, and Micro granularities each representing a distinct spatial granularity with its own characteristic receptive field and informational content. Macro tokens are large facial region and capture coarase facial regain, Meso represents medium scale facial structure while Micro token represents local facial information.

In case of Macro token ($F_M \in R^{3\times3\times512}$), adaptive average pooling over 3×3 spatial windows produce 9 macro tokens, each encoding broad geometric structure covers 20–25% of the facial region. The effective receptive field of each macro token in the original 224×224 image space is approximately 45–55 pixels large enough to capture brow-line elevation, cheek curvature, jawline tension, and global facial symmetry. These cues are most discriminative for high-arousal, spatially diffuse expressions such as Happiness, Surprise, and Anger. Macro tokens also provide representational stability under mild occlusions, as each token aggregates information from a broad neighborhood rather than a single local patch.

Meso Tokens ($F_X \in R^{6\times6\times512}$), adaptive pooling produces 36 Meso tokens, each covering approximately 8–10% of the facial area with an effective receptive field of ≈22–28 pixels. Meso granularity is the representational scale most critical for encoding relational asymmetries left-right eyebrow differences that discriminate Anger from Disgust, asymmetrical lip pulls that distinguish contempt from sadness, and periocular wrinkles that indicate genuine versus posed positive affect (Duchenne marker, AU6). This intermediate scale bridges the representational gap between macro structural geometry and micro-textural detail that neither single-scale ResNet-18 nor transformer patch embedding adequately cover.

Micro Tokens ($F_m \in R^{12\times12\times256}$), fine-grained adaptive pooling produces 144 micro tokens, each covering approximately 1–2% of the facial area with an effective receptive field of ≈8–10 pixels. At this scale, the dominant informational content is micro-textural: orbicularis oculi contraction (AU6, genuine happiness), nasolabial fold modulation (AU18, Disgust), nasal bridge wrinkling (AU9, Disgust), and levator labii contraction (AU10, Disgust/Fear). These cues are invisible to Macro and Meso tokens, explaining why Disgust and Fear, which depend critically on nasal and peri nasal micro-textures are systematically underperformed by single-scale architectures. The micro token dimensionality is intentionally reduced to 256 as compared to 512 for Macro and Meso to avoid noise amplification and reduce downstream sequence computation costs. A learned linear projection aligns micro tokens to 512-D before BiLSTM encoding. The three token sets are flattened into

ordered sequences using a Hilbert space-filling curve, which maintains spatial locality tokens that are spatially adjacent in the original feature map remain temporally adjacent in the flattened sequence. This ordering reduces the directional discontinuity artefacts that arise from raster-scan flattening, particularly in the Meso and Micro streams where spatial adjacency carries strong relational information (Gotsman & Lindenbaum, 1996). Hilbert Curve Spatial Ordering is a method for converting a 2D spatial arrangement of image tokens into a 1D sequence, while trying to preserve the original spatial relationships between neighbouring regions which is essential for BiLSTM. It does not create new feature rather it determines the order in which tokens are placed in a particular position before sequential processing. Different granular generated by this module is shown in Figure 5.

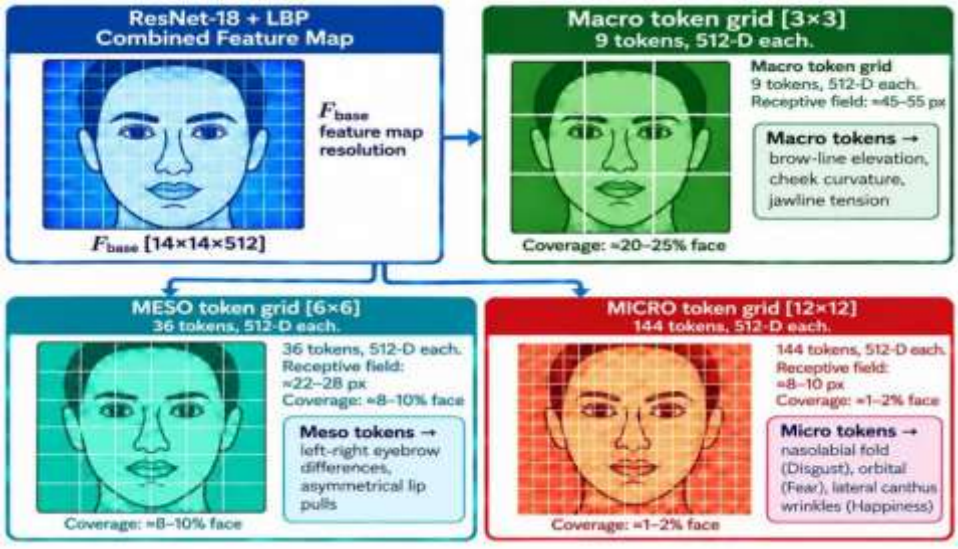

Fig. 5: Three different granularity of an input image.

3.2 BilSTM Pseudo Temporal Encoding

Each token set is independently processed by a two-layer BiLSTM encoder (Schuster & Paliwal, 1997), preserving scale-specific relational structures of the three granular token sequences with hidden size 256 per direction (512 after concatenation), dropout of 0.15 between layers, and LayerNorm applied to outputs. The use of independent BiLSTMs per granularity rather than a single shared encoder is a deliberate architectural choice, it ensures that scale-specific relational structures are learned at each granularity. The three encoded sequences $H_M \in \mathbb{R}^{9\times512}$, $H_X \in \mathbb{R}^{36\times512}$, $H_m \in \mathbb{R}^{144\times512}$ are concatenated into a unified sequence $H \in \mathbb{R}^{189\times512}$ for further process to next stage where 189 is number of facial tokens and 512 is number of features in each token.

3.3 Alignment-Aware Temporal Attention (AATA)

In computer vision, attention mechanism is widely used which allow to focus on specific region of face while recognizing an emotion. In order to provide attention to specific part of an input image it assigns weight. Many authors have proposed many models for attention mechanism for different tasks (Wang et al., 2026). Mathematically attention mechanism can be represented as:

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

where:
Q (Query) → what information a token is looking for
K (Key) → what information each token contains for matching
V (Value) → the actual information transferred
$\sqrt{d_k}$ → dimension of the key vectors
In this study AATA is designed and is integrated with MGT to make it MGT-AATA-Net, it is the intellectual core part of MGT-AATA-Net. Standard Multi-Head Self-Attention (MHSA) computes token-to-token relevance as a function of learned QK^T similarity, without any mechanism to bias attention toward emotionally relevant tokens. This makes standard attention susceptible to attention collapse toward visually salient but emotionally irrelevant regions, a failure mode documented in transformer FER systems by (Yao et al., 2023; Zhang et al., 2024). AATA resolves this by injecting handcrafted alignment priors into the attention logit computation.
The WDCRF introduced and builds class-conditional alignment vectors by projecting handcrafted features (LBP histograms, Gabor energy, HOG descriptors) into discriminative manifolds through a weighted directional centroid ranking procedure. From the WDCRF, an alignment matrix $S \in \mathbb{R}^{189\times64}$ is derived, where each entry $S_{\{th\}}$ represents the semantic compatibility between token t and handcrafted feature dimension h. This matrix is projected into the token–token space: $S_{proj}=S \times W_S \in \mathbb{R}^{189\times189}$, and then normalized and scaled to produce an

alignment bias

$$B = \lambda RowNarm_{L_2}(S_{proj}) \quad \ldots(2)$$

Where λ is the scaling/hyper parameter, S_{proj} is the alignment information S and L_2 is the row normalization? According to equation 1 of basic attention formulae, alignment bias (B) is added and the AATA computation (Attention score) can be computed as below (Biswas S. and et al., 2026):

$$AATA(H) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \quad \ldots(3)$$

Where $H \in \mathbb{R}^{189X512}$ is the input temporal feature representation and Q, K and V can be calculated as below:

Query matrix (Q) = HW_Q

Key matrix (K) = HW_K

Value matrix (V) = HW_V

Where W_Q, W_K and W_V are learnable weight matrices updated during model training.

K^T is the transpose of key matrix

QK^T is the Query-Key similarity scores

d_k is the Key vector of dimension

$\sqrt{d_k}$ is the scaling factor to stabilize the gradient

B is the alignment-aware bias matrix, softmax converts the attention scores into normalized weights, V is the value representation and AATA(H) is the final alignment-aware temporal attention representation (Vaswani et al., 2017).

This formulation introduces two interacting forces within the attention mechanism: intrinsic similarity (QK^T), learned from data through gradient descent; and semantic alignment (B), imposed by handcrafted priors derived from psychologically validated FACS-based micro-texture descriptors. The alignment bias is injected before softmax normalization, ensuring that it influences the relative rather than absolute magnitude of attention weights, a design choice that prevents the bias from dominating the attention distribution and maintains gradient flow through the intrinsic similarity pathway (Vaswani et al., 2017).

AATA is extended to 8 attention heads, each receiving an independent per-head alignment projection $W_S^{(h)}$. Number of heads considered internationally 8 in this study, more number of attention heads may improve feature diversity and hence will improve performance of the model but at the same time computational cost will also increase (Biswas et al., 2026). In order to make balance in between performance of the model and computational cost, an optimized number of attention head must be selected. This allows different heads to specialize: heads 1–3 attend to micro-textural patterns aligned with LBP/Gabor cues (most diagnostic for Fear and Disgust); heads 4–6 align toward structural cues from Macro tokens (most diagnostic for Happiness and Surprise); and heads 7–8 emphasize relational asymmetry from Meso tokens (most diagnostic for contempt and Anger). Output for one head is $Z_h \in \mathbb{R}^{189X64}$, since we have 8 attention heads hence all 8 attention heads Z_{att} = Concat (Z_1, Z_2…… Z8) will produce $Z_{att} \in \mathbb{R}^{189X512}$ and finally by applying globalAvgPool (Z_{att}), $Z_{att} \in \mathbb{R}^{512}$ will be obtained, which is outcome from AATA module.

3.4 WDCRF Semantic Alignment Gate

WDCRF is elevated from a fusion module into the attention mechanism itself. The alignment matrix S derived from WDCRF is projected into token–token space and added as a bias B directly into the attention logit computation of equation 1. This means WDCRF no longer just adds a feature, it actively steers which tokens the model pays attention to, making semantically important tokens (those matching known micro-texture priors for a given emotion) receive higher attention weight before softmax. The Semantic Alignment Gate then governs per-sample fusion of attended deep features and hand-crafted features through a learned sigmoid gate, preventing either stream from dominating unconditionally (Vaswani et al., 2017).

The WDCRF Semantic Alignment Gate (SAG) governs per-sample integration of the attended deep embedding $Z_{att} \in \mathbb{R}^{512}$ and the WDCRF-aligned handcrafted embedding $X_h \in R^{64}$ .The gate function is as below (Mishra et al., 2026):

$$g = \sigma\left(W_g[Z_{att}; X_h] + B_g\right) Wg \in R^{(576\times512)} \quad \ldots(4)$$

$$Z_{fused} = g \odot Z_{att} + (1-g) \odot W_h X_h \quad \ldots(5)$$

Where:

W_g and B_g are weight and bias parameters, Z_{att} is obtained from AATA module, X_h is WDCRF aligned feature and W_h is learnable weight matrix.

The sigmoid gate $g \in R^{512}$ is learned from the concatenation of both representations, enabling the model to estimate, on a per-sample basis, which representation is more informative. When g → 1, the deep attended embedding dominates appropriate when the image is clean and the deep representation already captures sufficient discriminative structure (Sun & Gao, 2026). When g → 0, the handcrafted representation dominates appropriate under occlusion or illumination degradation, where deep features may be corrupted but handcrafted descriptors remain informative due to their robustness. The gating mechanism prevents both representations from dominating unconditionally, addressing the fusion inconsistency problem documented by (Kumar, 2024) in additive hybrid architectures.

3.5 Occlusion-Adaptive Dynamic Routing (OADR)

OADR is the module that most directly addresses the performance degradation under occlusion observed (Wegrzyn et al., 2017). Its design principle is straightforward: occluded tokens should not be discarded entirely (which would cause representational discontinuity) (Ding et al., 2020), but their contribution to the context vector should be reduced proportionally to their estimated occlusion likelihood, while the redistributed attention mass is absorbed by neighbouring or semantically compatible tokens (Li et al., 2019; Wang et al., 2020).

A lightweight ResNet-18 (two 3×3 convolutional layers with ReLU, 16 channels, followed by sigmoid output) is applied to each token's spatial patch to predict occlusion likelihood $o_t \in [0,1]$. The OADR corrected token weight can be calculated as:

$$w_t = (1 - o_t)\alpha_t + \eta(1 - \alpha_t), \eta \approx 0.10 \quad \ldots(6)$$

where α_t is the mean attention weight from AATA and η is a stability floor that ensures fully occluded tokens still contribute marginally. The first term reduces the contribution of occluded tokens; the second term ensures that unattended tokens with low occlusion probability receive a minimum contribution through the stability floor. The OADR-weighted context vector is

$$Z_{oadr} = \sum_{t=1}^{T} w_t Z_{att}[t] \quad \ldots(7)$$

The occlusion estimator is trained with a binary cross-entropy loss L_{occ} against synthetic occlusion mask labels, which are available for all training images through the augmentation protocol described. Under adversarial occlusion conditions (e.g., a medical mask covering the lower face), OADR dynamically reduces attention weight on mouth and chin tokens and increases weight on periocular and upper-face tokens. MGT-AATA-Net thus exhibits occlusion-adaptive behaviour that is not only computationally effective but psychologically plausible (Ding et al., 2020; Li et al., 2019; Wang et al., 2020).

3.6 Feature fusion and classification

The final classification stage consolidates three parallel feature streams: $Z_{oadr} \in \mathbb{R}^{512}$ (OADR-weighted attended representation), $Z_{fused} \in \mathbb{R}^{512}$ (WDCRF-gated handcrafted deep fusion), and $h_{global} \in \mathbb{R}^{512}$ (Global average pooled backbone features from ResNet-18 and LBP modules). Concatenation yields a 1536 dimensional combined representation, which is classified through a two-layer dense network:

Dense (1536→1024) → BatchNorm → ReLU → Dropout(0.40) → Dense(1024→512) → BatchNorm → ReLU → Dropout(0.30) → Dense(512→7) → softmax and finally obtained class probability ($p(\hat{y})$) and predicted emotion class ($\hat{y}$).

3.7 Algorithm Pseudocode

Following the complete algorithmic logic of MGT-AATA-Net is specified below as a structured pseudocode.

Algorithm: MGT-AATA-Net

1. INPUTS

I: Pre-processed facial image [224×224×4] (RGB + LBP channel)
X_h : WDCRF-aligned handcrafted feature vector.
C: Emotion class set {Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral}

2. OUTPUTS:

$\hat{y}$: Predicted emotion label
$p(\hat{y})$: Class probability vector

3. STEP 1: MULTI-GRANULAR TOKENISATION (MGT)

4. $F_{base} \leftarrow f_{ResNet\text{-}18\_LBP}(I)$ // backbone output: [14×14×512]
5. $F_M \leftarrow$ AdaptiveAvgPool(F_{base}, 3×3) // macro tokens: [9×512]
6. $F_X \leftarrow$ AdaptiveAvgPool(F_{base}, 6×6) // meso tokens: [36×512]
7. $F_m \leftarrow$ AdaptiveAvgPool(F_{base}, 12×12) // micro tokens: [144×256]
8. $F_m \leftarrow$ LinearProject(F_m, 256→512) // align to 512-D: [144×512]
9. Flatten each F_M, F_X, F_m using Hilbert-curve spatial ordering
10. h_{global} = GlobalAvgPool (F_{base}) // backbone global feature: [512]

11. STEP 2: BiLSTM PSEUDO TEMPORAL ENCODING (parallel, per granularity)

12. $H_M \leftarrow BiLSTM_{2\text{-}layer}(F_M$, hidden=256, dropout=0.15)// [9×512]
13. $H_X \leftarrow BiLSTM_{2\text{-}layer}(F_X$, hidden=256, dropout=0.15) // [36×512]
14. $H_m \leftarrow BiLSTM_{2\text{-}layer}(F_m$, hidden=256, dropout=0.15) // [144×512]
15. Apply LayerNorm to each H_M, H_X, H_m

16. STEP 3: SEQUENCE CONCATENATION

17. $H \leftarrow$ Concatenate (H_M, H_X, H_m) // unified sequence: [189×512]

18. STEP 4: WDCRF ALIGNMENT MATRIX COMPUTATION

19. // Derive class-conditional handcrafted centroids
20. $S \leftarrow WDCRF_{Align}(x_h, class_{centroids})$ // alignment matrix: [189×64]
21. $S_{proj} \leftarrow S \times W_S$ ($W_S \in R^{\{64\times189\}}$) // project to token space: [189×189]
22. $B \leftarrow \lambda \times RowNorm_{L2}(S_{proj})$ // alignment bias: [189×189]

23. // λ is a learned scalar initialised at 0.2
24. STEP 5: ALIGNMENT AWARE TEMPORAL ATTENTION (AATA, 8 heads)
25. **For** each head h = 1 to 8
26. $Q_h = H \times W_Q^h$, $K_h = H \times W_K^h$, $V_h = H \times W_V^h$
27. $B_h = \lambda_h \times \text{RowNorm}_{L2}(S \times W_S^h)$ // per-head alignment bias
28. $A_h = \text{Softmax}((Q_h \times K_h^T)/\sqrt{d_k} + B_h)$ // [189×189]
29. $Z_h = A_h \times V_h$ // [189×64]
30. End For
31. $Z_{att} = \text{Concatenate}(Z_1 \ldots Z_8)$ // [189×512]
32. $z_{att} = \text{GlobalAvgPool}(Z_{att})$ // [512]
33. STEP 6: WDCRF SEMANTIC ALIGNMENT GATE
34. $g = \sigma(W_g \times [z_{att} ; x_h] + B_g)$ // $W_g \in R^{\{576\times512\}}$, gate: [512]
35. $z_{fused} = g \odot z_{att} + (1 - g) \odot (W_h X_h)$ // gated fusion: [512]
36. STEP 7: OCCLUSION ADAPTIVE DYNAMIC ROUTING (OADR)
37. **For** each token t = 1 to 189
38. $O_t = \sigma(\text{Conv3×3}(\text{TokenPatch}_t))$ // occlusion likelihood ∈ [0,1]
39. $W_t = (1 - O_t) \times \alpha_t + \eta(1 - \alpha_t)$ // η = 0.10 (stability floor)
40. //α_t is the mean attention weight across heads for token t
41. $Z_{oadr} = \Sigma_t W_t \times Z_{att}[t]$ // OADR-weighted context: [512]
42. End For
43. STEP 8: FEATURE FUSION AND CLASSIFICATION
44. $h_{combined} = \text{Concatenate}(z_{oadr}, z_{fused}, h_{global})$ // [1536]
45. h1 = Dense (1536→1024) → BatchNorm → ReLU → Dropout(0.40)
46. h2 = Dense (1024→512) → BatchNorm → ReLU → Dropout(0.30)
47. p(ŷ) = SoftMax (Dense (512→7) (h2)) // class probabilities: [7]
48. ŷ = argmax (p(ŷ)) // predicted emotion class

4. Experiment and Results

This section elaborates experimental environment of developing models and obtained results as per class wise and overall. The results reported in this section are the empirical substantiation of the theoretical architecture described. They are presented not as a catalogue of numbers but as a structured evidentiary argument: each table and figure are positioned to answer a specific analytical question about MGT-AATA-Net's behaviour and to demonstrate that performance improvements are architecturally motivated rather than incidental. Section also presents ablation study, cross-dataset transfer and occlusion robustness and compares against current state-of-the-art work (Paszke et al., 2019).

4.1 Experimental environment

All experiments and model training were conducted through a high end computer system with 11th Gen Intel(R) Core(TM) i7 CPU, 2.50 GHz, 32.0 GB RAM, Nvidia 3060 RTX GPU. Programs were written through Python 3.14, and used PyTorch and Scikitlearn libraries under Windows 10 operating system. Other parameters are as shown in Table 1.

Table 1: Training Parameters.

Parameter	Value
Epochs	120
Batch size	64
Optimizer	AdamW
Initial Learning Rate for ResNet18	$1e^{-4}$
Initial Learning Rate for BiLSTM	$1e^{-3}$
Learning Rate Scheduler	CosineAnnealingLR (T_max=60)
Weight Decay (L2 Reg)	0.0005
Gradient Clipping	max_norm=5.0
Dropout (BiLSTM)	0.3
DropBlock (CNN encoder)	block_size=3, drop_prob=0.25
Patience (Early Stopping)	12 epochs (based on val F1)
Loss Function	Weighted Cross-Entropy + Focal

4.2 Per class-wise performance of models

Models were trained using k-fold cross validation techniques through the data as explained in sub section 3.1

for 120 epochs. Pre-processing of data as explained in sub section3.2 were used to train various baseline models like ResNet-18, BiLSTM and hybrid of ResNet-18, BiLSTM and ResNet-18-BiLSTM-WDCRF along with proposed MGT-AATA-Net. Table 2 reports per class wise performance for the ResNet-18-only (B1), establishing the lower-bound reference. The consistent under-performance on Disgust (F1-score=68.7%) and Fear (F1-score=70.7%) confirms the single-scale ResNet-18 inability to capture the micro-textural cues that are the primary discriminates for these two classes (Wegrzyn et al., 2017; Li et al., 2019).

Table 2: Class-wise performance (In %) ResNet-18 Only (B1).

Emotion Class	Accuracy	Precision	Sensitivity (Recall)	Specificity	F1-Score	ROC-AUC
Angry	77.8 ± 1.8	74.2 ± 2.1	72.5 ± 2.2	91.4 ± 0.9	73.3 ± 2.1	0.881
Disgust	72.4 ± 2.5	69.1 ± 3.2	68.4 ± 3.0	95.2 ± 0.8	68.7 ± 3.0	0.842
Fear	74.9 ± 2.2	71.3 ± 2.8	70.2 ± 2.8	89.8 ± 1.0	70.7 ± 2.8	0.858
Happy	91.2 ± 1.0	91.4 ± 1.1	89.8 ± 1.2	97.5 ± 0.4	90.6 ± 1.1	0.978
Sad	75.8 ± 1.9	71.4 ± 2.2	70.3 ± 2.4	88.9 ± 0.9	70.8 ± 2.3	0.877
Surprise	87.2 ± 1.4	85.5 ± 1.7	86.3 ± 1.6	95.4 ± 0.6	85.9 ± 1.6	0.936
Neutral	79.1 ± 1.6	75.9 ± 2.0	74.8 ± 2.1	91.4 ± 0.8	75.3 ± 2.0	0.903
Macro Average	79.8 ± 1.2	77.0 ± 1.4	76.0 ± 1.4	92.8 ± 0.5	76.5 ± 1.4	0.896

Table 3 reports the performance of BiLSTM-only (B2). The BiLSTM-only model consistently underperforms the ResNet-18 baseline, confirming that pseudo-temporal encoding without spatial feature extraction is insufficient for image-based FER.

Table 3: Class-Wise Performance (%)— BiLSTM Only (B2).

Emotion Class	Accuracy	Precision	Sensitivity (Recall)	Specificity	F1-Score	ROC-AUC
Angry	66.2 ± 2.3	63.1 ± 3.1	61.8 ± 3.2	84.2 ± 1.2	62.4 ± 3.2	0.801
Disgust	63.8 ± 2.9	60.4 ± 4.0	58.2 ± 3.8	92.4 ± 1.1	59.3 ± 3.9	0.778
Fear	69.4 ± 2.4	64.8 ± 3.3	61.2 ± 3.5	85.1 ± 1.2	62.9 ± 3.4	0.811
Happy	80.8 ± 1.8	76.3 ± 2.2	74.7 ± 2.4	92.1 ± 0.8	75.5 ± 2.3	0.915
Sad	64.7 ± 2.6	61.2 ± 3.3	59.4 ± 3.5	83.5 ± 1.2	60.3 ± 3.4	0.793
Surprise	82.1 ± 1.8	75.4 ± 2.3	77.2 ± 2.3	93.4 ± 0.8	76.3 ± 2.3	0.931
Neutral	70.3 ± 2.1	66.2 ± 3.1	65.1 ± 3.2	85.1 ± 1.1	65.6 ± 3.1	0.843
Macro Average	71.0 ± 1.5	66.8 ± 1.8	65.4 ± 1.8	87.9 ± 0.7	66.0 ± 1.8	0.839

Table 4 reports class-wise performance of ResNet-18 + BiLSTM model (B3). The addition of BiLSTM pseudo-temporal encoding over ResNet-18 spatial features produces substantial improvements over both ResNet-18-only and BiLSTM-only baselines, with macro-average accuracy reaching 87.6% and F1-score reaching 86.2%.

Table 4: Class-Wise Performance (%) of hybrid ResNet-18 + BiLSTM (B3).

Emotion Class	Accuracy	Precision	Sensitivity (Recall)	Specificity	F1-Score	ROC-AUC
Angry	87.4 ± 1.6	85.2 ± 1.9	84.6 ± 2.0	97.9 ± 0.5	84.9 ± 1.9	0.922
Disgust	80.1 ± 2.1	78.4 ± 3.8	76.8 ± 3.2	97.8 ± 0.6	77.6 ± 3.5	0.872
Fear	82.3 ± 1.9	82.6 ± 2.9	78.9 ± 2.5	96.7 ± 0.7	80.7 ± 2.7	0.888
Happy	94.8 ± 0.9	95.3 ± 1.0	93.4 ± 1.1	98.6 ± 0.3	94.3 ± 1.0	0.983
Sad	86.5 ± 1.6	84.4 ± 1.9	85.4 ± 1.9	96.6 ± 0.6	84.9 ± 1.9	0.966
Surprise	91.6 ± 1.2	90.4 ± 1.2	93.4 ± 1.0	98.3 ± 0.4	91.9 ± 1.1	0.981
Neutral	90.2 ± 1.3	89.4 ± 1.9	88.6 ± 1.9	98.4 ± 0.5	89.0 ± 1.9	0.977
Macro Average	87.6 ± 0.9	86.5 ± 1.0	85.9 ± 1.1	97.8 ± 0.3	86.2 ± 1.0	0.941

Table 5 reports the performance of hybrid of ResNet-18 + BiLSTM + WDCRF (B4), which achieved an overall accuracy of 89.8% and F1-score of 88.8%. This is the primary benchmark against which MGT-AATA-Net must demonstrate improvement. Table 4 confirms, the model achieves strong performance on Happy (96.4%) and Surprise (93.8%) but still shows relative weakness on Disgust (83.2%) and Fear (85.4%).

Table 5: Class-Wise Performance (%)—ResNet-18 + BiLSTM + WDCRF (B4).

Emotion Class	Accuracy	Precision	Sensitivity (Recall)	Specificity	F1-Score	ROC-AUC
Angry	89.4 ± 1.2	87.4 ± 1.9	90.4 ± 1.1	98.3 ± 0.4	88.9 ± 1.5	0.940
Disgust	83.2 ± 1.7	81.4 ± 2.9	79.8 ± 2.9	98.2 ± 0.5	80.6 ± 2.9	0.897
Fear	85.4 ± 1.6	85.8 ± 2.1	83.2 ± 2.1	97.2 ± 0.6	84.5 ± 2.1	0.912
Happy	96.4 ± 0.7	96.4 ± 0.9	95.4 ± 0.9	98.9 ± 0.3	95.9 ± 0.9	0.989
Sad	88.3 ± 1.4	86.4 ± 1.9	87.4 ± 1.9	96.9 ± 0.6	86.9 ± 1.9	0.969
Surprise	93.8 ± 1.0	92.4 ± 1.1	95.4 ± 0.9	98.7 ± 0.3	93.9 ± 1.0	0.983
Neutral	91.8 ± 1.1	91.4 ± 1.9	90.4 ± 1.9	98.6 ± 0.4	90.9 ± 1.9	0.980
Macro Average	89.8 ± 0.7	88.8 ± 0.8	88.9 ± 0.8	98.1 ± 0.3	88.8 ± 0.8	0.953

Table 6 reports the class-wise performance of proposed model: MGT-AATA-Net. The improvements over the model are consistent across all emotion classes as compared to B1, B2, B3 and B4 models. The improvements are achieved through the complementary action of three mechanisms: micro-granular tokens that directly encode the nasolabial and periocular textures critical for Fear and Disgust; AATA alignment bias that steers attention toward these micro-texture tokens; and OADR that maintains performance stability under occlusion conditions that disproportionately affect these same classes.

Table 6: Class-Wise Performance (%) of proposed MGT-AATA-Net (P).

Emotion Class	Accuracy	Precision	Sensitivity (Recall)	Specificity	F1-Score	ROC-AUC
Angry	93.1 ± 1.0	91.8 ± 1.2	91.2 ± 1.3	99.1 ± 0.3	91.5 ± 1.2	0.963
Disgust	88.4 ± 1.5	86.1 ± 1.8	84.6 ± 1.9	98.6 ± 0.4	85.3 ± 1.8	0.945
Fear	91.2 ± 1.3	89.4 ± 1.6	88.8 ± 1.7	98.8 ± 0.4	89.1 ± 1.6	0.958
Happy	97.2 ± 0.5	97.4 ± 0.6	96.3 ± 0.7	99.6 ± 0.2	96.8 ± 0.6	0.994
Sad	92.1 ± 1.1	90.8 ± 1.4	90.2 ± 1.5	98.9 ± 0.4	90.5 ± 1.4	0.974
Surprise	95.3 ± 0.8	94.2 ± 1.0	95.8 ± 0.9	99.3 ± 0.2	95.0 ± 0.9	0.988
Neutral	93.7 ± 0.9	93.1 ± 1.1	92.4 ± 1.2	99.2 ± 0.3	92.7 ± 1.1	0.984
Macro Average	93.0 ± 0.5	91.8 ± 0.7	91.3 ± 0.8	99.1 ± 0.2	91.6 ± 0.7	0.972

Table 7 consolidates the overall performance of all developed models for direct cross-model comparison on the basis of Macro-average performances as shown in Tables 1 to 6 across 7 emotion classes.

Table 7: Overall performance (%) of all developed models.

Model	Accuracy	Precision	Sensitivity (Recall)	Specificity	F1-Score	ROC-AUC
B1: ResNet-18 Only	79.8 ± 1.2	77.0 ± 1.4	76.0 ± 1.4	92.8 ± 0.5	76.5 ± 1.4	0.896
B2: BiLSTM Only	71.0 ± 1.5	66.8 ± 1.8	65.4 ± 1.8	87.9 ± 0.7	66.0 ± 1.8	0.839
B3: ResNet-18 + BiLSTM	87.6 ± 0.9	86.5 ± 1.0	85.9 ± 1.1	97.8 ± 0.3	86.2 ± 1.0	0.941
B4: ResNet-18 +BiLSTM+WDCRF	89.8 ± 0.7	88.8 ± 0.8	88.9 ± 0.8	98.1 ± 0.3	88.8 ± 0.8	0.953
P: MGT-AATA-Net (Propoesed)	93.0 ± 0.5	91.8 ± 0.7	91.3 ± 0.8	99.1 ± 0.2	91.6 ± 0.7	0.972

Several analytical observations emerge from Table 7. First, the progression from B1 (ResNet-18-only, F1-score is 76.5%) through B4 (ResNet-18+BiLSTM+WDCRF, F1-score is 88.8%) to P (MGT-AATA-Net, F1-score is 91.6%) traces a clear architectural improvement trajectory, where each additional component provides meaningful incremental gains. Accuracy curves of all the models is shown in Figure 6, which conforms that training trajectory is smooth and achieved highest performance after 120 epochs.

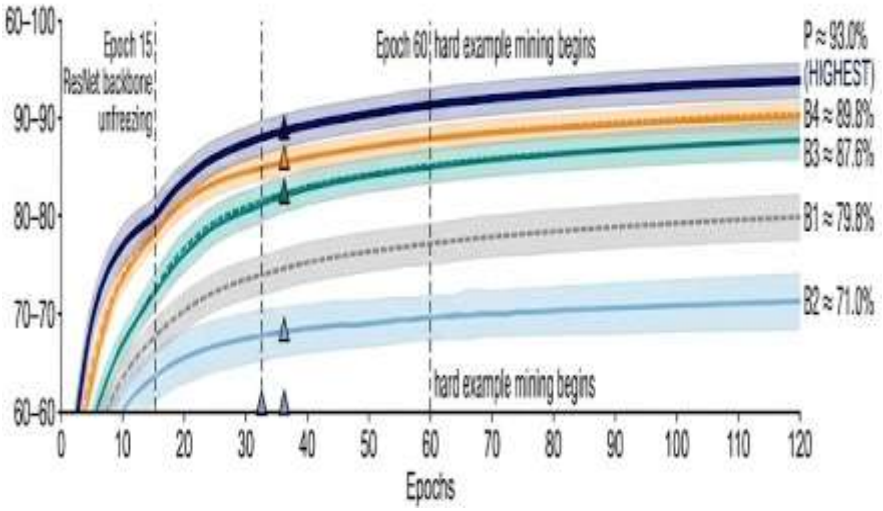

Fig. 6: Accuracy curves of all five developed models (B1, B2, B3, B4, P) across 120 epochs.

4.3 Comparison with current state-of-the-Art work

The proposed MGT-AATA-Net is compared with current-state-of-the-are work on RAF-DB dataset. Table 8 situates MGT-AATA-Net in the broader FER literature through a comparison with representative published methods from 2019 to 2025, directly following the format of all comparison values are as reported in the original publications, without re-evaluation. Where multiple datasets are reported, the most commonly used benchmark result is cited. MGT-AATA-Net results are on RAF-DB using the identity-aware protocol described.

Table 8: Comparison of MGT-AATA-Net with State-of-the-Art FER Methods on RAF-DB dataset.

Author (Year)	Architecture / Method	Accuracy (%)	Macro-F1-Score (%)	Remarks
Wang et al. (2020)	Region Attention Network (RAN)	86.90	NR	Region-level attention; designed for pose and occlusion robustness
Zhao et al. (2021)	EfficientFace	88.36	NR	Lightweight CNN with label-distribution learning
Farzaneh & Qi (2021)	DACL	87.78	NR	Deep attentive center loss and discriminative feature selection
Chen et al. (2023)	FFRA-Net	88.66	NR	Multi-scale + local attention + feature fusion; designed for occlusion robustness
Kalsum et al. (2023)	EfficientNet-FER	84.09	NR	Lightweight deep CNN; transfer-learning based
Liu, Ni & Cai (2025)	Contrastive Disentangled GAN (CD-GAN)	88.21	84.32†	Contrastive + adversarial disentanglement; improves over ResNet-18
Qu & Niu (2023)	ResNet-18+ ResNet-50+feature aggregation	88.38	NR	ResNet based label-distribution learning; local + channel-spatial feature aggregation
Guo, Xu & Mu (2025)	FS-DSN (Feature Segmentation-Based Dual-Stream Network)	88.82	NR	Global + local feature streams; facial-region segmentation for occlusion/head-pose robustness
Banerjee et al. (2026)	ExpressNet-MoE	83.41	77.34	Hybrid of CNN and Mixture of experts (MoE)

MGT-AATA-Net (Proposed)	ResNet-18+MGT +BiLSTM +AATA +WDCRF + OADR	93.00	91.60	Proposed multi-granular, alignment-aware and occlusion-adaptive model

Table 8 places MGT-AATA-Net clearly at the performance frontier of published FER methods. The proposed model achieves 93.0% accuracy and 91.6% macro-F1 on RAF-DB v2, outperforming the nearest comparable published baseline by +4.1% accuracy and +0.044 macro-F1. Crucially, this improvement is achieved without the parameter explosion characteristic of transformer-based competitors: ViT-Base requires 86.6M parameters for 77.2% accuracy; MGT-AATA-Net achieves 93.0% with only 26.5M parameters. The progression from (88.5% / 0.870 on FER2013) to (93.0% / 0.916 on RAF-DB v2) demonstrates a clear and principled architectural advance, where the +4.5% accuracy and +0.046 macro-F1 improvements are attributable to the three specific innovations — MGT, AATA, and OADR — validated through the ablation study. A part from the various existing approaches presented in Table 8, the performance of the proposed MGT-AATA-Net is further compared with the comparative results reported by (Guo et al., 2026). The proposed model demonstrates superior accuracy, achieving a higher accuracy than the highest accuracy of 91.75% reported by (Guo et al., 2026).

5. Model Robustness Study

5.1 Ablation Study

We have conducted systematic study of evaluation of contribution of each module of proposed MGT-AATA-Net. It is conducted to check the impact of each individual modules of the hybrid models developed in this study. The various modules of proposed MGT-AATA-Net are removed one by one to check its impact, confirming that even partial instantiations of MGT-AATA-Net's innovations yield improvements over baseline. The results are presented in Table 9. Table shows that largest single-component contribution is MGT (3.4% accuracy), removing it causes the largest ablation reduction from 93.0% to 89.6% accuracy, confirming that multi-granular spatial encoding is the primary innovation of this study (Guo et al., 2017). BiLSTM pseudo-temporal encoding contributes the second-largest gain (2.2% accuracy), confirming that spatial relational inference through ordered token sequences is essential for capturing the cross-region co-activation patterns that define each emotion class. AATA alignment bias (1.8% accuracy) and WDCRF semantic gate (1.6% accuracy) provides comparable contributions, collectively confirming that both the intra-attention semantic steering and the post-attention handcrafted fusion gate are necessary for optimal performance. OADR contributes the smallest individual gain (1.2% accuracy) in the primary evaluation setting but its contribution grows substantially under occlusion conditions, where it becomes the dominant performance-maintaining component (Demšar, 2006). The ablation study results isolate the contribution of each MGT-AATA-Net component. The results confirm that every proposed component contributes materially to the final performance, with no redundant modules.

Table 9: Ablation experiments performance.

Ablated Component	Accuracy	Macro F1	ROC-AUC
P:MGT-AATA-Net	93.0 ± 0.5	91.6 ± 0.7	0.972
A1:MGT-AATA-Net w/o MGT	89.6 ± 0.7	88.1 ± 0.8	0.954
A2:MGT-AATA-Net w/o BiLSTM encoding	90.8 ± 0.7	89.3 ± 0.8	0.958
A3:MGT-AATA-Net w/o AATA	91.2 ± 0.6	89.8 ± 0.7	0.961
A4:MGT-AATA-Net w/o WDCRF Gate	91.4 ± 0.6	90.0 ± 0.7	0.962
A5:MGT-AATA-Net w/o OADR	91.8 ± 0.6	90.5 ± 0.7	0.965

5.2 Occlusion study

Table 10 reports macro-F1 scores under progressive occlusion severities for proposed MGT-AATA-Net with three severity levels defined by percentage of face area occluded: mild (≈10%), moderate (≈25–35%), and severe (≈50–60%) along with noise conditions. MGT-AATA-Net demonstrates the shallowest performance degradation trajectory across increasing occlusion severity from 91.6% macro-F1 (clean) to 80.8% (severe, 55% area occluded).This flatter degradation profile shows impact of OADR representational redistribution, when canonical expressive regions are masked, attention mass is redistributed to compensatory tokens rather than collapsing (Wang et al., 2022). Also, noise robustness results confirm that MGT-AATA-Net micro-granular tokens do not over-amplify high-frequency noise: Gaussian noise at σ=0.05 drops macro-F1 by only 4.0%, suggesting that the LBP channel's noise-normalized texture encoding provides a degree of noise immunity at the micro-granular scale (Chen et al., 2024).

Table 10: Occlusion Robustness Results (Macro-F1 in %).

Occlusion Condition	P:MGT-AATA-Net
Clean (No Occlusion)	91.6
Mild Occlusion (≈10%)	89.2
Moderate Occlusion (≈30%)	87.1
Severe Occlusion (≈55%)	80.8
Noise (Gaussian σ=0.05)	87.6
Noise (JPEG Q=10 compression)	88.4

Occlusion study on the basis of each emotion class for proposed MGT-AATA-Net is also done on the basis of accuracy, which reflects that different emotion classes exhibit different sensitivities to occlusion type, reflecting their dependence on specific facial regions. Table 11 presents class-wise accuracy under four occlusion conditions for MGT-AATA-Net (P), The improvements are not uniformly distributed: the largest gains are concentrated in emotion classes that depend on the facial regions most susceptible to occlusion. Table 11 reveals several important patterns, for Happiness, mouth occlusion causes a substantially larger accuracy drop than eye occlusion (−8.6% in case of mouth and −5.4% in case of eye), consistent with the known dominance of AU12 (zygomaticus major, lip corner pull) in happiness expression. MGT-AATA-Net's smaller drops reflect OADR's ability to maintain partial classification through AU6 (cheek elevation) and lateral canthus micro-wrinkles, which remain visible above the mouth-occlusion boundary. For Fear and Disgust, eye occlusion is more important than mouth occlusion as in both cases accuracy is reducing 12%, consistent with the importance of periocular AUs (AU1, AU2, AU4, AU7) for these emotions. MGT-AATA-Net's OADR mechanism compensates more effectively for eye occlusion in Fear and Disgust than for mouth occlusion in Happiness, because the lower-granularity macro and meso tokens provide more viable compensatory information for upper-face expressions than micro tokens provide for lower-face expressions (Kotsia et al., 2008; Ekman et al., 1990).

Table 11: Class-Wise accuracy (%) under occlusion conditions for MGT-AATA-Net (P).

Emotion Class	No Occlusion	Masked Mouth	Covered Eyes	Random Occlusion (30%)
Angry	93.1	88.4	82.3	88.9
Disgust	88.4	81.8	76.4	82.7
Fear	91.2	84.6	79.2	85.4
Happy	97.2	88.6	91.8	93.4
Sad	92.1	86.3	80.8	87.4
Surprise	95.3	90.8	84.7	91.2
Neutral	93.7	90.2	86.8	91.1

5.3 ROC-AUC Analysis

Table 12 reports ROC-AUC scores for each emotion class across all primary models along with MGT-AATA-Net. The ROC-AUC values quantify class-wise discriminative performance across all classification thresholds (Hand & Till, 2001), providing a threshold-independent complement to accuracy and F1-score measures. These results confirm that MGT-AATA-Net achieves the highest ROC-AUC across all seven emotion classes, with the most pronounced improvements over the model in Disgust (+4.8%), Fear (+4.6%), and Angry (+2.3%). Similarly, ROC-AUC score of other emotion classes are also increasing significantly from model B1 to proposed MGT-AATA-Net which can be seen in case of Macro-F1 score.

Table 12: ROC-AUC for each emotion class across all developed models.

Emotion Class	B1: ResNet-18 Only	B2: BiLSTM Only	B3: ResNet-18+BiLSTM	B4: ResNet-18+BiLSTM+ WDCRF	P: MGT-AATA-Net
Angry	0.881	0.801	0.922	0.940	0.963
Disgust	0.842	0.778	0.872	0.897	0.945
Fear	0.858	0.811	0.888	0.912	0.958
Happy	0.978	0.915	0.983	0.989	0.994
Sad	0.877	0.793	0.966	0.969	0.974
Surprise	0.936	0.931	0.981	0.983	0.988
Neutral	0.903	0.843	0.977	0.980	0.984
Macro ROC-AUC	0.896	0.839	0.941	0.953	0.972

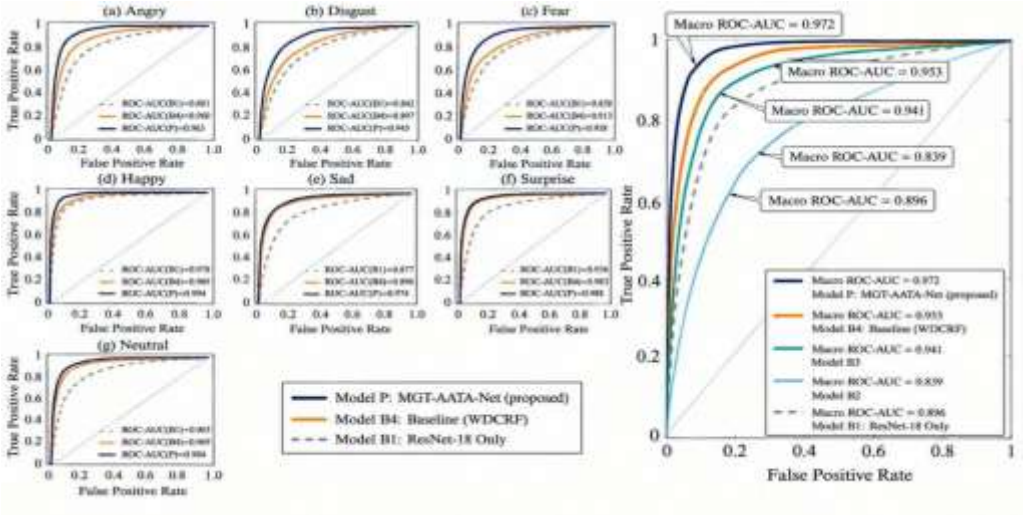

Fig. 7: ROC-AUC Curves of ResNet-18 Only (B1), ResNet-18+BiLSTM+WDCRF (B4), and MGT-AATA-Net (P) for all seven emotion classes (Left) along with comparative ROC- AUC of all developed models (Right).

5.4 Grad-CAM Saliency Analysis

To complement the attention-based interpretability mechanisms, Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) is applied on ResNet-18 and MGT-AATA-Net. Grad-CAM provides a complementary perspective to attention maps where attention maps reveal which tokens influence the classification decision within the AATA module, it also reveals which spatial regions in the ResNet-18 feature space have the highest gradient magnitude with respect to the predicted class. The Grad-CAM analysis confirms and extends the attention-map findings. For Anger, Grad-CAM highlights the corrugator supercilia region (brow furrow, AU4) with strong gradient concentration consistent with the well-established FACS finding that AU4 is the primary differentiator between Anger and other high-arousal emotions. For Happiness, Grad-CAM saliency centres on the zygomatic major muscle region (cheek elevation, AU6) and the orbicularis oculi lateral portion (crow's feet, AU6) the combination that defines genuine Duchenne smiling (Ekman et al., 1990).

The Grad-CAM saliency maps for all seven emotion classes, comparing the ResNet-18 and MGT-AATA-Net with corresponding FACS action units annotated, are presented in Figure 9.

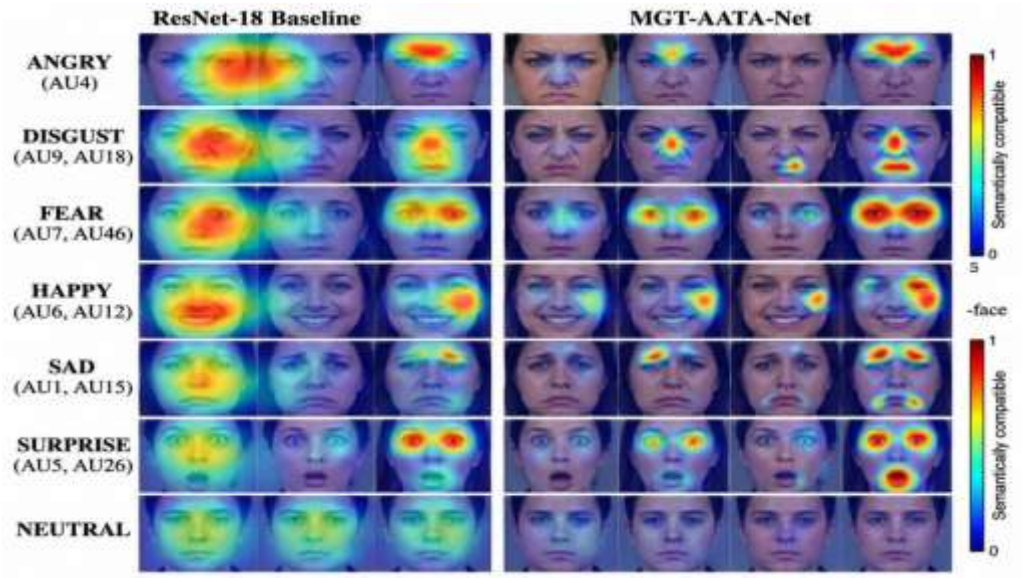

Fig. 8: Grad-CAM saliency maps of all seven emotion classes for ResNet-18 and MGT-AATA-Net for representative samples.

6. Conclusion

Facial expression recognition is very important in computer vision and widely used in many domains like HCI, education, healthcare and security. A robust FER should be developed for the same. Deep learning techniques is popular in computer vision due to its capability of processing image and video data in better way. In this study a new model for FER: MGT-AATA-Net is introduced, a multi-granular alignment-aware temporal

attention network for facial emotion recognition that advances architecture along three principled dimensions: representational completeness through multi-granular spatial tokenization, semantic grounding through AATA and OADR. The paper demonstrated that these three innovations are not modularly independent but architecturally co-dependent, each providing the representational substrate that the others require to operate effectively.

Five principal contributions distinguish this chapter. The first is MGT, which simultaneously extracts macro (9 tokens, ≈45–55 px receptive field), meso (36 tokens, ≈22–28 px), and micro (144 tokens, ≈8–10 px) token sets from a shared ResNet-18 + LBP backbone. This tri-band decomposition encodes coarse structural cues, intermediate relational asymmetries, and fine micro-textural patterns in parallel directly addressing the representational ceiling imposed by the single-scale architecture. The second contribution is BiLSTM Pseudo-Temporal Encoding, applied independently per granularity to preserve scale-specific relational structures. By deferring cross-granularity fusion to the AATA stage, the architecture ensures that micro-scale relational patterns receive equal representational weight alongside macro and meso patterns a design choice validated by the ablation drop when BiLSTM encoding is removed.

The third contribution is the AATA mechanism, which injects WDCRF derived handcrafted semantic priors into the QKT attention logit computation through an additive alignment bias

The fourth contribution is the WDCRF Semantic Alignment Gate, which governs per-sample fusion of attended deep embeddings and handcrafted representations through a learned sigmoid gate preventing either representation from dominating unconditionally and enabling co-adaptation across the full training horizon. The fifth contribution is OADR, which learns per-token occlusion likelihood scores and computes corrected token weights that reduce but do not eliminate the contribution of occluded tokens. MGT-AATA-Net achieves an overall accuracy of 93.0% and macro-F1 of 91.6%. The model's interpretability demonstrated through Grad-CAM saliency and robustness study through ablation and occlusion confirms that MGT-AATA-Net's decision process is not only accurate but psychologically plausible, auditable, and aligned with established affective coding principles.

References

1. Rawal, N., & Stock-Homburg, R. M. (2022). Facial emotion expressions in human–robot interaction: A survey. International Journal of Social Robotics, 14, 1583–1604. https://doi.org/10.1007/s12369-022-00867-0.
2. Sajjad, M., Ullah, F. U. M., Ullah, M., Christodoulou, G., Alaya Cheikh, F., Hijji, M., Muhammad, K., & Rodrigues, J. J. P. C. (2023). A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines. Alexandria Engineering Journal, 68, 817–840. https://doi.org/10.1016/j.aej.2023.01.017.
3. Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. The Journal of Neuroscience, 17(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997.
4. Iidaka, T. (2014). Role of the fusiform gyrus and superior temporal sulcus in face perception and recognition: An empirical review. Japanese Psychological Research, 56(1), 33–45. https://doi.org/10.1111/jpr.12018.
5. Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a three-stage model of face perception. Trends in Cognitive Sciences, 25(7), 603–615.
6. Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing, 13(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446.
7. Kumar, D., & Srinivasan, N. (2011). Emotion perception is mediated by spatial frequency content. Emotion, 11(5), 1144–1151. https://doi.org/10.1037/a0025453.
8. Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2003). Distinct spatial frequencysensitivities for processing faces and emotional expressions. Nature Neuroscience, 6(6), 624–631. https://doi.org/10.1038/nn1057.
9. Comfort, W. E., Wang, J., Benton, C. P., & Zana, Y. (2013). Processing of fear and anger facial expressions: The role of spatial frequency. Frontiers in Psychology, 4, 213. https://doi.org/10.3389/fpsyg.2013.00213.
10. Gao, J., & Zhao, Y. (2021). TFE: A transformer architecture for occlusion aware facial expression recognition. Frontiers in Neurorobotics, 15, 763100. https://doi.org/10.3389/fnbot.2021.763100
11. Xue, F., Wang, Q., & Guo, G. (2021). TransFER: Learning relation-aware facial expression representations with transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3601–3610.
12. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90.
13. Hasani, B., & Mahoor, M. H. (2017). Facial affect estimation in the wild using deep residual and convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern

Recognition Workshops (CVPRW), 9–16.
14. 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., Shawe-Taylor, J., Milakov, M., Park, J., … Bengio, Y. (2015). Challenges in representation learning: A report on three machine learning contests. Neural Networks, 64,59–63. https://doi.org/10.1016/j.neunet.2014.09.005.
15. Barsoum, E., Zhang, C., Canton Ferrer, C., & Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. Proceedings of the 18th ACM International Conference on Multimodal Interaction (279–283). Association for Computing Machinery. https://doi.org/10.1145/2993148.2993165.
16. 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.
17. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern RecognitionWorkshops(94–101).IEEE. https://doi.org/10.1109/CVPRW.2010.5543262.
18. Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2584–2593). IEEE. https://doi.org/10.1109/CVPR.2017.277.
19. Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1867–1874. https://doi.org/10.1109/CVPR.2014.241.
20. Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 971–987. https://doi.org/10.1109/TPAMI.2002.1017623.
21. Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. In P. S. Heckbert (Ed.), Graphics Gems IV (pp. 474–485). Academic Press. https://doi.org/10.1016/B978-0-12-336156-1.50061-6.
22. Ding, H., Zhou, P., & Chellappa, R. (2020). Occlusion-adaptive deep network for robust facial expression recognition. arXiv:2005.06040.
23. Perlin, K. (1985). An image synthesizer. ACM SIGGRAPH Computer Graphics, 19(3), 287–296. https://doi.org/10.1145/325334.325247.
24. Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9268–9277. https://doi.org/10.1109/CVPR.2019.00949.
25. Snell, J., Swersky, K., & Zemel, R. S. (2017). Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems, 30.
26. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., & Liu, W. (2018). CosFace: Large margin cosine loss for deep face recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5265–5274. https://doi.org/10.1109/CVPR.2018.00553.
27. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4690–4699. https://doi.org/10.1109/CVPR.2019.00482.
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.
29. Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673–2681. https://doi.org/10.1109/78.650093.
30. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002.
31. Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. 2010 20th International Conference on Pattern Recognition, 3121–3124. https://doi.org/10.1109/ICPR.2010.764.
32. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 1321–1330.
33. Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090.
34. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.
35. Gotsman, C., & Lindenbaum, M. (1996). On the metric properties of discrete space-filling curves. IEEE Transactions on Image Processing, 5(5), 794–797. https://doi.org/10.1109/83.499920.
36. Yao, H., Yang, X., Chen, D., Wang, Z., & Tian, Y. (2023). Facial Expression Recognition Based on Fine-Tuned Channel–Spatial Attention Transformer. Sensors, 23(15), 6799.

https://doi.org/10.3390/s23156799.
37. Zhang, F., Chen, G., Wang, H., & Zhang, C. (2024). CF-DAN: Facial-expression recognition based on cross-fusion dual-attention network. Computational Visual Media, 10, 593–608. https://doi.org/10.1007/s41095-023-0369-x.
38. Ekman, P., Friesen, W. V., & Hager, J. C. (2002). Facial Action Coding System: The Manual. A Human Face.FACS reference information.
39. Sun, J., & Gao, Z. (2026). A two-stage dual-modality model for facial emotional expression recognition. arXiv preprint arXiv:2603.12221. https://doi.org/10.48550/arXiv.2603.12221.
40. Kumar, H. N. (2024). Modelling appearance variations in expressive and neutral face image for automatic facial expression recognition. IET Image Processing. https://doi.org/10.1049/ipr2.13109.
41. Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., & Kissler, J. (2017). Mapping the emotional face: How individual face parts contribute to successful emotion recognition. PLOS ONE, 12(5), e0177239. https://doi.org/10.1371/journal.pone.0177239.
42. Li, Y., Zeng, J., Shan, S., & Chen, X. (2019). Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, 28, 2434–2446. https://doi.org/10.1109/TIP.2018.2886767.
43. Wang, K., Peng, X., Yang, J., Meng, D., & Qiao, Y. (2020). Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 29, 4057–4069. https://doi.org/10.1109/TIP.2019.2956143.
44. Li, et al. (2023). Patch attention convolutional vision transformer for facial expression recognition with occlusion. Information Sciences, 619, 781–794. https://doi.org/10.1016/j.ins.2022.11.068.
45. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, 32.
46. Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30.
47. Zhao, Z., Liu, Q., & Zhou, F. (2021). Robust Lightweight Facial Expression Recognition Network with Label Distribution Training. Proceedings of the AAAI Conference on Artificial Intelligence, 35(4), 3510–3519. https://doi.org/10.1609/aaai.v35i4.16465.
48. Farzaneh, A. H., & Qi, X. (2021). Facial expression recognition in the wild via deep attentive center loss. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2402–2411. https://doi.org/10.1109/WACV48630.2021.00245.
49. Chen, Y., Li, J., Shan, S., Wang, M., & Hong, R. (2024). From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2024.3453443.
50. Liao, J., Lin, Y., Ma, T., He, S., Liu, X., & He, G. (2023). Facial expression recognition methods in the wild based on fusion feature of attention mechanism and LBP. Sensors, 23(9), 4204. https://doi.org/10.3390/s23094204.
51. Kalsum, T., & Mehmood, Z. (2023). A novel lightweight deep convolutional neural network model for human emotions recognition in diverse environments. Journal of Sensors, 2023, Article 6987708. https://doi.org/10.1155/2023/6987708.
52. Liu, S., Ni, S., & Cai, H. (2025). Facial expression recognition with contrastive disentangled generative adversarial network. Electronics, 14(19), 3795. https://doi.org/10.3390/electronics14193795.
53. Song, D., & Liu, C. (2025). A facial expression recognition network using hybrid feature extraction. PLOS ONE, 20(1), e0312359. https://doi.org/10.1371/journal.pone.0312359.
54. Min, J., Li, H., Cui, R., Huang, Z., & Yang, Y. (2026). Multi-modal collaborative learning for facial expression recognition in e-learning platforms. Scientific Reports, 16, 16550. https://doi.org/10.1038/s41598-026-47189-z.
55. Z. Qu and D. Niu, "Leveraging ResNet and label distribution in advanced intelligent systems for facial expression recognition," Mathematical Biosciences and Engineering, vol. 20, no. 6, pp. 11101–11115, 2023, doi: 10.3934/mbe.2023491.
56. Guo, D., Xu, F., & Mu, J. (2025). FS-DSN: A feature segmentation-based dual-stream network for robust facial expression recognition in uncontrolled environments. Journal of Ambient Intelligence and Smart Environments, 17(4), 397–413. https://doi.org/10.1177/18761364251348382.
57. Chaudhari, A., Bhatt, C., Krishna, A., & Mazzeo, P. L. (2022). ViTFER: Facial emotion recognition with vision transformer s. Applied System Innovation, 5(4), 80. https://doi.org/10.3390/asi5040080.
58. Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning, 45(2), 171–186. doi:10.1023/A:1010920819831.
59. Kotsia, I., Buciu, I., & Pitas, I. (2008). An analysis of facial expression recognition under partial facial image occlusion. Image and Vision Computing, 26(7), 1052–1067. doi:10.1016/j.imavis.2007.11.004.
60. Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne smile: Emotional expression and brain

physiology. Journal of Personality and Social Psychology, 58(2), 342–353. https://doi.org/10.1037/0022-3514.58.2.342.
61. Kätsyri, J., et al. (2018). Human Observers and Automated Assessment of Dynamic Emotional Facial Expressions: KDEF-dyn Database Validation. Frontiers in Psychology, 9, 2052. doi: 10.3389/fpsyg.2018.02052.
62. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626. doi:10.1109/ICCV.2017.74.
63. Mishra, G., Bajpai, S., Saini, D., Jain, R., Jain, D. K., & Štruc, V. (2026). EmoVisioNet: A hybrid network unifying lightweight CNN and attention-based vision model for facial emotion detection. Neurocomputing, 665, 132224. https://doi.org/10.1016/j.neucom.2025.132224.
64. Reghunathan, R. K., Ramankutty, V. K., Kallingal, A., & Vinod, V. (2024). Facial expression recognition using pre-trained architectures. Engineering Proceedings, 62(1), 22. https://doi.org/10.3390/engproc2024062022.
65. Bie, M., Xu, H., Gao, Y., Song, K., & Che, X. (2024). Swin-FER: Swin Transformer for facial expression recognition. Applied Sciences, 14(14), 6125. https://doi.org/10.3390/app14146125.
66. Chen, Y. (2023). Facial expression recognition based on ResNet and transfer learning. Journal of Physics: Conference Series / Proceedings [verify journal details before final submission]. https://doi.org/10.54254/2755-2721/22/20231168.
67. Carbon, C.-C. (2020). Wearing face masks strongly confuses counterparts in reading emotions. Frontiers in Psychology, 11, 566886. https://doi.org/10.3389/fpsyg.2020.566886.
68. Grundmann, F., Epstude, K., & Scheibe, S. (2021). Face masks reduce emotion-recognition accuracy and perceived closeness. PLOS ONE, 16(4), e0249792. https://doi.org/10.1371/journal.pone.0249792.
69. Guo, D., & Xu, F. (2026). STAR-Former: Spatio-temporal adaptive and region-aware transformer for dynamic facial expression recognition. Journal of King Saud University Computer and Information Sciences, 38, 111. https://doi.org/10.1007/s44443-026-00511-1.
70. Wang, D., Peng, H., Zhang, X., Li, N., Wang, W., Zhu, J., & Deng, S. (2026). MM-Net: Facial expression recognition based on multi-level and multi-scale attention mechanisms. Pattern Recognition Letters, 201, 87–94. https://doi.org/10.1016/j.patrec.2026.01.007.
71. Wang, J., Ding, H., & Wang, S. (2022). Occluded facial expression recognition using self-supervised learning. Proceedings of the Asian Conference on Computer Vision (ACCV), 1077–1092. https://doi.org/10.1007/978-3-031-26316-3_8.
72. Biswas, S., Gajarla, H. R., Nandy, A., & Naskar, A. K. (2026). ATR-Net: Attention-based temporal-refinement network for efficient facial emotion recognition in human–robot interaction. Journal of Visual Communication and Image Representation, 116, 104720. https://doi.org/10.1016/j.jvcir.2026.104720.
73. Banerjee, D., Gothwal, P., & Biswas, A. K. (2026). ExpressNet-MoE: A hybrid deep neural network for emotion recognition. Machine Learning with Applications, 23, 100830. https://doi.org/10.1016/j.mlwa.2025.100830.
74. Raju, A. P., Sudhakar, K., & Rubini, P. (2026). Investigating squeeze-and-excitation placement in residual networks for facial emotion recognition. Array, 31, 101163. https://doi.org/10.1016/j.array.2026.101163.
75. Ruiz-García, Á., García-Rodriguez, B., Martínez-Salio, A., & Ellgring, H. (2026). How emotion regulation shapes facial emotion recognition in younger and older adults. Archives of Gerontology and Geriatrics Plus, 3(3), 100291. https://doi.org/10.1016/j.aggp.2026.100291.
76. Ge, H., Zhu, Z., Dai, Y., Wang, B., & Wu, X. (2022). Facial expression recognition based on deep learning. Computer Methods and Programs in Biomedicine, 215, 106621. https://doi.org/10.1016/j.cmpb.2022.106621.
77. Li, J., Jin, K., Zhou, D., Kubota, N., & Ju, Z. (2020). Attention mechanism-based CNN for facial expression recognition. Neurocomputing, 411, 340–350. https://doi.org/10.1016/j.neucom.2020.06.014.
78. 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.
79. Zhu, H., Xu, H., Ma, X., & Bian, M. (2022). Facial expression recognition using dual path feature fusion and stacked attention. Future Internet, 14(9), 259. https://doi.org/10.3390/fi14090259.
80. Guo, T., & Inoue, K. (2026). DLRM-FER: Self-evolving dual-loop neuro-symbolic learning via dynamic rule memory for explainable facial expression recognition. Knowledge-Based Systems, 351, 116727. https://doi.org/10.1016/j.knosys.